Computers and Chemical Engineering 177 (2023) 108345

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/cace

Integrating model-based design of experiments and computer-aided solvent design

Lingfeng Gui [a], Yijun Yu [a], Titilola O. Oliyide [a], Eirini Siougkrou [a], Alan Armstrong [b], Amparo Galindo [a], Fareed Bhasha Sayyed [c], Stanley P. Kolis [d], Claire S. Adjiman [a,*]

[a] *Department of Chemical Engineering, The Sargent Centre for Process Systems Engineering and Institute for Molecular Science and Engineering, Imperial College London, London, SW7 2AZ, UK*

[b] *Department of Chemistry and Institute for Molecular Science and Engineering, Imperial College London, Molecular Sciences Research Hub, White City Campus, London, W12 0BZ, UK*

[c] *Synthetic Molecule Design and Development, Eli Lilly Services India Pvt Ltd, Devarabeesanahalli, Bengaluru, 560103, India*

[d] *Synthetic Molecule Design and Development, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, 46285, IN, United States*

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.7839710

MSC:
62K05
80A30
90C11

Keywords:
Model-based design of experiments
Computer-aided molecular design
Solvent design
Solvatochromic equation

ABSTRACT

Computer-aided molecular design (CAMD) methods can be used to generate promising solvents with enhanced reaction kinetics, given a reliable model of solvent effects on reaction rates. Herein, we use a surrogate model parameterised from computer experiments, more specifically, quantum-mechanical (QM) data on rate constants. The choice of solvents in which these computer experiments are performed is critical, considering the cost and difficulty of these QM calculations. We investigate the use of model-based design of experiments (MBDoE) to identify an information-rich solvent set and integrate this within a QM-CAMD framework. We find it beneficial to consider a wide range of solvents in designing the solvent set, using group contribution techniques to predict missing solvent properties. We demonstrate, via three case studies, that the use of MBDoE yields surrogate models with good statistics and leads to the identification of solvents with enhanced predicted performance with few iterations and at low computational cost.

1. Introduction

Reaction kinetics is a key factor in the manufacture of chemical products. For instance, in the pharmaceutical industry, reaction kinetics affect productivity, yield and product purity in the production of active pharmaceutical ingredients and synthetic intermediates, which in turn influence the cost and complexity of downstream separations (Grom et al., 2016). When producing functional polymers, such as polysiloxanes for the manufacture of adhesives and lubricants (Hill, 2005), the reaction kinetics at the different stages of polymerisation determine the properties of the final product and its suitability for application (Issa and Luyt, 2019). Likewise, in the food industry, controlling the kinetics of reactions involved in food processing, such as the well-known Maillard reaction, can produce desired aromas and colours in processed food (Martins et al., 2000). It can be seen from these examples that an ability to control the reaction kinetics of a system is important for achieving high product quality, low process cost and high atom efficiency (Song et al., 2017).

In this context, the factors that affect the reaction kinetics, such as temperature, pH, catalysts and solvents, are of interest. In particular, the choice of solvent is especially important for liquid-phase reactions as the reaction medium alters the free energy landscape of the system. The Menschutkin reaction (Menschutkin, 1890a,b), for example, is well-known for its sensitivity to solvent effects as the rate constant of the reaction can vary by several orders of magnitude in different solvents. Usually solvents with larger dielectric constants favour the Menshutkin reaction due to the formation of charged products from neutral reactants (Reinheimer et al., 1963). In a more recent example, in the coupling of amino acids, changing the reaction solvent was found to not only accelerate the coupling reaction but also to suppress the side reaction between two amino-acid activation reagents, ethyl cyano(hydroxyimino)acetate (Oxyma) and diisopropylcarbodiimide (DIC), which produces hydrogen cyanide (HCN) (Erny et al., 2020). Solvent selection can therefore be a very rewarding, albeit by no means trivial, task in many situations where reaction kinetics play an important role. Despite its central role, the investigation of solvent effects on reaction kinetics and selectivity, and the use of the knowledge gained from this for solvent selection still often relies heavily on

[*] Corresponding author.
E-mail address: c.adjiman@imperial.ac.uk (C.S. Adjiman).

https://doi.org/10.1016/j.compchemeng.2023.108345
Received 18 April 2023; Received in revised form 28 June 2023; Accepted 1 July 2023
Available online 8 July 2023

experimental trial-and-error approaches, as has been the case in the two examples cited.

Although performing experiments is often seen as the most straightforward way to test reaction and solvent performance, this requires time and resources to purchase or synthesise the necessary materials, to set up and run the experiments, and to perform characterisation and analysis. The emergence of high-throughput experimentation technology (Potyrailo et al., 2011; Coley et al., 2020a,b) can greatly reduce the time needed by making it possible to conduct a large number of experiments in parallel, but it still requires the availability of materials and specialised equipment that may not be easily accessed. For certain classes of reactions involving highly energetic or toxic compounds, the associated safety and health risks need to be minimised (Cao et al., 2020; Erny et al., 2020), even when high-throughput techniques are available.

One approach to reduce the experimental effort is to resort to computer simulation. With the development of advanced modelling techniques for the calculation of liquid-phase rate constants such as density functional theory (DFT) (Jalan et al., 2013; Diamanti et al., 2021) along with appropriate solvation models (Miertus et al., 1981; Miertus and Tomasi, 1982; Marenich et al., 2009) and reactive molecular dynamics (Meuwly, 2019), *in-silico* "experiments" can be performed with more readily available resources. However, the accuracy and reliability of such models are also affected by a number of factors and cannot always be guaranteed (Harvey et al., 2019). Furthermore, while more sophisticated computational methods may provide better accuracy for certain systems, they usually incur large computational expense which requires many CPUs or even GPUs to work in parallel, with large associated operating and energy costs. These complex calculations can be prone to failure, requiring human intervention or running the risk of missing promising solvents. Simply transferring experimental trial-and-error approaches to a computer and applying a brute-force method to identify good reaction solvents offers limited scope.

In view of this, computer-aided molecular design (CAMD) has emerged as a novel computational tool for efficient solvent selection and design by taking advantage of powerful modelling and optimisation techniques. CAMD has been successfully applied to solvent selection and design for various applications, such as crystallisation (Karunanithi et al., 2006; Watson et al., 2021), liquid–liquid extraction (Scheffczyk et al., 2017) and reactions (Folić et al., 2008; Struebing et al., 2013; Zhou et al., 2015b; Austin et al., 2016; Gertig et al., 2019; Liu et al., 2019; Gertig et al., 2020). When CAMD is applied to solvent design for chemical reactions, a kinetic metric, such as rate constant or selectivity, is often selected as the objective function to be optimised. Folić et al. (2008) formulated a CAMD problem to maximise the rate constant of a Menschutkin reaction. In their formulation, rate constants were predicted from several solvent descriptors (Abraham's hydrogen bond acidity, Abraham's hydrogen bond basicity, polarity/dipolarisability (Abraham, 1993), Hildebrand's solubility parameter and a correction parameter denoting whether a solvent molecule is aromatic and/or halogenated) using a linear free-energy relationship (solvatochromic equation) (Abraham et al., 1987a,b) regressed to a small set of experimental rate constants in different solvents. The multivariate linear regression (MLR) formalism of the solvatochromic equation facilitates its incorporation into a CAMD framework. Struebing et al. (2013) built on the work of Folić et al. (2008) and developed a quantum-mechanical computer-aided molecular design (QM-CAMD) method in which the experimental rate constants for model training are replaced with computed rate constants that are calculated via DFT combined with a continuum solvation model. In the work of Folić et al. (2008) and Struebing et al. (2013), the solvatochromic descriptors can be calculated using group contribution (GC) methods (Sheldon et al., 2005; Folić et al., 2008).

In the same spirit, Zhou et al. (2015a,b) constructed another type of quantitative structure–activity relationship (QSPR) model that uses a set of quantum mechanical descriptors derived from so-called σ-profiles (Klamt, 1995) generated by the conductor-like screening model (COSMO) approach, a type of continuum solvation model in which an infinite dielectric is used; GC methods were also developed to calculate the σ-profile-based descriptors. Similar to the solvatochromic equation, the σ-profile-based QSPR model was also regressed via MLR using a small number of experimental rate constants and incorporated into a mixed-integer nonlinear programming (MINLP) formulation where the rate constant is maximised. The approach was applied to a Diels–Alder reaction. As an alternative to the QSPR model of Zhou et al. (2015a,b), Liu et al. (2019) identified a set of solvent descriptors from thermodynamic derivations within conventional transition state theory (CSTS) in combination with additional knowledge-based solvent descriptors which can be calculated using GC methods. Austin et al. (2017, 2018) used COSMO-RS, a COSMO post-processing method that estimates mixture thermodynamics from σ-profiles and cavity volumes of species of interest, so as to calculate rate constants in solvent mixtures. In their work, σ-profiles and cavity volumes were also calculated using GC methods to circumvent the direct evaluation using expensive QM methods. They decomposed the mixture design problem into (1) single-molecule design problems solved by a derivative-free optimisation algorithm over a lower-dimensional space of so-called σ-moments which are also calculated using GC methods and (2) a simplified mixture-design problem for optimal mixture composition with molecular identities fixed to be the best molecules generated by the single-molecule design problems. Gertig et al. (2019) proposed another COSMO-based solvent design method that does not use a surrogate model but may incur higher computational expense to explore a large design space.

Most CAMD approaches directed at reaction solvents thus fit within a general solvent design framework that utilises a surrogate model to replace more expensive experimental or computational evaluations of reaction rate constants. The solvent descriptors used in these surrogate models are often linked to molecular structure through GC methods. It is beneficial to use such a model-based approach since the discrete design space of solvent molecular structures can be projected onto the continuous (latent) space of solvent descriptors. In addition, the simple MLR formalism of the surrogate models facilitates the use of optimisation algorithms to solve the CAMD problem efficiently. Bayesian optimisation, in which Gaussian processes are often adopted as a type of non-parametric surrogate models, has also attracted attention for optimising chemical properties of molecules/materials in recent years as it can guide the sampling process and quantify prediction uncertainty (Pollice et al., 2021; Aldeghi and Coley, 2022; Wang and Dowling, 2022). Beyond surrogate models, the use of chemometric techniques, such as principal component analysis (PCA) and partial least squares regressions (PLS) (Wold, 1995), and other machine learning methods, such as artificial neural network (ANN) and its derivatives, have also been gaining popularity for the prediction of properties including rate constants (Komp and Valleau, 2020; Lu et al., 2021; Komp et al., 2022). However, these chemometric and machine learning methods often require large amount of training data to achieve good accuracy.

In a data-poor context, it becomes important to choose an optimal set of conditions at which a limited number of experimental or computational training data are collected such that the performance of the surrogate model obtained is maximised based on some metric. The relevant experimental conditions in the context of CAMD for reaction solvents include the identities of the initial solvents used to regress the surrogate model, assuming that the reaction temperature is fixed. In this setting, model-based design of experiments (MBDoE), in which a statistical criterion that represents the information content of an initial solvent set is maximised, can be a useful tool. Wicaksono et al. (2014) used the condition number criterion to maximise the diversity of an initial set of solvents in which the experimental rate constants of the solvolysis reaction of *tert*-butyl chloride were obtained.

To facilitate their study, they considered a large set of solvents for which experimental rate constants had been reported in the literature. These experimental rate constants were used to train a solvatochromic equation in order to predict the rate constants of the solvolysis reaction for computer-aided solvent screening. A similar approach using the condition number criterion was taken by Tsichla et al. (2019) to build the solvatochromic equation for predicting the rate constants of the amination reaction of ethyl trichloroacetate with liquefied ammonia. Through experimental verification, it was found that despite some mismatch between the experimental and predicted solvent rankings, the predicted best solvent by the solvatochromic equation indeed maximises the rate constant of the amination reaction. Oliyide (2014) investigated the condition number criterion as well as two other statistical criteria: the A-optimality criterion and the D-optimality criterion. She found that the D-optimality criterion consistently led to the best-performing solvatochromic equation for the solvolysis reaction of tert-butyl chloride and the Menschutkin reaction of tripropylamine and methyl iodide.

Recently, we reported an enhanced version of the QM-CAMD method of Struebing et al. (2013) which we called DoE-QM-CAMD (Gui et al., 2022). In the QM-CAMD approach and in the experimental analogue (Folić et al., 2007, 2008; Grant et al., 2018), an initial set of 6–8 solvents was selected to construct a dataset that was used to build the first surrogate model, which was then refined iteratively through the acquisition of additional data. The initial solvent set was hitherto selected by chemical intuition (Struebing et al., 2017). In our more recent work (Gui et al., 2022), the initial solvents used for model regression were chosen such that the D-optimality criterion value (John and Draper, 1975), which measures the information content of an experimental design, is maximised. The solvatochromic equation resulting from the MBDoE-selected solvents leads to improved performance over that from the solvents selected by chemical intuition in the original QM-CAMD work (Struebing et al., 2013).

In the current paper, we expand the results of our previous work on DoE-QM-CAMD for reaction solvent design (Gui et al., 2022). We explore two formulations of the MBDoE problem which differ in the way that the solvent design space is constructed, one using a predefined list of solvents for which experimental values of the (latent) properties are available and the other using an extended solvent list with both predefined molecules and additional molecules represented by atom groups so that the (solvent) properties can be predicted by GC methods. The two formulations result in different sizes of the design space and thus different D-optimality criterion values. The more appropriate formulation is then chosen for the generation of the initial solvent set (denoted as the “MBDoE set”) to be used in the DoE-QM-CAMD framework. Three case studies are investigated (Fig. 1). In the first case study, solvents are designed to accelerate the Menschutkin reaction of phenacyl bromide and pyridine. Menschutkin reactions are a classic type of S_N2 reaction that have been used many times as a solvent design case study (Folić et al., 2008; Struebing et al., 2013; Austin et al., 2018; Liu et al., 2019; Gertig et al., 2019) due to their sensitivity to solvent effects. The second case study focuses on the reagent combination of ethyl cyano(hydroxyimino)acetate (Oxyma) and diisopropylcarbodiimide (DIC) for amino acid activation in peptide synthesis, which has recently been shown to generate harmful HCN via a side reaction (McFarland et al., 2019). Erny et al. (2020) have shown that solvent effects can affect the amount of HCN produced but only a small number of solvents/solvent mixtures has been tested to date. The third case study involves the Williamson ether synthesis reaction (O-alkylation) of sodium β-naphthoxide and benzyl bromide and its competition with a side reaction (C-alkylation) (Diamanti et al., 2021). The selectivity of the Williamson ether synthesis reaction changes drastically in different solvents. In this case study, the rate constant of the O-alkylation is maximised and the rate constant of the C-alkylation is minimised.

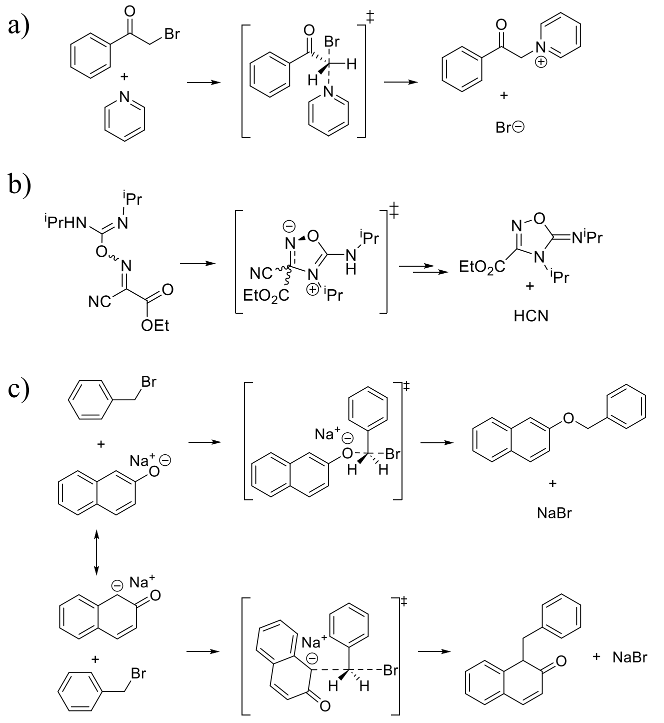

Fig. 1. The case studies in this work, (a) the Menschutkin reaction of phenacyl bromide with pyridine, (b) The cyclisation reaction of the Oxyma/DIC adduct, (c) the O-alkylation (the Williamson ether synthesis reaction) and the C-alkylation of sodium β-naphthoxide with benzyl bromide.

We calculate the rate constants of the reactions using different quantum mechanical levels of theory and the SMD model (Marenich et al., 2009) and use the calculated values as the training data to regress a solvatochromic equation. The performance of the solvatochromic equations generated with MBDoE set of solvents is compared to that of the solvatochromic equations generated with solvents selected by chemical intuition in the original work of QM-CAMD of Struebing et al. (2013), denoted as the “Div” set, i.e., the MBDoE approach is benchmarked against an expert chemist’s intuition. Finally, the solvatochromic equations are incorporated into a mixed-integer linear programming (MILP) problem to optimise the reaction kinetics in each case study (with one or two objectives as appropriate). The results of the CAMD problem obtained at each iteration when using the MBDoE set and the Div set are compared.

The remainder of this paper is organised as follows: in Section 2, the components of the methodology are introduced including the principles of the MBDoE technique using the D-optimality criterion, the formulations of the MBDoE problem, the approach used to calculate liquid-phase rate constants, the formulation of the CAMD problem and the workflow of the DoE-QM-CAMD method. Results are presented and discussed in Section 3, wherein the two formulations of the MBDoE problem are compared, the regression and validation of the solvatochromic equations are presented and the application of the DoE-QM-CAMD method to the three case studies is explored. In Section 4, the conclusions are summarised and perspectives for the future work are discussed.

2. Methodology

The target problem that we aim to address in our work is defined as follows: “Given a selection space of l solvents in which computer experiments can be performed, identify an information-rich set of p experiments, i.e., solvents, such that a multivariate linear regression model regressed from rate constants calculated in the p solvents is more likely to show good performance”. Due to the discrete nature of the selection space, there are C_p^l possible combinations of design choices, a number that can become unmanageable when the selection space is large. To overcome this, we project the discrete space of solvents into the continuous (latent) space of solvent properties. This projection makes it possible to use the D-optimality criterion value to quantify the information content of a selected set of solvents. We nevertheless retain the chemical identity of the solvents through mixed-integer constraints to ensure that only combinations of latent variables that correspond to allowable solvent molecules are considered. Thus, the problem of computer experiment design can be formulated as an MINLP problem solved using a state-of-the-art optimisation solver.

2.1. Model-based design of experiments

2.1.1. D-optimality criterion

Consider a set of q descriptors, one output variable that has been measured in p experiments, and an associated multivariate linear regression (MLR) model,

$$\boldsymbol{Y} = \boldsymbol{F}^{*}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y}$ is a p-dimensional vector of response variables at each measurement, $\boldsymbol{F}^{*}$ is a $p \times q$ matrix in which all elements in the first column, $F^{*}_{i,1}$, $i = 1, \ldots, p$, are equal to 1 and the element in the ith row and the jth column ($j \geq 2$), $F^{*}_{i,j}$, is equal to the value of the $(j-1)$th descriptor at the ith measurement, $\boldsymbol{\beta}$ is a q-dimensional vector of coefficients that need to be estimated, and $\boldsymbol{\epsilon}$ is a p-dimensional vector of random errors. Given the p measurements of the response variables Y_i, $i = 1, \ldots, p$, and the descriptor values at every measurement, $F^{*}_{i,j}$, $i = 1, \ldots, p; j = 2, \ldots, q$, the least-square estimator of the coefficients $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^{*T}\boldsymbol{F}^{*})^{-1}\boldsymbol{F}^{*T}\boldsymbol{Y} = \mathcal{I}^{-1}\boldsymbol{F}^{*T}\boldsymbol{Y}, \tag{2}$$

where $\mathcal{I} = \boldsymbol{F}^{*T}\boldsymbol{F}^{*}$ is defined as the Fisher information matrix (Atkinson et al., 2007). The variance–covariance matrix associated with the coefficients $\hat{\boldsymbol{\beta}}$ is

$$\mathbf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^{2}\mathcal{I}^{-1}, \tag{3}$$

where σ^2 is variance of the random errors $\boldsymbol{\epsilon}$. The D-optimality criterion minimises $\det \mathcal{I}^{-1}$ or equivalently maximises $\det \mathcal{I}$. The determinant of the information matrix can be interpreted geometrically as the volume of the ellipsoid describing the joint confidence region of the estimated coefficients $\hat{\boldsymbol{\beta}}$. It should be noted that in our work, the measured responses are reaction rate constants evaluated via QM calculations. The random errors $\boldsymbol{\epsilon}$ can be viewed as the inherent errors caused by the uncertainties in the GC methods that are employed and the inadequacy of the solvatochromic equation for the prediction of rate constants. Although these errors may violate the assumption of normally distributed random errors, the D-optimality criterion can nevertheless be used to determine an experimental design, and as we will show, this increases the likelihood of obtaining reliable models.

2.1.2. MINLP formulation of the MBDoE problem

The MBDoE problem for the selection of an initial solvent set is formulated as an MINLP problem. The formulation is introduced in this section. Two formulations are presented corresponding to different solvent design spaces: a list of candidate solvents with known experimental property values, and an extended list with both experimental and GC-predicted property values for the solvents, i.e., in which solvent molecules are represented by atom groups.

Formulation 1: list of candidate solvents. First, the user provides a list of l solvents and their associated descriptors, which are stored in the $l \times q$ matrix $\boldsymbol{F}$. In each row k of $\boldsymbol{F}$, the elements in columns 2 to q are the descriptor values of the kth candidate solvent and column 1 is the identity vector. The model matrix $\boldsymbol{F}^{*}$ is then constructed by selecting p rows from matrix $\boldsymbol{F}$. Element (i, j) of the model matrix is given by:

$$F^{*}_{i,j} = \sum_{k=1}^{l} z_{i,k}F_{k,j}, \quad i = 1, \ldots, p, j = 2, \ldots, q, \tag{4}$$

where $z_{i,k}$ is a binary variable denoting whether candidate solvent k is selected as experiment i ($z_{i,k} = 1$) or not ($z_{i,k} = 0$). To express the model matrix more explicitly,

$$\boldsymbol{F}^{*} = \begin{bmatrix} 1 & A_1 & B_1 & S_1 & \delta_1 & \delta^2_{H,1} \\ 1 & A_2 & B_2 & S_2 & \delta_2 & \delta^2_{H,2} \\ 1 & A_3 & B_3 & S_3 & \delta_3 & \delta^2_{H,3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & A_p & B_p & S_p & \delta_p & \delta^2_{H,p} \end{bmatrix}, \tag{5}$$

where $q = 6$ and A_i, B_i, S_i, $\delta^2_{H,i}$ are Abraham's hydrogen bond acidity, Abraham's hydrogen bond basicity, dipolarity/dipolarisability (Abraham, 1993) and the squared Hildebrand solubility parameter of the selected initial solvent i, respectively. δ_i is an additional correction parameter denoting whether the solvent molecule i is aromatic ($\delta_i = 1$), halogenated aliphatic ($\delta_i = 0.5$) or neither ($\delta_i = 0$). Then, the information matrix $\mathcal{I}$ can be calculated as

$$\mathcal{I}_{j,j^*} = \sum_{i=1}^{p} F^{*}_{i,j}F^{*}_{i,j^*}, \qquad j = 1, \ldots, q; j^* = 1, \ldots, q. \tag{6}$$

The D-optimality criterion requires the calculation of the determinant of the information matrix $\mathcal{I}$ which is difficult to formulate directly. Instead, the LDL decomposition of the information matrix $\mathcal{I} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ is used. $\boldsymbol{L}$ is a lower unit triangular matrix and $\boldsymbol{D}$ is a diagonal matrix. $\boldsymbol{L}$ and $\boldsymbol{D}$ can be calculated as below (Watkins, 1991):

$$\begin{aligned} D_j &= \mathcal{I}_{j,j} - \sum_{j^*=1}^{j^*<j} L^2_{j,j^*}D_{j^*}, \qquad j = 1, \ldots, q, \\ L_{j,j^*}D_{j^*} &= \mathcal{I}_{j,j^*} - \sum_{j^{**}=1}^{j^{**}<j^*} L_{j,j^{**}}L_{j^*,j^{**}}D_{j^{**}}, \qquad \text{for } j > j^*; \\ &\quad j = 2, \ldots, q; \;\; j^* = 1, \ldots, q-1. \end{aligned} \tag{7}$$

Then the determinant of the information matrix $\mathcal{I}$, as well as the objective function of the MINLP problem, OF, can be expressed as

$$\mathrm{OF} = \det \mathcal{I} = \prod_{j=1}^{q} D_j. \tag{8}$$

Several logical constraints are required to complete the exposition of the MINLP problem and are presented in the following. One solvent vacancy in the model matrix can be taken up by one candidate solvent only,

$$\sum_{k=1}^{l} z_{i,k} = 1, \qquad i = 1, \ldots, p. \tag{9}$$

Each candidate solvent can be selected at most once,

$$\sum_{i=1}^{p} z_{i,k} \leq 1, \qquad k = 1, \ldots, l. \tag{10}$$

Finally, the initial solvents should be selected in the same order as they are (arbitrarily) arranged in the candidate solvent list to avoid degeneracy:

$$z_{i,k} + z_{i^*,k^*} \leq 1, \qquad \forall i < i^*, \forall k > k^*, i^* = 1, \ldots, p; k^* = 1, \ldots, l-1. \tag{11}$$

Eqs. (4)–(11) complete the formulation of the MBDoE problem where solvents are selected from a predefined candidate list with given experimental property values (Formulation 1).

Formulation 2: an extended list of candidate solvents with both experimental and GC-predicted property values. Formulation 2 is constructed by incorporating additional solvents with missing solvent descriptor values to extend the predefined list in Formulation 1 as shown by the workflow in Fig. 2. These missing solvent descriptor values are predicted by GC methods based on fragmenting molecules into atom groups. A set G of 46 groups is considered in our work, which are appropriately partitioned into subsets (e.g., the set G_A of aromatic groups). The list of atom groups in set G as well as the lists of atom groups in the

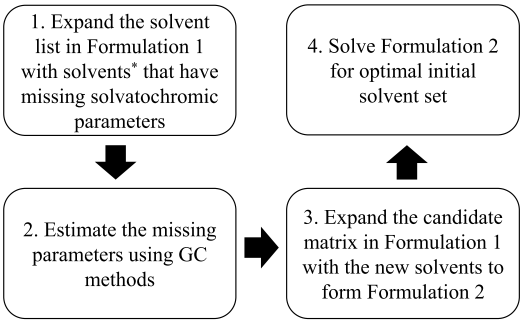
1. Expand the solvent list in Formulation 1 with solvents* that have missing solvatochromic parameters
2. Estimate the missing parameters using GC methods
3. Expand the candidate matrix in Formulation 1 with the new solvents to form Formulation 2
4. Solve Formulation 2 for optimal initial solvent set

Fig. 2. Workflow for developing Formulation 2. *The solvent list in Formulation 2 is that used in Folić et al. (2007) in which not all the experimental property values are available. Therefore, those solvents with complete property values constitute the solvent list in Formulation 1, and the solvents with missing property values are used as the additional solvents in Formulation 2.

subsets of set G defined in the remainder of the paper are provided in the corresponding GAMS code in the Zenodo open repository (see Data availability).

Firstly, Abraham's hydrogen bond acidity (Abraham, 1993) for solvent i defined by vector $n_{i,g}$ of group occurrences can be calculated as below (Folić et al., 2007):

$$A_i = \begin{cases} 0.010641 + \sum_{g\in G} n_{i,g} a_g & \text{if } 0.010641 + \sum_{g\in G} n_{i,g} a_g > 0.029 \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where $n_{i,g}$ denotes the number of atom group g present in molecule i, g belongs to the set G and a_g is the contribution of atom group g to hydrogen bond acidity. Hydrogen bond basicity (Abraham, 1993) can be similarly calculated as below (Folić et al., 2007):

$$B_i = \begin{cases} 0.12371 + \sum_{g\in G} n_{i,g} b_g & \text{if } 0.12371 + \sum_{g\in G} n_{i,g} b_g > 0.124 \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where b_g is the contribution of atom group g to hydrogen bond basicity. The dipolarity/polarisability (Abraham, 1993) of initial solvent i is determined by Folić et al. (2007):

$$S_i = 0.325675 + \sum_{g\in G} n_{i,g} s_g, \tag{14}$$

where s_g is the contribution of atom group g to dipolarity/polarisability. The Hildebrand solubility parameter $\delta^2_{H,i}$ (Hildebrand, 1929) of solvent i can be related to the enthalpy of vaporisation of solvent i at 298 K in kJ mol^{-1}, $\Delta H_{\upsilon,i}$, and the liquid molar volume of solvent i at 298 K in m^3 kmol^{-1}, $V_{m,i}$, using the following equation (Sheldon et al., 2005):

$$\delta^2_{H,i} = 0.238846 \frac{\Delta H_{\upsilon,i} - 10^{-3} RT}{V_{m,i}}, \tag{15}$$

where R is the ideal gas constant in J mol^{-1} K^{-1} and T is the temperature in K. $\Delta H_{\upsilon,i}$ and $V_{m,i}$ can be calculated using the group contribution methods below (Hukkerikar et al., 2012b),

$$\Delta H_{\upsilon,i} = \sum_{g\in G} n_{i,g} h_{\upsilon,g} + 10.4327, \tag{16}$$

$$\Delta V_{m,i} = \sum_{g\in G} n_{i,g} \upsilon_{m,g} + 0.0123, \tag{17}$$

where $h_{\upsilon,g}$ and $\upsilon_{m,g}$ are the contributions to the enthalpy of vaporisation and the liquid molar volume of group g, respectively.

These structure–property relationships are applied to calculate all the missing values to complete the expanded candidate matrix $\boldsymbol{F}$ in the prepossessing stage instead of being activated as constraints in the process of optimisation. Therefore, Formulation 2 has the same level of complexity as Formulation 1 but the size of the design space in Formulation 2 is larger. In addition, design constraints on the melting points and boiling points of solvents can be included in Formulation 2. A dimensionless melting point of solvent i, $T_{me,i}$, can be calculated by Hukkerikar et al. (2012b)

$$T_{me,i} = \exp\left(\frac{T_{m,i}}{T_{m,0}}\right) = \sum_{g\in G} n_{i,g} t_{me,g}, \tag{18}$$

where $T_{m,i}$ is the melting point of solvent i in K, $T_{m,0} = 144.0977$ K is a constant and $t_{me,g}$ is the contribution of group g to the dimensionless melting point. Similarly, a dimensionless boiling point, $T_{be,i}$, is defined as (Hukkerikar et al., 2012b)

$$T_{be,i} = \exp\left(\frac{T_{b,i}}{T_{b,0}}\right) = \sum_{g\in G} n_{i,g} t_{be,g}, \tag{19}$$

where $T_{b,i}$ is the boiling point in K, $T_{b,0} = 244.7889$ K is a constant and $t_{be,g}$ is the contribution to the dimensionless boiling point. Two associated design constraints are imposed on the melting point and the boiling point of solvent i:

$$T_{m,i} \leq 318.15, \tag{20}$$

$$T_{b,i} \geq 278.15. \tag{21}$$

It should be noted that in this formulation the two bounds are more relaxed than what is needed to ensure that the solvents designed are in the liquid phase at room temperature. This is in recognition of the uncertainties in melting and boiling point predictions of Eqs. (18) and (19) and also allows more flexibility in the operating temperature. The relaxed bounds also result in a larger design space, increasing the chance of achieving greater D-optimality criterion values.

Finally, it is possible to exclude specific groups that may react with reactants or products by removing solvents containing these groups from the matrix $\boldsymbol{F}$.

2.2. Quantum mechanical calculation of the reaction rate constants using the thermodynamic cycle approach

In this section, the quantum mechanical methods used in our work for the generation of rate constant data are introduced. The liquid-phase rate constant $k^{\mathrm{L,QM}}$ can be expressed via transition-state theory (Eyring, 1935; Laidler and King, 1983) as:

$$k^{\mathrm{L,QM}} = \kappa \frac{k_B T}{h} \left(c^{\circ,\mathrm{L}}\right)^{1-\sum_{r\in D} \upsilon_r} \exp\left(-\frac{\Delta^{\neq} G^{\circ,\mathrm{L}}}{RT}\right), \tag{22}$$

where $\Delta^{\neq} G^{\circ,\mathrm{L}}$ is the liquid-phase activation Gibbs free energy, κ is the transmission coefficient for which the Wigner tunnelling correction factor (Wigner, 1937) is used, k_B is the Boltzmann constant, T is the temperature (298.15 K in our work), h is the Planck constant, $c^{\circ,\mathrm{L}}$ is the molar concentration at the standard state, D is the set of reactant(s) and υ_r is the stoichiometric coefficient of reactant r. We note that due to the exponential dependence of the rate constant on the activation free energy, k^{L} is very sensitive to small errors in the free energies. In our work, the solution environment is simulated by the SMD continuum solvation model (Marenich et al., 2009). The mean unsigned errors in solvation free energies of the SMD model were reported by Marenich et al. (2009) to be 2.5–4.2 kJ mol^{-1} for neutral species and 16.7 kJ mol^{-1} for ionic species on average in terms of reproducing experimental solvation energies. Moreover, k^{L} can vary by several orders of magnitude from solvent to solvent. For these reasons, $\log k^{\mathrm{L}}$ or $\ln k^{\mathrm{L}}$ (used in our work) is normally considered in developing data-driven models and in comparisons with experimental data. The calculation of liquid-phase activation free energy $\Delta^{\neq} G^{\circ,\mathrm{L}}$ is detailed for each case study in the remainder of this section.

2.2.1. The Menschutkin reaction and the HCN reaction

For the Menschutkin reaction and the HCN reaction, the liquid-phase activation Gibbs free energy $\Delta^{\neq}G^{\circ,\mathrm{L}}$ for the conversion from the reactant(s) to the transition state is calculated using the thermodynamic cycle (TC) method (Ho and Ertem, 2016),

$$\Delta^{\neq}G^{\circ,\mathrm{L}} = \Delta^{\neq}G^{\circ,\mathrm{IG}} + \Delta G_{\mathrm{TS}}^{\circ,\mathrm{solv}} + \sum_{r\in D} \upsilon_r \Delta G_r^{\circ,\mathrm{solv}} + (1 + \sum_{r\in D} \upsilon_r) RT \ln \frac{RT}{P_0}, \quad (23)$$

where $\Delta^{\neq}G^{\circ,\mathrm{IG}}$ is the ideal gas activation Gibbs free energy, $\Delta G_{\mathrm{TS}}^{\circ,\mathrm{solv}}$ is the solvation free energy of the transition state, $\Delta G_r^{\circ,\mathrm{solv}}$ is the solvation free energy of reactant r and P_0 is the reference pressure. The last term is the standard-state correction from the gas-phase standard state defined by $T = 298.15$ K and $P_0 = 1$ atm to the solution-phase standard state of 1 mol L^{-1}. $\Delta^{\neq}G^{\circ,\mathrm{IG}}$ can be calculated as follows

$$\Delta^{\neq}G^{\circ,\mathrm{IG}} = G_{\mathrm{TS}}^{\circ,\mathrm{IG}} + \sum_{r\in D} \upsilon_r G_r^{\circ,\mathrm{IG}} = (E_{\mathrm{TS}}^{\mathrm{el,IG}} + G_{\mathrm{TS}}^{\mathrm{therm,IG}}) + \sum_{r\in D} \upsilon_r (E_r^{\mathrm{el,IG}} + G_r^{\mathrm{therm,IG}}), \quad (24)$$

where $G_{\mathrm{TS}}^{\circ,\mathrm{IG}}$ and $G_r^{\circ,\mathrm{IG}}$ are the ideal gas Gibbs free energies of the transition state and reactant r, respectively, $E_{\mathrm{TS}}^{\mathrm{el,IG}}$ and $E_r^{\mathrm{el,IG}}$ are the single-point energies of the transition state and reactant r, respectively, calculated using the G3MP2 method (Curtiss et al., 1999), and $G_{\mathrm{TS}}^{\mathrm{therm,IG}}$ and $G_r^{\mathrm{therm,IG}}$ are the thermal corrections to the free energies of the transition state and species r, respectively. In the calculation of $\Delta^{\neq}G^{\circ,\mathrm{IG}}$, the structures of the transition state and reactant r are optimised at the M062X/6-31+G(d) level of theory (Zhao and Truhlar, 2008) in the ideal gas phase using the Gaussian 16 software (Frisch et al., 2016). After each calculation, we verify that each of the reactants has no imaginary frequency, and the transition state has only one imaginary frequency. The same verification is applied to all liquid-phase optimised structures in this work.

The solvation energy $\Delta G_r^{\circ,\mathrm{solv}}$ is calculated by applying the SMD solvation model (Marenich et al., 2009) in the Gaussian 16 software and is given as:

$$\begin{aligned} \Delta G_r^{\circ,\mathrm{solv}} &= E_r^{\mathrm{el,L}} - E_r^{\mathrm{el,IG}}, \quad r \in D, \\ \Delta G_{\mathrm{TS}}^{\circ,\mathrm{solv}} &= E_{\mathrm{TS}}^{\mathrm{el,L}} - E_{\mathrm{TS}}^{\mathrm{el,IG}}. \end{aligned} \quad (25)$$

where $E_r^{\mathrm{el,L}}$ and $E_r^{\mathrm{el,IG}}$ are the single-point energies of reactant r calculated using M062X/6-31+G(d) in the liquid phase (with the SMD solvation model) and in the ideal gas phase, respectively, and $E_{\mathrm{TS}}^{\mathrm{el,L}}$ and $E_{\mathrm{TS}}^{\mathrm{el,IG}}$ are the single-point energies of the transition state calculated at the same level of theory in the liquid phase and in the ideal gas phase, respectively. The liquid-phase energies, $E_r^{\mathrm{el,L}}$ and $E_{\mathrm{TS}}^{\mathrm{el,L}}$, are calculated at geometries optimised in the liquid phase.

In calculating the transition states for the two reactions, we use the S$_\mathrm{N}$2 mechanism with the transition state structure reported by Struebing et al. (2013) for the Menschutkin reaction. For the HCN reaction, we use the mechanism recently elucidated by Gui et al. (2023) and the corresponding transition state.

2.2.2. The Williamson ether synthesis reaction

For the study of the Williamson ether synthesis reaction, the liquid-phase activation Gibbs free energy is directly calculated within the SMD solvation model at one level of theory, B3LYP/6-31+G(d), as this approach was found to perform well for this specific reaction (Diamanti et al., 2021). The liquid-phase activation free energy, $\Delta^{\neq}G_{\mathrm{direct}}^{\circ,\mathrm{L}}$, calculated by this direct method, without recourse to a thermodynamic cycle, can be expressed as:

$$\Delta^{\neq}G_{\mathrm{direct}}^{\circ,\mathrm{L}} = G_{\mathrm{TS}}^{\circ,\mathrm{L}} + \sum_{r\in D} \upsilon_r G_r^{\circ,\mathrm{L}} = (E_{\mathrm{TS}}^{\mathrm{el,L}} + G_{\mathrm{TS}}^{\mathrm{therm,L}}) + \sum_{r\in D} \upsilon_r (E_r^{\mathrm{el,L}} + G_r^{\mathrm{therm,L}}), \quad (26)$$

where $G_{\mathrm{TS}}^{\circ,\mathrm{L}}$ and $G_r^{\circ,\mathrm{L}}$ are the liquid-phase Gibbs free energies of the transition state and reactant r, respectively, $E_{\mathrm{TS}}^{\mathrm{el,L}}$ and $E_r^{\mathrm{el,L}}$ are the single-point energies of the transition state and reactant r, respectively (calculated using B3LYP/6-31+G(d)), and $G_{\mathrm{TS}}^{\mathrm{therm,L}}$ and $G_r^{\mathrm{therm,L}}$ are the thermal corrections to the free energies of the transition state and species r, respectively, obtained directly in the liquid phase. The structures of the transition state and reactants $r \in D$ are also optimised at the B3LYP/6-31+G(d) level of theory in the liquid phase.

2.3. Formulation of the computer-aided molecular design problem to optimise reaction kinetics

In this section, the MILP formulation of the CAMD problem wherein solvents are constructed from atom groups to design novel solvents to optimise reaction kinetics, is detailed. As in the work of Struebing et al. (2013) and Grant et al. (2018), we introduce single-molecule groups to represent common solvents that are too small to be represented by atom groups (e.g., DMSO) and we set the relevant properties to be equal to measured values. In the MILP formulation, the index i is dropped as only one solvent is designed each time the CAMD problem is solved.

The objective function of the MILP problem to be maximised or minimised, depending on whether the chemical reaction of interest is a main reaction or a side reaction, is the rate constant of the reaction, which is represented by the solvatochromic equation (Abraham et al., 1987a,b),

$$\ln k^{\mathrm{L}} = c_0 + c_A A + c_B B + c_S S + c_\delta \delta + c_H \delta_H^2, \quad (27)$$

where k^{L} is the liquid-phase rate constant of the studied reaction in the designed solvent and c_0, c_A, c_B, c_S, c_δ and c_H are the coefficients that need to be estimated via MLR.

Specifically for the CAMD problem, constraints representing structure–property relationships, solvent properties, chemical feasibility and molecular complexity are taken from Grant et al. (2018). The constraints that are considered in our work are briefly described in this section, although not all auxiliary constraints are shown here for conciseness. The complete formulation of these constraints can be consulted in the work of Grant et al. (2018) and in the GAMS files provided in the Zenodo repository (see Data availability).

Firstly, constraints to calculate solvent properties using group contribution methods are included. We use the same group contribution methods as in the work of Grant et al. (2018) for the calculations of A, B, S, δ_H^2, ΔH_v, ΔV_m, T_{me} and T_{be}, i.e., Eqs. (12)–(19). The associated bounds imposed on the melting point and the boiling point of the designed solvent are tightened in comparison with the bounds in constrained Formulation 2 of the MBDoE problem in order to ensure that the solvents are liquid at room temperature:

$$T_m \le 298.15, \quad (28)$$

$$T_b \ge 323.15. \quad (29)$$

Apart from these design constraints on the melting and boiling point, additional design constraints on the flash point F_p, octanol/water partition coefficient K_{OW} and the oral rat median lethal dose LD$_{50}$ of the designed solvent are added to take into account the factors of health, safety and environmental impact. The flash point of the designed molecule can be calculated by (Hukkerikar et al., 2012a)

$$F_{\mathrm{p}} = \sum_{g\in G} n_g F_{\mathrm{p},g} + 150.0218, \quad (30)$$

where $F_{\mathrm{p},g}$ is the contribution of atom group g to the flash point. It is bounded by

$$F_{\mathrm{p}} \ge 252. \quad (31)$$

The octanol/water partition coefficient K_{OW} is given by (Hukkerikar et al., 2012b)

$$\log K_{\mathrm{OW}} = \sum_{g\in G} n_g K_{\mathrm{OW},g} + 0.752, \quad (32)$$

where $K_{OW,g}$ is the contribution of group g to the logarithm of the octanol/water partition coefficient. It is constrained by

$$\log K_{OW} \geq 3. \tag{33}$$

The oral rat LD_{50} is calculated by (Hukkerikar et al., 2012a),

$$-\log LD_{50} = \sum_{g\in G} n_g LD_{50,g} + 1.9372 + 0.0016M, \tag{34}$$

where $LD_{50,g}$ is the contribution of group g to the negative logarithm of LD_{50}, and M is the molecular weight of the designed molecule. M is simply calculated by

$$M = \sum_{g\in G} n_g m_g, \tag{35}$$

where m_g is the molecular weight of group g. The oral rat LD_{50} is constrained by

$$-\log LD_{50} \leq 3. \tag{36}$$

A set of chemical feasibility constraints is introduced to ensure that the solvent molecules generated are valid chemical structures. Three binary variables, y_{ac}, y_{bi} and y_{mo}, denote whether the designed molecule is acyclic, bicyclic or monocyclic, respectively. Only one of these three binary variables can be 1 for the designed solvent, i.e., only one of the three types of molecules can be designed for one solvent vacancy, and this can be expressed by:

$$y_{ac} + y_{bi} + y_{mo} = 1. \tag{37}$$

Based on these three binary variables above, auxiliary variable m (Odele and Macchietto, 1993) can be defined as

$$m = y_{ac} - y_{bi}, \tag{38}$$

where m is equal to 1 for an acyclic molecule, 0 for a monocyclic molecule and −1 for a bicyclic molecule. The number of aromatic fragments is set to 6 for a monocyclic molecule and 10 for a bicyclic molecule. This is implemented with the equation below:

$$\sum_{g\in G_A} n_g - 6y_{mo} - 10y_{bi} = 0, \tag{39}$$

where G_A is the set of all aromatic atom groups (excluding aromatic single-molecule groups). The octet rule (Odele and Macchietto, 1993) needs to be satisfied to design structurally-feasible molecules. It is given by:

$$\sum_{g\in G} (2 - \upsilon_g) n_g - 2m = 0, \tag{40}$$

where υ_g is the valence of structural group g. The correct bonding between groups and the feasibility of single-molecule groups are ensured using the following bonding rule (Buxton et al., 1999; Struebing et al., 2013):

$$n_g(\upsilon_g - 1) + 2\Big(m - \sum_{g^*\in G_1} n_{g^*}\Big) - \sum_{g^*\in G} n_{g^*} \leq 0, \quad \forall g \in G, \tag{41}$$

where G_1 is the set of single-molecule groups. Also, in the case of solvent design from a single-molecule group, to ensure that for each vacancy, only one molecule is designed, the following constraints are imposed:

$$\sum_{g\in G_1} n_g \leq 1, \tag{42}$$

$$\sum_{g\in G} n_g - \sum_{g^*\in G_1} n_{g^*} \leq \Big(1 - \sum_{g^*\in G_1} n_{g^*}\Big) n_{G,\max}, \quad g^* \in G_1 \tag{43}$$

where $n_{G,\max}$ is the maximum number of atom groups allowed in a molecule. Alternative ways of representing chemical feasibility are available, and the reader is referred to Sahinidis et al. (2003) and Samudra and Sahinidis (2013) for other possible formulations.

Next, molecule complexity constraints are defined to limit the size and the complexity of designed molecules. First, the number of atom groups in each designed molecule should be larger than the minimum number of groups allowed, $n_{G,\min}$ (taken as 1 in this work), and smaller than the maximum number of groups, $n_{G,\max}$ (taken as 7 in this work):

$$\sum_{g\in G} n_g \geq n_{G,\min}, \tag{44}$$

$$\sum_{g\in G} n_g \leq n_{G,\max}. \tag{45}$$

For each atom group g, an upper limit, n_g^U, is also set, based on

$$n_g \leq n_g^U, \quad \forall g \in G. \tag{46}$$

The values of n_g^U can be found in the GAMS code provided. The number of "main" groups, i.e., groups that contain C and H atoms only, is constrained depending on whether the solvent molecule is acyclic or monocyclic:

$$\sum_{g\in G_M} n_g \leq 2y_{mo} + n_{G,\max} y_{ac}, \tag{47}$$

where G_M is the set of main groups. The number of functional groups in the designed molecule, i.e., groups that contain at least one atom other than C or H, is also constrained with

$$\sum_{g\in G_F} \frac{n_g}{n_g^U} \leq y_{mo} + y_{ac}, \tag{48}$$

where G_F is the set of functional groups. In addition, only one double bond is allowed in the designed solvent to ensure the solvent has sufficient chemical stability:

$$\sum_{g\in G_D} n_g \leq 1, \tag{49}$$

where G_D is the set of groups that contain a double bond. For the remaining constraints, new binary variables, y_g, are defined to denote whether atom group g occurs in the designed molecule:

$$y_g = \begin{cases} 1 & \text{if group } g \text{ is present in the designed molecule} \\ 0 & \text{otherwise.} \end{cases} \tag{50}$$

Another binary, y_M, is defined to denote whether the molecule is aromatic and monocyclic:

$$y_M = \begin{cases} 1 & \text{if } y_{aC} + y_{mo} = 2 \\ 0 & \text{otherwise.} \end{cases} \tag{51}$$

Then, the following constraint can be imposed to allow a monocyclic molecule to have side chains but prevent a bicyclic molecule from having any side chains:

$$2y_{bi} + y_M - n_{aC} = 0. \tag{52}$$

To further reduce the molecular complexity, at most one type of side-chain forming aromatic groups (here, aC, aCCH or $aCCH_2$) can appear in a monocyclic molecule:

$$y_M + y_{aCCH} + y_{aCCH_2} \leq 1. \tag{53}$$

For the aCCH group, one of the side chains is constrained to be a CH_3 group:

$$y_{aCCH} \leq n_{CH_3}. \tag{54}$$

Chain-ending groups can occur at most three times in an aliphatic molecule and once in an aromatic molecule:

$$\sum_{g\in G_{CE}} n_g \leq 3y_{ac} + y_M + y_{aCCH} + y_{aCCH_2}, \tag{55}$$

where G_{CE} is the set of chain-ending groups. The number of non-chain-ending groups in the designed molecule is constrained by:

$$\sum_{g\in G_{NCE}} n_g \leq 3y_{ac} + y_M + y_{aCCH_2}, \tag{56}$$

where G_{NCE} is the set of non-chain-ending groups. Eqs. (12)–(19) and Eqs. (27)–(56) complete the MILP formulation of the CAMD problem.

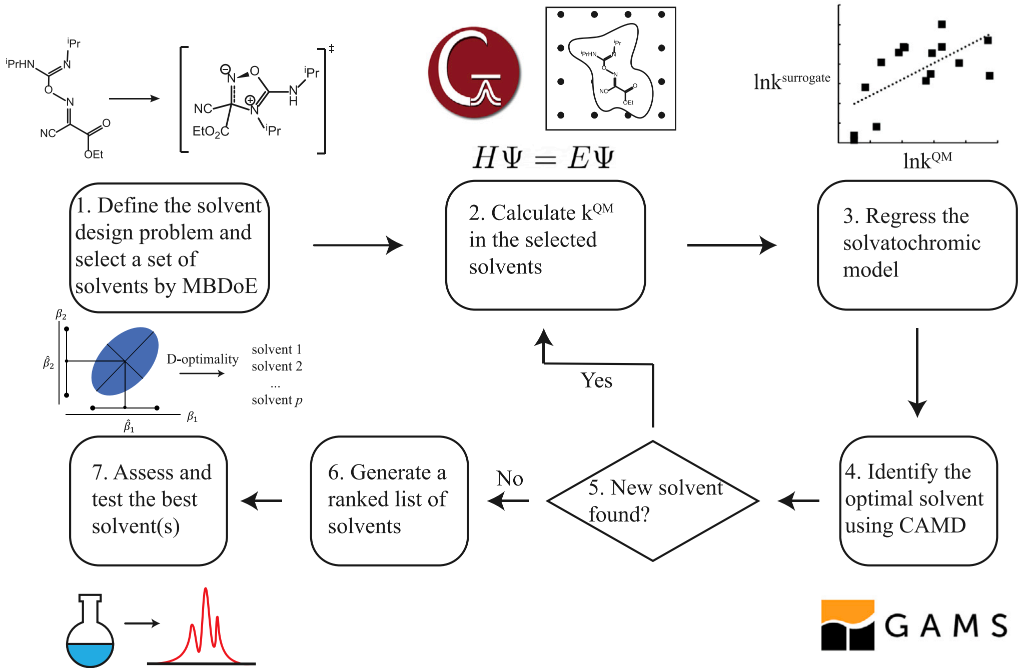
$H\Psi = E\Psi$
lnk^surrogate
lnk^QM
1. Define the solvent design problem and select a set of solvents by MBDoE
2. Calculate k^QM in the selected solvents
3. Regress the solvatochromic model
D-optimality
solvent 1
solvent 2
...
solvent p
Yes
No
7. Assess and test the best solvent(s)
6. Generate a ranked list of solvents
5. New solvent found?
4. Identify the optimal solvent using CAMD
GAMS

Fig. 3. Workflow of the DoE-QM-CAMD method for reaction solvent design.

2.4. *The DoE-QM-CAMD workflow*

The complete workflow for the DoE-QM-CAMD method is shown in Fig. 3. First (step 1), the design problem, including the objective function and all the design constraints, is defined and an initial solvent set is created using the MBDoE technique by solving an MINLP problem, where the determinant of the information matrix (D-optimality criterion) is maximised. Second (step 2), the rate constants of the reaction(s) of interest in the selected initial solvents are computed with the QM method. In the next step (step 3), the computed rate constants are used to regress a surrogate model, the solvatochromic equation. Once the regressed solvatochromic equation is obtained, it is incorporated into an MILP problem (step 4) where the objective function to be optimised is the rate constant of the reaction or other metrics derived from the rate constant. In step 5, the optimal solvent identified is checked against the training set used to build the solvatochromic equation. If it was not previously included in the set, then steps 2–4 are repeated with the optimal solvent added to the training set for the solvatochromic equation. The iterations end when the latest optimal solvent is found to have already been included in the training set. The last CAMD problem with the latest solvatochromic equation is then (step 6) solved repeatedly with integer cuts applied to generate a ranked list of candidate solvents. In the final step, (step 7) the best solvents generated are assessed against criteria not yet considered (e.g., cost or stability) and tested experimentally.

3. Results and discussion

3.1. *Comparing the two MBDoE formulations*

The two MBDoE formulations introduced in Section 2.1.2 are compared by generating two D-optimal designs for the initial solvent set. The solvents are characterised by five descriptors A_i, B_i, S_i, δ and $\delta^2_{H,i}$. In Formulation 1, 240 solvents are considered in the predefined list. In Formulation 2, the design space is extended from 240 solvents to 391 solvents; the missing solvent descriptor values are predicted using the GC methods for the additional 151 solvents, i.e., Formulation 2 uses a mix of experimental values and GC methods. These additional 151 solvents were also in the training data for the GC methods of Sheldon et al. (2005) and Folić et al. (2007), but not all the solvatochromic parameters are available for them. The lists in Formulation 1 and 2 as well as the list of atom groups used in this work can be found in the GAMS code provided in the Zenodo open repository (see Data availability).

These formulations are solved using the DICOPT solver (Kocis and Grossmann, 1989) in General Algebraic Modelling System (GAMS) Release 37.1 (https://www.gams.com/) with multiple randomly generated initial guesses. In Fig. 4, all the feasible solutions obtained from the solver for each formulation are shown in descending order of objective function values (D-optimality criterion). It can be seen that most of the solutions generated are only feasible solutions that are not certified as global optima. The largest objective function (OF = 10.85) is achieved with Formulation 2. For Formulation 1, the best solution gives an objective function of 8.42. It is not surprising that the best possible solution of Formulation 1 cannot exceed that of Formulation 2 as the design space of Formulation 2 is expanded relative to Formulation 1. For the same reason, all the feasible solutions in Formulation 1 are also feasible solutions of Formulation 2. Based on the calculated D-optimality criterion values, Formulation 2 is the better-performing formulation.

To enhance the feasibility of the selected solvents, extra constraints on the melting points and the boiling points of the solvents, Eqs. (20) and (21), are added to Formulation 2 (“constrained Formulation 2”) to ensure that the selected solvents are liquid at room temperature. This is especially important for real experiments but may be also helpful for computational experiments as the solvation model may not work well with solid or gas phase “solvents” even if they are represented as a continuum. Specific groups (see the GAMS code provided in the Zenodo open repository in Data availability) are excluded in order to avoid the selection of amines, carboxylic acids and pyridine and its derivatives due to their potential reactivity with the key species in the case studies. As shown in Fig. 4, adding these constraints greatly lowers the objective function values; the largest objective function achieved is 2.27. However, this is still larger than the D-optimality criterion value of the empirically chosen set (denoted as “Div”) used in Struebing et al. (2013) (OF = 1.62×10^{-7}).

To further evaluate the diversity of the solvents in the D-optimal solvent sets, four radar charts corresponding to four continuous-valued

Table 1
Initial sets of solvents generated by MBDoE and used in Struebing et al. ("Div") and the corresponding rate constants for the Menschutkin reaction, as computed by the QM thermodynamic cycle method. The solvents are listed in order of increasing rate constant in L mol-1 s-1.

MBDoE (constrained Formulation 2)			Div		
No	Solvent	$k^{L,QM}$	No	Solvent	$k^{L,QM}$
1	2,2,4-Trimethylpentane	4.99×10^{-7}	1	Toluene	1.71×10^{-6}
2	1-Phenyl-1-propanol	1.40×10^{-4}	2	Chlorobenzene	3.80×10^{-5}
3	3-Fluorophenol	1.50×10^{-4}	3	Ethyl acetate	4.88×10^{-5}
4	Nitrobenzene	5.15×10^{-4}	4	Tetrahydrofuran	8.77×10^{-5}
5	2-Methoxyethanol	7.06×10^{-4}	5	Acetone	2.99×10^{-4}
6	Adiponitrile	8.50×10^{-4}	6	Acetonitrile	5.02×10^{-4}
7	N-Methylformamide	2.68×10^{-3}	7	Nitromethane	6.50×10^{-4}
D-optimality value		2.27	D-optimality value		1.62×10^{-7}

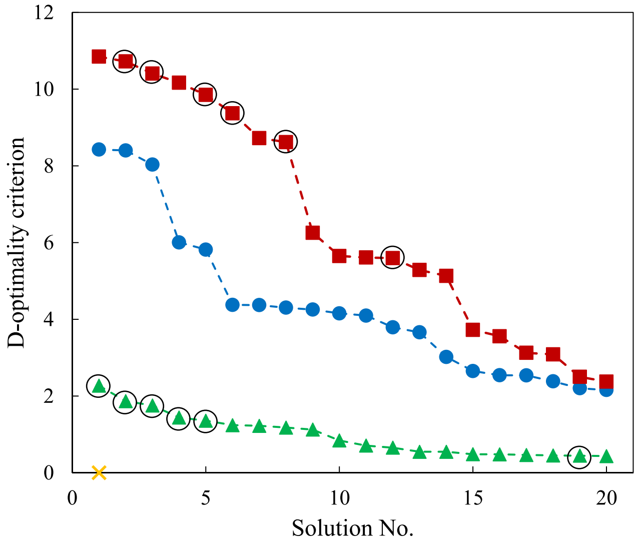

Fig. 4. The D-optimality criterion values of the solvent sets generated by the three MBDoE formulations and the Div set (blue circles: Formulation 1; red squares: Formulation 2; green triangles: constrained Formulation 2; yellow cross: the Div set). The solutions reported to be locally optimal by the solver are circled.

solvent descriptors (A_i, B_i, S_i and $\delta^2_{H,i}$) are plotted for the top D-optimal solvent set (with the highest D-optimality criterion value) generated by each of Formulation 1, Formulation 2 and constrained Formulation 2, as well as for the Div set (Fig. 5).

The Div set is the least diverse among all the solvent sets as the solvent descriptors span relatively small ranges and do not vary much compared to the MBDoE-generated solvent sets. Similarly, it can be seen that the sets generated by the formulations without the design constraints (Formulation 1, Formulation 2) perform better than constrained Formulation 2, which is consistent with the D-optimality criterion values of these sets. The set generated from constrained Formulation 2 generally shows similar variability to those from Formulation 1 and Formulation 2, except that the value ranges they span are slightly smaller. Thus, constrained Formulation 2 is still reasonably good. In this work, constrained Formulation 2 is chosen for the remaining steps in the DoE-QM-CAMD workflow to avoid possible poor performance of the SMD model when applied to non-liquid "solvents".

3.2. Case study 1: Menschutkin reaction - rate constant maximisation

3.2.1. Regression and validation of the solvatochromic equations

The identities of the selected solvents and the computed rate constants for the Menschutkin reaction in these solvents are listed in Table 1 along with those of the Div set for comparison. Two solvatochromic equations are regressed using the computed QM rate constants in the solvents of the MBDoE set (denoted by the superscript "MBDoE") and those in the solvents of the Div set (denoted by the superscript "Div"), respectively:

$$\ln k^{L,MBDoE\text{-}1} = -16.30 - 3.52A + 4.62B + 0.21S + 1.64\delta + 4.26\delta_H^2, \quad (57)$$

$$\ln k^{L,Div\text{-}1} = -18.18 - 12.09A + 5.09B + 12.05S - 0.48\delta - 0.78\delta_H^2. \quad (58)$$

The superscript "1" denotes the iteration number in the QM-CAMD procedure. The MBDoE model shows better statistics (adjusted R^2 = 0.83) than the Div model (adjusted R^2 = 0.16) and covers a wider range of rate constant values, with four orders of magnitude for MBDoE vs two orders of magnitude for Div.

To examine the accuracy of these two solvatochromic equations further, the rate constants of the Menschutkin reaction in the 326 solvents in the solvent design space are computed using the MBDoE model, the Div model and the QM method. With the QM model as the benchmark, the MBDoE model has a mean absolute deviation (MAD) of 1.90 log units, outperforming the Div model, which has a MAD of 3.51 log units (Fig. 6). While model accuracy is low, as can be expected from the small size of the data set used in building the model, the MBDoE-based model provides a better assessment of "good" solvents, that accelerate the Menschutkin reaction, and "poor" solvents, that slow it down.

3.2.2. Solution to the computer-aided molecular design problem

The two solvatochromic equations are incorporated into two CAMD problems, the MBDoE CAMD problem and the Div CAMD problem, which are then solved using the CPLEX solver (CPLEX, IBM ILOG, 2009) in GAMS. The objective function is the maximisation of the rate constant, while the constraints are introduced in Section 2.3. The progress of the QM-CAMD iterations is shown in Table 2. In contrast, at the first iteration of the Div CAMD problem, the rate constant for the optimal solvent identified, Dimethyl sulfoxide, is predicted by the solvatochromic equation to be 3.88×10^2 L mol-1 s-1 which is greater than the QM prediction (1.09×10^{-3} L mol-1 s-1) by five orders of magnitude, a behaviour consistent with that observed in Struebing et al. (2013). At the first iteration of the MBDoE QM-CAMD problem, nitromethanol is identified as the optimal solution with the rate constant predicted to be 2.28×10^{-1} L mol-1 s-1. The prediction is much closer to the QM value (1.56×10^{-3} L mol-1 s-1) with a deviation of two orders of magnitude.

At the second iteration, the optimal solutions of the MBDoE CAMD problem and the Div CAMD problem in the first iteration are added into the respective solvent sets, to update the solvatochromic equations. The following MBDoE model is obtained:

$$\ln k^{L,MBDoE\text{-}2} = -14.57 + 0.21A + 4.61B + 1.53S + 1.16\delta + 1.34\delta_H^2, \quad (59)$$

and the updated Div model is:

$$\ln k^{L,Div\text{-}2} = -8.44 + 9.68A - 5.79B + 5.95S - 4.35\delta - 2.14\delta_H^2. \quad (60)$$

The two CAMD problems are then solved again using these updated models. In this specific case, the MBDoE CAMD problem generates the same optimal solution as in the previous iteration and terminates,

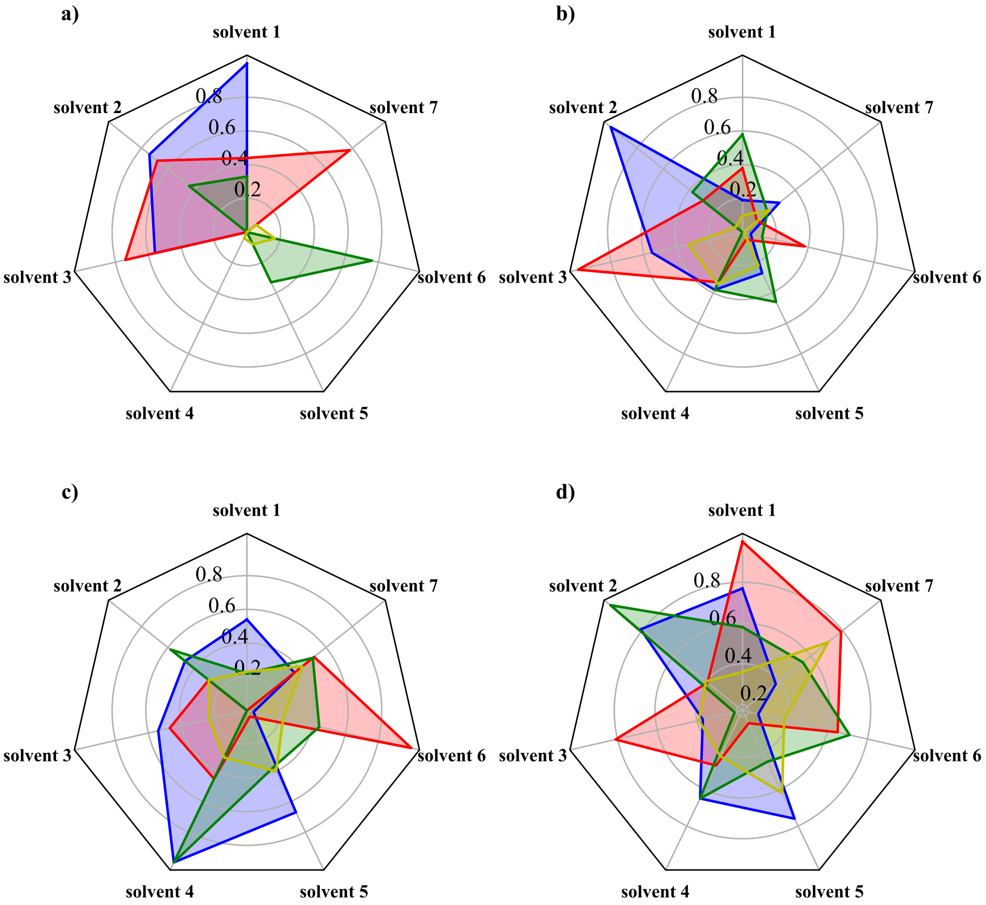

Fig. 5. Radar charts of the four normalised solvent properties for the seven D-optimal solvents identified by each formulation: (a) hydrogen bond acidity, (b) hydrogen bond basicity, (c) dipolarity/polarisability and (d) Hildebrand solubility parameter (blue: Formulation 1, red: Formulation 2, green: Constrained Formulation 2, yellow: the Div set).

while the Div CAMD problem generates a new solution. It takes another four iterations for the Div CAMD problem to reach convergence and generate the same optimal solution, nitromethanol, as the MBDoE CAMD problem. The converged Div model is:

$$\ln k^{\mathrm{L,Div\text{-}5}} = -10.32 + 0.32A - 2.69B + 2.53S - 3.32\delta + 0.84\delta_H^2. \tag{61}$$

It is noted that the quality of the predictions from the Div model has greatly improved. In addition to the best solvents, the top 10 solvents generated by the converged MBDoE (Eq. (59)) and the Div (Eq. (61)) models are presented in Table 3. It can be seen that most of the solvents identified contain nitro group and hydroxyl group which greatly contribute to the polarity of a molecule. These solvents are not typically used in chemical reactions but it is reasonable for them to be identified as solvents that can accelerate the Menschutkin reaction, since as a rule of thumb, a reaction involving a charged transition state proceeds faster in polar protic solvents, as the transition state can be stabilised in such solvents. Despite deviations between the MLR models and QM data, almost all the top 10 solvents of the MBDoE model and the Div model (except 2-methyl-1-nitropropan-2-ol from the MBDoE model) are found to be in the top 10 solvents predicted by the QM model. It should be noted that the top solvent, nitromethanol, in which the nitro group and the hydroxyl group connect to the same carbon atom, is chemically unstable (Winey and Gupta, 1997). Therefore, nitromethanol would be excluded in a further examination of chemical stability. Consequently, dimethyl sulfoxide and N-methylformamide are the best solvents suitable for experimental tests from the MBDoE CAMD problem and the Div CAMD problem, respectively.

3.3. Case study 2: HCN formation minimisation in peptide synthesis

The same procedures are repeated for the HCN formation reaction. Firstly, the solvatochromic equations are regressed using the MBDoE set and the Div set in Table 1 with the rate constants recalculated for the HCN formation reaction. As with the Menschutkin reaction, the MBDoE model shows better statistics (adjusted $R^2 = 0.85$) than the Div model (adjusted $R^2 = 0.26$). The QM rate constants of the HCN formation in eight solvents (1,4-dioxane, chloroform, dichloromethane, benzyl alcohol, methyl isobutyl ketone, ethanol, dimethylformamide and dimethyl sulfoxide) commonly found in chemistry labs are computed and used as a validation set to assess the MBDoE model and the Div model. The model coefficients of the solvatochromic equation and the QM data used for regression and validation can be found in the Excel spread sheet provided in the Zenodo open repository (see Data availability). The MBDoE model yields a MAD of 2.55 log units while the Div model yields a larger MAD of 6.97 log units on the validation set.

In the design problem, the rate constant is *minimised* to suppress HCN formation. In the first iteration of the MBDoE CAMD problem, 2,3,4-trimethyl-2-pentene is identified as the optimal solution with the rate constant predicted to be 3.09×10^{-6} s^{-1} which is close to the QM value of 1.34×10^{-5} s^{-1}, with a deviation of only one order of magnitude. In the Div CAMD problem, the rate constant for the optimal solvent identified, 2,3-dimethylpentane, is predicted by the corresponding solvatochromic equation to be 1.16×10^{-8} s^{-1} which is three orders of magnitude smaller than the QM prediction (2.26×10^{-5} s^{-1}).

In the second iteration, both CAMD problems generate the same optimal solutions as in the respective previous iterations and reach convergence. However, according to the QM predictions, 2,3,4-trimethyl-2-pentene, which is identified by the MBDoE CAMD problem is a better

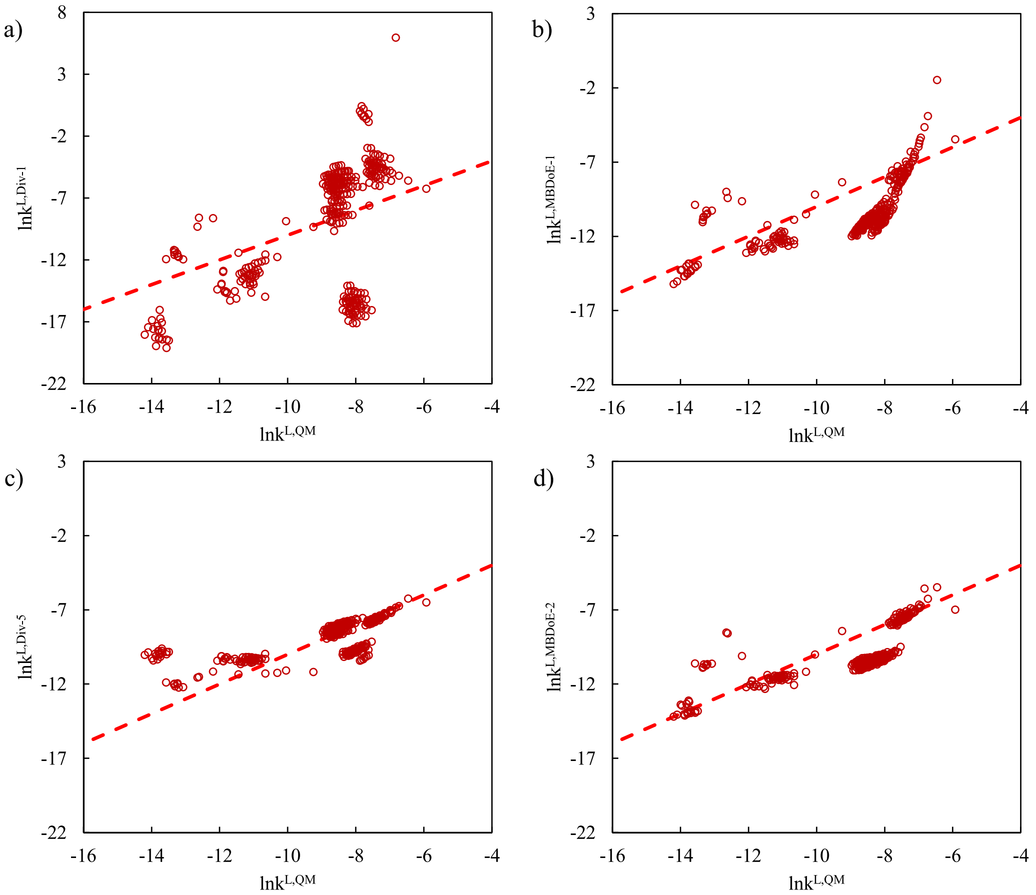

Fig. 6. Parity plots for the logarithm of the rate constant of the Menschutkin reaction in the 326 solvents in the design space (the dashed line is $y = x$ line), (a) Div model Eq. (58) vs. the QM model, (b) the MBDoE model Eq. (57) vs. the QM model, (c) converged Div model Eq. (61) vs. the QM model, (d) converged MBDoE model Eq. (59) vs. the QM model.

solvent compared to 2,3-dimethylpentane identified by the Div CAMD problem as it leads to a smaller rate constant, by nearly one order of magnitude, which enables better suppression of HCN formation. In both CAMD problems, most of the solvents identified are either alkenes or alkanes which are non-polar solvents. These solvents are not typically used in peptide synthesis but it is reasonable for them to be identified as solvents that can suppress HCN formation, since in general, a reaction involving a charged transition state proceeds more slowly in non-polar solvents. The main reactions of peptide synthesis need to be taken into consideration to identify more suitable solvents for suppressing the side reaction as well as supporting amino acid activation and amidation. It is necessary for the reaction solvents not only to promote the kinetics of the main reactions but also to provide adequate solubility of the reacting species.

3.4. *Case study 3: Williamson ether synthesis reaction – investigation of the trade-offs between the main reaction rate and selectivity*

In case study 3, a solvent is designed to accelerate the Williamson ether synthesis reaction, i.e., the O-alkylation of sodium β-naphthoxide and benzyl bromide, as well as to suppress its accompanying side reaction, the C-alkylation reaction. Thus, the trade-offs between two objectives, the reaction rate of the main reaction and reaction selectivity, are investigated. Similar to the previous two case studies, solvatochromic equations are regressed to the computed QM rate constants of the O-alkylation reaction and the C-alkylation reaction. Statistics consistent with the two previous case studies are obtained for the C-alkylation reaction as the MBDoE set leads to a higher adjusted R^2 of 0.89 while the Div set leads to negative adjusted R^2 of −0.40. However, better regression statistics are obtained for the O-alkylation reaction using the Div set. The adjusted R^2 resulting from the MBDoE set and the Div set for the O-alkylation reaction are −0.38 and 0.70, respectively. Despite the poor adjusted R^2, the MBDoE model is still found to have better accuracy for both the O-alkylation (MAD: 1.18 log units) and the C-alkylation (MAD: 0.92 log units) reactions than the Div model (O-alkylation MAD: 2.11 log units, C-alkylation MAD: 2.88 log units) based on the MAD calculated using the same validation set as that used in case study 2.

Before investigating the trade-offs between the reaction rate of the main reaction and reaction selectivity, we consider two single-objective CAMD problems. A MBDoE CAMD problem and a Div CAMD problem are formulated for each of the objectives, i.e., maximising the rate constant of the O-alkylation reaction and minimising the rate constant of the C-alkylation reaction. Similarly to the previous two case studies, the MBDoE models provide more reliable predictions of the QM rate constants for both the O-alkylation and C-alkylation reactions. The use of these more accurate MLR models in the first iterations reduces the total number of total iterations required for the QM CAMD procedures to terminate; both the MBDoE CAMD problems for the O-alkylation and C-alkylation reactions reach convergence in three iterations. Four iterations are needed for the Div CAMD problems of these two objectives. After completion of the QM-CAMD procedures, nitromethanol is identified by the MBDoE CAMD problem as the optimal solution to maximise the O-alkylation rate with a QM rate constant of $k^{L,QM} = 7.13 \times 10^{-2}$ L mol^{-1} s^{-1} while the Div CAMD problem identifies the optimal solvent to be DMSO, which provides a slightly larger rate constant ($k^{L,QM} = 7.53\times10^{-2}$ L mol^{-1} s^{-1}) than nitromethanol. In the case of C-alkylation, where the reaction rate is minimised, chlorostyrene is

Table 2
Design results of the MBDoE CAMD problem and Div CAMD problem for the Menschutkin reaction. The rate constants are in the unit of L mol-1 s-1. The mean absolute deviations (MAD), in log units, are calculated for the 326 solvents in the design space. †Multiple isomers exist for the same group combination.

First iteration	DoE CAMD	Div CAMD
Optimal solvent name	Nitromethanol	Dimethyl sulfoxide
Optimal solvent structure	OH × 1 CH2NO2 × 1	C2H6SO × 1
$k^{L,MLR-1}$	2.28×10^{-1}	3.88×10^{2}
$k^{L,QM}$	1.56×10^{-3}	1.09×10^{-3}
MAD	1.90	3.51
Second iteration	**DoE CAMD**	**Div CAMD**
Optimal solvent name	Nitromethanol	4-Methyl-3-nitromethylpent-3-en-1-ol†
Optimal solvent structure	OH × 1 CH2NO2 × 1	CH3 × 2 CH2 × 2 C=C × 1 OH × 1 CH2NO2 × 1
$k^{L,MLR-2}$	4.13×10^{-3}	1.50×10^{-2}
$k^{L,QM}$	1.56×10^{-3}	4.77×10^{-4}
MAD	1.43	1.62
Third iteration	**DoE CAMD**	**Div CAMD**
Optimal solvent name	Converged	2,3-Dimethyl-1-nitro-but-2-ene
Optimal solvent structure		CH3 × 3 C=C × 1 CH2NO2 × 1
$k^{L,MLR-3}$		3.33×10^{-3}
$k^{L,QM}$		1.76×10^{-4}
MAD		2.05
Fourth iteration	**DoE CAMD**	**Div CAMD**
Optimal solvent name	Converged	Nitromethanol
Optimal solvent structure		OH × 1 CH2NO2 × 1
$k^{L,MLR-4}$		7.68×10^{-3}
$k^{L,QM}$		1.56×10^{-3}
MAD		0.79
Fifth iteration	**DoE CAMD**	**Div CAMD**
Optimal solvent name	Converged	Nitromethanol
Optimal solvent structure		OH × 1 CH2NO2 × 1
$k^{L,MLR-5}$		1.93×10^{-3}
$k^{L,QM}$		1.56×10^{-3}
MAD		0.94

Table 3
The top 10 ranked solvents generated by the converged MBDoE and Div models for the Menschutkin reaction. †Solvents that do not appear in both lists.

Ranking	Solvent name	Solvent structure	$\ln k^{L,MBDoE-2}$	$\ln k^{L,QM}$
1	Nitromethanol	OH × 1 CH2NO2 × 1	−5.49	−6.46
2	Dimethyl sulfoxide	C2H6SO × 1	−5.57	−6.82
3	2-Nitroethanol	CH2 × 1 OH × 1 CH2NO2 × 1	−6.26	−6.72
4	3-Nitroprop-1-en-2-ol	CH2=C × 1 OH × 1 CH2NO2 × 1	−6.65	−6.98
5	3-Nitropropanol	CH2 × 2 OH × 1 CH2NO2 × 1	−6.73	−6.93
6	1-Nitropropan-2-ol	CH3 × 1 CH × 1 OH × 1 CH2NO2 × 1	−6.87	−6.96
7	2-Methyl-1-nitropropan-2-ol†	CH3 × 2 C × 1 OH × 1 CH2NO2 × 1	−6.90	−7.23
8	3-Nitroprop-1-en-1-ol	CH=CH × 1 OH × 1 CH2NO2 × 1	−6.91	−7.00
9	2-Nitromethyl-prop-2-en-1-ol	CH2 × 1 CH2=C × 1 OH × 1 CH2NO2 × 1	−6.97	−7.15
10	N-Methylformamide	C2H5NO × 1	−6.99	−5.92
Ranking	**Solvent name**	**Solvent structure**	$\ln k^{L,Div-5}$	$\ln k^{L,QM}$
1	Nitromethanol	OH × 1 CH2NO2 × 1	−6.25	−6.46
2	N-Methylformamide	C2H5NO × 1	−6.51	−5.92
3	2-Nitroethanol	CH2 × 1 OH × 1 CH2NO2 × 1	−6.74	−6.72
4	Dimethyl sulfoxide	C2H6SO × 1	−6.85	−6.82
5	3-Nitroprop-1-en-2-ol	CH2=C × 1 OH × 1 CH2NO2 × 1	−7.02	−6.98
6	3-Nitropropanol	CH2 × 2 OH × 1 CH2NO2 × 1	−7.04	−6.93
7	3-Nitroprop-1-en-1-ol	CH=CH × 1 OH × 1 CH2NO2 × 1	−7.14	−7.00
8	1-Nitropropan-2-ol	CH3 × 1 CH × 1 OH × 1 CH2NO2 × 1	−7.21	−6.96
9	2-Nitromethyl-prop-2-en-1-ol	CH2 × 1 CH2=C × 1 OH × 1 CH2NO2 × 1	−7.23	−7.15
10	4-Nitrobutanol†	CH2 × 3 OH × 1 CH2NO2 × 1	−7.25	−7.10

generated by the MBDoE CAMD problem with a QM rate constant of $k^{L,QM} = 1.02 \times 10^{-4}$ L mol-1 s-1. This is larger than the QM rate constant of 2,3,4-trimethyl-2-pentene ($k^{L,QM} = 8.32 \times 10^{-5}$ L mol-1 s-1) which is the optimal solvent identified by the Div CAMD problem. It can be seen from these results that in this case the MBDoE problems do not identify better solvents than those identified by the Div problems though the rate constants are similar in each scenario.

The trade-offs between the rate constant of the main reaction (O-alkylation) and reaction selectivity are investigated as shown in Fig. 7. To gain insights into the trade-offs, the full space of solutions is explored. The reaction selectivity of the designed solvent, α, is calculated as

$$\ln \alpha = \ln k_{O}^{L,MBDoE} - \ln k_{C}^{L,MBDoE} \tag{62}$$

where $k_{O}^{L,MBDoE}$ and $k_{C}^{L,MBDoE}$ are the rate constants predicted by the converged MBDoE models for the O-alkylation reaction and the C-alkylation reaction, respectively. All the feasible molecules in the design space are generated using the MILP formulation of CAMD with

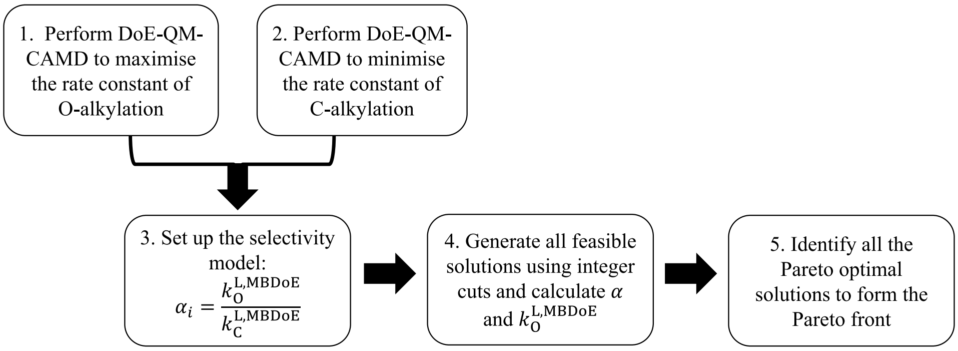
1. Perform DoE-QM-CAMD to maximise the rate constant of O-alkylation
2. Perform DoE-QM-CAMD to minimise the rate constant of C-alkylation
3. Set up the selectivity model: $\alpha_i = \frac{k_{\mathrm{O}}^{\mathrm{L,MBDoE}}}{k_{\mathrm{C}}^{\mathrm{L,MBDoE}}}$
4. Generate all feasible solutions using integer cuts and calculate α and $k_{\mathrm{O}}^{\mathrm{L,MBDoE}}$
5. Identify all the Pareto optimal solutions to form the Pareto front

Fig. 7. Workflow for generating the Pareto front in case study 3.

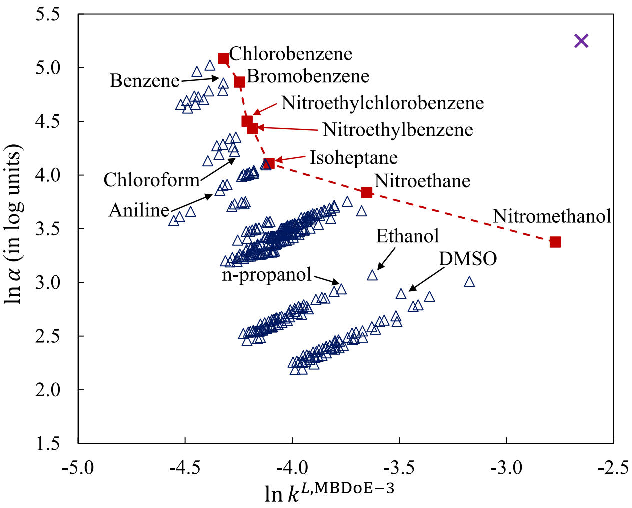

Fig. 8. Natural logarithm of selectivity vs logarithm of the O-alkylation rate constant for all the feasible solvent molecules in the design space. The Pareto front is highlighted with a red dashed line connecting all the Pareto optimal solutions (red squares), the blue triangles represent the dominated solutions and the purple cross represents an ideal solution that achieves both high selectivity and large rate constant.

integer cuts. The calculated rate constant of the main reaction and reaction selectivity of the 326 solvents are shown in Fig. 8, where the trade-offs between the two objectives, i.e., the molecules on the Pareto front, are highlighted. Chlorobenzene is found to be the solvent that maximises the selectivity but shows low rate constant for the O-alkylation while nitromethanol gives the largest rate constant for the O-alkylation but leads to low selectivity. As the rate of the O-alkylation reaction increases, selectivity drops rapidly until $\ln k_{\mathrm{O}}^{\mathrm{L,MBDoE}}$ reaches −4.10 and then it decreases gradually. It is also observed that the data points in Fig. 8 form clusters separated by gaps where the associated combinations of selectivity and $\ln k_{\mathrm{O}}^{\mathrm{L,MBDoE}}$ are infeasible. After careful examination of the solvent molecules in each cluster, it is found that each cluster corresponds to a family of molecules that share one or more atom group(s). For example, the cluster in the region ($\ln k_{\mathrm{O}}^{\mathrm{L,MBDoE}} \in [-4.3, -3.6]$, $\alpha \in [2.4, 3.1]$) corresponds to the alcohol family. This phenomenon reflects the discrete nature of CAMD problems. These infeasible regions between the clusters in Fig. 8 could potentially be reached by taking into account more types of solvents or atom groups or by considering solvent mixtures.

4. Conclusions and future work

An enhanced CAMD method for reaction solvent design, DoE-QM-CAMD, has been developed. Two MBDoE formulations have been proposed and tested, either using a list of solvents with known (measured) properties or a (larger) list of solvents with known or predicted properties, to expand the design space of candidate solvents. This latter approach has been found to be effective in identifying a set of initial solvents with a large degree of diversity, as evidenced by a high D-optimality value. Using the initial solvent set thus identified, an MLR surrogate model, the solvatochromic equation, can be parameterised based on QM-computed reaction rate constants and used within a QM-CAMD framework to identify solvents that improve reaction performance. The integration of MBDoE and QM-CAMD was applied to three case studies and compared with the original QM-CAMD approach in which the initial solvent set is based on chemical intuition (the Div set). The MBDoE-derived model was found to offer consistently higher predictive accuracy in the first iteration of the QM-CAMD design loop than the Div-derived model. As a result, the MBDoE CAMD problem was generally found to converge within fewer steps than the Div CAMD problem. Regardless of the model used, both approaches have been shown to lead to the identification of solvents that can considerably improve reaction performance, whether based on a single objective (reaction rate) or two objectives (reaction rate and selectivity). Given that the computational expense of implementing the MBDoE method is small compared to that of QM calculations, the use of the proposed systematic MBDoE methodology for initial solvent set design is preferred over the *ad hoc* selection of such a set on the basis of perceived diversity. Overall, the integration of the MBDoE technique into the QM-CAMD framework improves the performance of the surrogate kinetic model without introducing extra complexity in the simple MLR formalism.

Given the promising results obtained with MBDoE relative to the use of solvents selected based on expert intuition, more work needs to be done to understand the relationship between the D-optimality criterion values and model performance. One of the underlying assumptions in our work is that the model errors can be viewed as random and normally distributed. Errors in physics-based computer experiments are often systematic. Therefore, it would be interesting to take this behaviour into account more formally.

We anticipate that the developed method can be applied to optimise the rates of many other chemical reactions and other properties that can be predicted using a linear free energy relationship, such as solubility or octanol-water partition coefficients. Further improvements to the surrogate model will also be investigated using MBDoE formulations in which solvents are designed from atom groups rather than selected from a list or in which hypothetical or alchemical solvents are defined by the values of the continuous descriptors without reference to a specific molecular structure. The impact of using different types of surrogate model, other sets of solvent properties/descriptors, and alternative strategies to update the surrogate model will be examined with the aim to achieve even greater reliability.

CRediT authorship contribution statement

Lingfeng Gui: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Visualization. **Yijun Yu:** Conceptualization, Methodology, Software, Investigation, Writing – review & editing. **Titilola O. Oliyide:** Conceptualization, Methodology, Software, Investigation, Writing – review & editing. **Eirini Siougkrou:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Alan Armstrong:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Amparo Galindo:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Funding acquisition. **Fareed Bhasha Sayyed:** Conceptualization, Writing – review & editing, Supervision. **Stanley P. Kolis:** Conceptualization, Writing – review & editing, Supervision. **Claire S. Adjiman:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Claire Adjiman reports financial support was provided by Eli Lilly and Company. Amparo Galindo reports financial support was provided by Eli Lilly and Company. Alan Armstrong reports financial support was provided by Eli Lilly and Company. Lingfeng Gui reports financial support was provided by Eli Lilly and Company. Claire Adjiman is a member of the editorial of Computers & Chemical Engineering - CSA.

Data availability

Data are available on the Zenodo repository (https://doi.org/10.5281/zenodo.7839710).

Acknowledgements

Funding from Eli Lilly and Company, United States and the UK EPSRC, through the PharmaSEL-Prosperity Programme (EP/T005556/1), is gratefully acknowledged. AG acknowledges the funding of a Research Chair by the Royal Academy of Engineering and Eli Lilly and Company (RCSRF1819/7/33). The Imperial College Research Computing Service (DOI: 10.14469/hpc/2232) is also gratefully acknowledged for providing support and resources for the quantum mechanical calculations.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compchemeng.2023.108345.

References

Abraham, M.H., 1993. Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes. Chem. Soc. Rev. 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. J. Chem. Soc., Perkin Trans. 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. J. Chem. Soc., Perkin Trans. 2 1097–1101.

Aldeghi, M., Coley, C.W., 2022. A focus on simulation and machine learning as complementary tools for chemical space navigation. Chem. Sci. 13, 8221–8223.

Atkinson, A.C., Donev, A.N., Tobias, R.D., 2007. Optimum Experimental Designs, with SAS. Oxford University Press, Oxford.

Austin, N.D., Sahinidis, N.V., Konstantinov, I.A., Trahan, D.W., 2018. COSMO-based computer-aided molecular/mixture design: A focus on reaction solvents. AIChE J. 64 (1), 104–122.

Austin, N.D., Sahinidis, N.V., Trahan, D.W., 2016. Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques. Chem. Eng. Res. Des. 116, 2–26.

Austin, N.D., Sahinidis, N.V., Trahan, D.W., 2017. A COSMO-based approach to computer-aided mixture design. Chem. Eng. Sci. 159, 93–105.

Buxton, A., Livingston, A.G., Pistikopoulos, E.N., 1999. Optimal design of solvent blends for environmental impact minimization. AIChE J. 45 (4), 817–843.

Cao, H.-Q., Duan, Q.-L., Chai, H., Li, X.-X., Sun, J.-H., 2020. Experimental study of the effect of typical halides on pyrolysis of ammonium nitrate using model reconstruction. J. Hazard. Mater. 384, 121297.

Coley, C.W., Eyke, N.S., Jensen, K.F., 2020a. Autonomous discovery in the chemical sciences part I: Progress. Angew. Chem. Int. Edn 59 (51), 22858–22893.

Coley, C.W., Eyke, N.S., Jensen, K.F., 2020b. Autonomous discovery in the chemical sciences part II: Outlook. Angew. Chem. Int. Edn 59 (52), 23414–23436.

CPLEX, IBM ILOG, 2009. V12. 1: User's Manual for CPLEX. International Business Machines Corporation 46 (53), 157.

Curtiss, L.A., Redfern, P.C., Raghavachari, K., Rassolov, V., Pople, J.A., 1999. Gaussian-3 theory using reduced Møller-Plesset order. J. Chem. Phys. 110 (10), 4703–4709.

Diamanti, A., Ganase, Z., Grant, E., Armstrong, A., Piccione, P.M., Rea, A.M., Richardson, J., Galindo, A., Adjiman, C.S., 2021. Mechanism, kinetics and selectivity of a williamson ether synthesis: elucidation under different reaction conditions. React. Chem. Eng. 6, 1195–1211.

Erny, M., Lundqvist, M., Rasmussen, J.H., Ludemann-Hombourger, O., Bihel, F., Pawlas, J., 2020. Minimizing HCN in DIC/Oxyma-mediated amide bond-forming reactions. Org. Process Res. Dev. 24 (7), 1341–1349.

Eyring, H., 1935. The activated complex in chemical reactions. J. Chem. Phys. 3 (2), 107–115.

Folić, M., Adjiman, C.S., Pistikopoulos, E.N., 2007. Design of solvents for optimal reaction rate constants. AIChE J. 53 (5), 1240–1256.

Folić, M., Adjiman, C.S., Pistikopoulos, E.N., 2008. Computer-aided solvent design for reactions: Maximizing product formation. Ind. Eng. Chem. Res. 47 (15), 5190–5202.

Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Scalmani, G., Barone, V., Petersson, G.A., Nakatsuji, H., Li, X., Caricato, M., Marenich, A.V., Bloino, J., Janesko, B.G., Gomperts, R., Mennucci, B., Hratchian, H.P., Ortiz, J.V., Izmaylov, A.F., Sonnenberg, J.L., Williams-Young, D., Ding, F., Lipparini, F., Egidi, F., Goings, J., Peng, B., Petrone, A., Henderson, T., Ranasinghe, D., Zakrzewski, V.G., Gao, J., Rega, N., Zheng, G., Liang, W., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Throssell, K., Montgomery, J.A., Peralta, J.E., Ogliaro, F., Bearpark, M.J., Heyd, J.J., Brothers, E.N., Kudin, K.N., Staroverov, V.N., Keith, T.A., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A.P., Burant, J.C., Iyengar, S.S., Tomasi, J., Cossi, M., Millam, J.M., Klene, M., Adamo, C., Cammi, R., Ochterski, J.W., Martin, R.L., Morokuma, K., Farkas, O., Foresman, J.B., Fox, D.J., 2016. Gaussian 16 revision c.01.

Gertig, C., Fleitmann, L., Schilling, J., Leonhard, K., Bardow, A., 2020. Rx-COSMO-CAMPD: enhancing reactions by integrated computer-aided design of solvents and processes based on quantum chemistry. Chem. Ing. Tech. 92 (10), 1489–1500.

Gertig, C., Kröger, L., Fleitmann, L., Scheffczyk, J., Bardow, A., Leonhard, K., 2019. Rx-COSMO-CAMD: computer-aided molecular design of reaction solvents based on predictive kinetics from quantum chemistry. Ind. Eng. Chem. Res. 58 (51), 22835–22846.

Grant, E., Pan, Y., Richardson, J., Martinelli, J.R., Armstrong, A., Galindo, A., Adjiman, C.S., 2018. Multi-objective computer-aided solvent design for selectivity and rate in reactions. In: Eden, M.R., Ierapetritou, M.G., Towler, G.P. (Eds.), 13th International Symposium on Process Systems Engineering (PSE 2018). In: Computer Aided Chemical Engineering, Vol. 44, Elsevier, pp. 2437–2442.

Grom, M., Stavber, G., Drnovšek, P., Likozar, B., 2016. Modelling chemical kinetics of a complex reaction network of active pharmaceutical ingredient (API) synthesis with process optimization for benzazepine heterocyclic compound. Chem. Eng. J. 283, 703–716.

Gui, L., Adjiman, C.S., Galindo, A., Sayyed, F.B., Kolis, S.P., Armstrong, A., 2023. Uncovering the most kinetically influential reaction pathway driving the generation of HCN from Oxyma/DIC adduct: a theoretical study. Ind. Eng. Chem. Res. 62 (2), 874–880.

Gui, L., Armstrong, A., Galindo, A., Sayyed, F.B., Kolis, S.P., Adjiman, C.S., 2022. Computer-aided solvent design for suppressing HCN generation in amino acid activation. In: Montastruc, L., Negny, S. (Eds.), 32nd European Symposium on Computer Aided Process Engineering. In: Computer Aided Chemical Engineering, Vol. 51, Elsevier, pp. 607–612.

Harvey, J.N., Himo, F., Maseras, F., Perrin, L., 2019. Scope and challenge of computational methods for studying mechanism and reactivity in homogeneous catalysis. ACS Catal. 9 (8), 6803–6813.

Hildebrand, J.H., 1929. Solubility. XII. Regular solutions[1]. J. Am. Chem. Soc. 51 (1), 66–80.

Hill, R.G., 2005. 10 - Biomedical polymers. In: Hench, L.L., Jones, J.R. (Eds.), Biomaterials, Artificial Organs and Tissue Engineering. In: Woodhead Publishing Series in Biomaterials, Woodhead Publishing, pp. 97–106.

Ho, J., Ertem, M.Z., 2016. Calculating free energy changes in continuum solvation models. J. Phys. Chem. B 120 (7), 1319–1329.

Hukkerikar, A.S., Kalakul, S., Sarup, B., Young, D.M., Sin, G., Gani, R., 2012a. Estimation of environment-related properties of chemicals for design of sustainable processes: development of group-contribution+ (GC+) property models and uncertainty analysis. J. Chem. Inf. Model. 52 (11), 2823–2839.

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

Issa, A.A., Luyt, A.S., 2019. Kinetics of alkoxysilanes and organoalkoxysilanes polymerization: A review. Polymers 11 (3), 537.

Jalan, A., Alecu, I.M., Meana-Pañeda, R., Aguilera-Iparraguirre, J., Yang, K.R., Merchant, S.S., Truhlar, D.G., Green, W.H., 2013. New pathways for formation of acids and carbonyl products in low-temperature oxidation: The korcek decomposition of γ-ketohydroperoxides. J. Am. Chem. Soc. 135 (30), 11100–11114.

John, R.C.S., Draper, N., 1975. D-optimality for regression designs: A review. Technometrics 17 (1), 15–23.

Karunanithi, A.T., Achenie, L.E., Gani, R., 2006. A computer-aided molecular design framework for crystallization solvent design. Chem. Eng. Sci. 61 (4), 1247–1260.

Klamt, A., 1995. Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena. J. Phys. Chem. 99 (7), 2224–2235.

Kocis, G.R., Grossmann, I.E., 1989. Computational experience with dicopt solving MINLP problems in process systems engineering. Comput. Chem. Eng. 13 (3), 307–315.

Komp, E., Janulaitis, N., Valleau, S., 2022. Progress towards machine learning reaction rate constants. Phys. Chem. Chem. Phys. 24, 2692–2705.

Komp, E., Valleau, S., 2020. Machine learning quantum reaction rate constants. J. Phys. Chem. A 124 (41), 8607–8613.

Laidler, K.J., King, M.C., 1983. Development of transition-state theory. J. Phys. Chem. 87 (15), 2657–2664.

Liu, Q., Zhang, L., Liu, L., Du, J., Meng, Q., Gani, R., 2019. Computer-aided reaction solvent design based on transition state theory and COSMO-SAC. Chem. Eng. Sci. 202, 300–317.

Lu, J., Zhang, H., Yu, J., Shan, D., Qi, J., Chen, J., Song, H., Yang, M., 2021. Predicting rate constants of hydroxyl radical reactions with alkanes using machine learning. J. Chem. Inf. Model. 61 (9), 4259–4265.

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. J. Phys. Chem. B 113 (18), 6378–6396.

Martins, S.I., Jongen, W.M., van Boekel, M.A., 2000. A review of Maillard reaction in food and implications to kinetic modelling. Trends Food Sci. Technol. 11 (9), 364–373.

McFarland, A.D., Buser, J.Y., Embry, M.C., Held, C.B., Kolis, S.P., 2019. Generation of Hydrogen Cyanide from the reaction of Oxyma (Ethyl Cyano(hydroxyimino)acetate) and DIC (Diisopropylcarbodiimide). Org. Process Res. Dev. 23 (9), 2099–2105.

Menschutkin, N., 1890a. Beiträge zur Kenntnis der Affinitätskoeffizienten der Alkylhaloide und der organischen Amine. Z. Phys. Chem. 5U (1), 589–600.

Menschutkin, N., 1890b. Über die Affinitätskoeffizienten der Alkylhaloide und der Amine: Zweiter Teil. Über den Einfluss des chemisch indifferenten flüssigen Mediums auf die Geschwindigkeit der Verbindung des Triäthylamins mit den Alkyljodiden. Z. Phys. Chem. 6U (1), 41–57.

Meuwly, M., 2019. Reactive molecular dynamics: From small molecules to proteins. Wiley Interdiscip. Rev. Comput. Mol. Sci. 9 (1), e1386.

Miertus, S., Scrocco, E., Tomasi, J., 1981. Electrostatic interaction of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects. Chem. Phys. 55 (1), 117–129.

Miertus, S., Tomasi, J., 1982. Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes. Chem. Phys. 65 (2), 239–245.

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

Oliyide, T.O., 2014. Design of molecules to build optimal models of solvent effects on reactions. (Master's thesis). Imperial College London.

Pollice, R., dos Passos Gomes, G., Aldeghi, M., Hickman, R.J., Krenn, M., Lavigne, C., Lindner-D'Addario, M., Nigam, A., Ser, C.T., Yao, Z., Aspuru-Guzik, A., 2021. Data-driven strategies for accelerated materials design. Acc. Chem. Res. 54 (4), 849–860.

Potyrailo, R., Rajan, K., Stoewe, K., Takeuchi, I., Chisholm, B., Lam, H., 2011. Combinatorial and high-throughput screening of materials libraries: Review of state of the art. ACS Comb. Sci. 13 (6), 579–633.

Reinheimer, J.D., Harley, J.D., Meyers, W.W., 1963. Solvent effects in the Menschutkin reaction. J. Org. Chem. 28 (6), 1575–1579.

Sahinidis, N.V., Tawarmalani, M., Yu, M., 2003. Design of alternative refrigerants via global optimization. AIChE J. 49 (7), 1761–1775.

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

Scheffczyk, J., Fleitmann, L., Schwarz, A., Lampe, M., Bardow, A., Leonhard, K., 2017. COSMO-CAMD: A framework for optimization-based computer-aided molecular design using COSMO-RS. Chem. Eng. Sci. 159, 84–92.

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 Equilib. 231 (1), 27–37.

Song, Q.-W., Zhou, Z.-H., He, L.-N., 2017. Efficient, selective and sustainable catalysis of carbon dioxide. Green Chem. 19, 3707–3728.

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 Chem. 5, 952–957.

Struebing, H., Obermeier, S., Siougkrou, E., Adjiman, C.S., Galindo, A., 2017. A QM-CAMD approach to solvent design for optimal reaction rates. Chem. Eng. Sci. 159, 69–83.

Tsichla, A., Severins, C., Gottfried, M., Marquardt, W., 2019. An experimental assessment of model-based solvent selection for enhancing reaction kinetics. Ind. Eng. Chem. Res. 58 (30), 13517–13532.

Wang, K., Dowling, A.W., 2022. Bayesian optimization for chemical products and functional materials. Curr. Opin. Chem. Eng. 36, 100728.

Watkins, D.S., 1991. Fundamentals of Matrix Computations. Wiley, New York, p. 84.

Watson, O.L., Jonuzaj, S., McGinty, J., Sefcik, J., Galindo, A., Jackson, G., Adjiman, C.S., 2021. Computer aided design of solvent blends for hybrid cooling and antisolvent crystallization of active pharmaceutical ingredients. Org. Process Res. Dev. 25 (5), 1123–1142.

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

Wigner, E., 1937. Calculation of the rate of elementary association reactions. J. Chem. Phys. 5 (9), 720–725.

Winey, J.M., Gupta, Y.M., 1997. Shock-induced chemical changes in neat nitromethane: use of time-resolved Raman spectroscopy. J. Phys. Chem. B 101 (50), 10733–10743.

Wold, S., 1995. Chemometrics; what do we mean with it, and what do we want from it? Chemom. Intell. Lab. Syst. 30 (1), 109–115.

Zhao, Y., Truhlar, D.G., 2008. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. Theor. Chem. Acc. 120 (1), 215–241.

Zhou, T., Lyu, Z., Qi, Z., Sundmacher, K., 2015a. Robust design of optimal solvents for chemical reactions—A combined experimental and computational strategy. Chem. Eng. Sci. 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 J. 61 (1), 147–158.