A model-based framework assisting the design of vapor-liquid equilibrium experimental plans

Belmiro P.M. Duarte[a,d], Anthony C. Atkinson[b], José F.O. Granjo[c], Nuno M.C. Oliveira[d]

[a]*ISEC, Department of Chemical & Biological Engineering, Polytechnic Institute of Coimbra, Rua Pedro Nunes, 3030–199 Coimbra, Portugal.*
[b]*Department of Statistics, London School of Economics, London WC2A 2AE, United Kingdom.*
[c]*CERENA, Department of Chemical Engineering, Instituto Superior Técnico, Av. Rovisco Pais 1, 1049–001, Lisbon, Portugal.*
[d]*CIEPQPF, Department of Chemical Engineering, University of Coimbra, Rua Sílvio Lima — Polo II, 3030–790 Coimbra, Portugal.*

Abstract

In this paper we propose a framework for Model-based Sequential Optimal Design of Experiments to assist experimenters involved in Vapor-Liquid equilibrium characterization studies to systematically construct thermodynamically consistent models. The approach uses an initial continuous optimal design obtained via semidefinite programming, and then iterates between two stages (i) model fitting using the information available; and (ii) identification of the next experiment, so that the information content in data is maximized. The procedure stops when the number of experiments reaches the maximum for the experimental program or the dissimilarity between the parameter estimates during two consecutive iterations is below a given threshold. This methodology is exemplified with the D-optimal de-

*Corresponding author. Tel.: +351 239 798 200.
Email addresses: bduarte@isec.pt (Belmiro P.M. Duarte), a.c.atkinson@lse.ac.uk (Anthony C. Atkinson), josegranjo9@tecnico.ulisboa.pt (José F.O. Granjo), nuno@eq.uc.pt (Nuno M.C. Oliveira)

sign of isobaric experiments, for characterizing binary mixtures using the NRTL and UNIQUAC thermodynamic models for liquid phase. Significant reductions of the confidence regions for the parameters are achieved compared with experimental plans where the observations are uniformly distributed over the domain.
Keywords: Sequential optimal design of experiments, Vapor-Liquid Equilibrium, Semidefinite Programming, NRTL model, Nonlinear Programming.

1. Introduction

Numerous vapor-liquid equilibrium (VLE) experiments are undertaken to (i) build models that are subsequently used for process design and optimization; and (ii) improve the understanding of the VLE system. VLE models are crucial in designing, optimizing and controlling process equipment (e.g., distillation or flash operation), finding application in many chemical industries, including petrochemicals, pharmaceuticals and food processing. Constructing adequate mathematical models from VLE data typically involves seven steps or variants thereof: (i) choosing the appropriate thermodynamic modeling framework, according to the nature of the components and the operating region [50]; (ii) setting the experimental apparatus to extract the data with the required accuracy and precision [59, 48]; (iii) determining the experimental plan [47]; (iv) setting the criterion used for parameter estimation [32]; (v) performing the experiments and collecting the data [62]; (vi) performing the parameter estimation given the thermodynamic model [34, 42] and the criterion previously chosen, which, in turn, requires a phase stability analysis to prevent spurious phase predictions [53], and consistency tests to identify data anomalies [39, 75]; (vii) if necessary, iterating from (v) using the results of further experiments to refine the parameter estimates, until the estimated

model satisfies the experimental objectives. While there is a rich body of knowledge to tackle steps (i-ii) and (iv-vi), to the best of our knowledge, methods to deal with (iii) in the light of (iv) and with (vii) are practically non-existent. Very often the choice of experiments relies on an uniform grid temperature (or pressure) conditions. Consequently, even correctly obtained and accurate VLE data may be insufficient to support the fitting model, leading to high covariances in parameters and lack of identifiability which undermine its use in process simulation and optimization environments.

The experimental effort required for gathering VLE data depends on the system and operating conditions considered. While this procedure can now be routinely performed for low pressure and non-azeotropic systems, it can also involve more significant planning and effort when high pressure-temperature conditions are involved. Hence, it is surprising that experimental plans are still customarily established without the use of tools that can maximize the information about the system, subject to the available resources [17]. As a result, the experiments may yield limited information for the accurate estimation of the parameters of the chosen thermodynamic model [55]. In addition, the models used to describe VLE behavior are nonlinear by nature, so that good designs paradoxically require accurate knowledge of the parameters to be estimated. Given the complexity of finding robust Bayesian or minimax optimal designs to overcome parametric uncertainty, the optimal designs prescribed would be *local* [5]. In such cases, the information is maximized assuming a given set of operating parameter estimates. Consequently, locally determined optimal designs may perform poorly if the true parameter values are appreciably different from those used to calculate the design [69]. A good way to circumvent this issue is the use of sequential optimal de-

signs, where the experimental plan includes an initial step where a locally optimal design is obtained and carried out, followed by a phase where the set of parameters is iteratively re-estimated after each set of new results [80]. Specifically, the subsequent experiment is chosen so that the amount of information available in the augmented data set is maximized [38]. The procedure terminates when the model parameters are accurate enough or the resources available have been depleted. The rationale just described is the basis for the model-based sequential optimal design of experiments (M-bSODE) [23]. An application of M-bSODE to model parametrization appears in [5, §17.7].

M-bSODE helps to identify and plan the set of experiments required to improve the precision or significance of the estimated parameters, providing a steady incremental increase in the accuracy of the parameter estimates and guiding the experimental work with model-based designs based on these estimates. Metrics, such as the reduction of the volume of the confidence region for the parameters, can then be computed using the additional information available from any new experiment. In this way, successive optimal experiment(s) can be chosen. Practically, the method also allows for gradual refinement of the experimental region.

The minimum number of data measurements required to derive the model parameters with a given accuracy is difficult to forecast because the results of the Model-based Optimal Design of Experiments (M-bODE) procedure are dependent on the currently assumed parameter values. If the current values are known to be exact, then we can apply the proposed procedure to determine how many and what are the new experimental conditions for the required experiments. However, if the current estimates are inaccurate (which is usually the case), these predictions would probably change, as new information about the system is obtained. This is

the reason for the incremental structure of the proposed algorithm. Nevertheless, once the parameters become approximate, this question can still be approximately answered in the framework of M-bSODE.

The use of Model-based Optimal Design of Experiments is well established in the Chemical Engineering literature. Most of the applications aim at finding the optimal sequence of actions on input variables and/or time instants at which sampling is required so that the information obtained from dynamic experiments is maximized [33, 4]. The problem is formulated as an optimal control problem [79, 57] and numerically handled with dynamic optimization techniques [72, 51, 46]. Recent applications include systems with continuous measurement [41], online redesign of experiments considering the amount of information gathered previously and the model inaccuracy [40], the design of robust experiments taking into account the uncertainty of the model and violation of the constraints [68] and an application to a real case study where local identifiability is simultaneously monitored and used to transform the problem into a well-conditioned equivalent form [9]. All these references focus on dynamic experiments; the literature on M-bODE of static experiments applied to thermodynamic models is, from our knowledge, limited to Dechambre et al. [25] where they use a Wynn-Fedorov algorithm to prescribe optimal experimental for Liquid-Liquid Equilibrium characterization considering the Non-Random Two-Liquid (NRTL) and the Universal Quasichemical (UNIQUAC) and Duarte et al. [30] where Semidefinite Programming is used.

The works of Galvanin et al. [40] and Barz et al. [9] involve the redesign of experiments which can be seen as an automated variant of M-bSODE in dynamic experiments. However, the application of M-bSODE frameworks within static ex-

perimental plans where human resources are required is rather scarce. The rationale is using the knowledge previously obtained to prescribe the next experiment to maximize the information. Among the references in the Chemical Engineering literature are Brendel et al. [16], Thompson et al. [69] who addressed the identification of complex reaction kinetics in chemical reactors, Goujot et al. [43] whose application is a compartmental model describing the convective drying of rice, and Soepyan et al. [66] where the methodology couples the sequential design of experiments with a software tool to build surrogate models for learning about the operation of a solvent-based CO_2 capture pilot plant. M-bSODE frameworks were also consistently applied in model discrimination, see Buzzi-Ferraris et al. [19] and Buzzi-Ferraris et al. [18] among others. Further examples of M-bSODE can be found primarily in drug development; specifically in dose-response studies using prediction models, see Dragalin and Fedorov [26], Dragalin et al. [27], Wang et al. [73], Leonov and Miller [52] among others.

While a larger number of measurements can help increasing the confidence degree in the parameter values produced, extensive sampling of the domain might not be required or correspond to the most efficient method of achieving a reasonable description of the underlying physical system, as demonstrated in the examples in §4 and §5 (see also [56]). This is the key concept of Optimal Experimental Design (OED), which has been successfully applied in many scientific domains, most remarkably when new experiments are expensive. The extensive sampling methodology still corresponds to the traditional approach for conducting experimental studies in the VLE area. But the current situation can be improved by using an assumed model for measuring the information content and iteratively maximizing the information value of each new experiment by choosing the conditions that

minimize a measure of the size of the parametric confidence region.

This paper addresses the M-bSODE for thermodynamic models describing VLE data which, to the best of our knowledge, has not been demonstrated in the open literature. We aim at establishing a computational framework capable of guiding the experimenters in planning measurements to build and refine thermodynamic models for VLE characterization, with a minimal use of resources. M-bSODE complements other well established tools, such as the NIST ThermoData Engine [39] which are focused on dynamic data validation rather than on designing experiments to improve thermodynamic models precision. The algorithms proposed herein can be implemented in a user friendly software package, so that they become amenable and helpful for researchers involved in VLE characterization. This will ensure that a large group of researchers currently involved in experimental studies will have a suitable planning tool available to support their work, in order to improve the accuracy of the resulting models, and help in rationalizing the corresponding resource usage.

One important feature of the proposed methodology is that it allows anticipating whether it will produce significant improvements, compared with a more traditional regular domain sampling. The answer is related to the optimal location of the new experimental points. If the locations of the optimal sampling points deviate significantly from a regular (equidistributed) sampling plan, then it is more plausible that differences in the results can be expected (and the opposite, in the remaining case). The proposed methodology is grounded on mathematical programming, relying on algorithms which have improved substantially over the last two decades and are now able to handle complex, large-scale optimization problems. Practically, our framework requires solving nonlinear optimization prob-

lems to (i) fit the available data set; and (ii) find the next experimental conditions that maximize the information obtained. Semidefinite Programming is used for constructing the initial design.

Herein, we consider the D-optimality criterion for measuring the amount of information in data and apply the approach to isobaric measurements of binary mixtures. Although the study assumes that the vapor phase is ideal and that the non-ideal behavior in the liquid phase is modeled with NRTL [61] or UNIQUAC [1], the proposed approach can be generalized to other continuously differentiable optimality criteria, isothermal experimental setups and thermodynamic models through the adaptation of VLE description used (Appendix B).

1.1. Paper organization and nomenclature

Section 2 presents the mathematical representation of the VLE model, the background required for building the M-bSODE framework, and the general characteristics of the mathematical programming tools used. Section 3 presents the framework algorithm and analyzes each step in depth. Section 4 demonstrates its application to the methanol-water system where the thermodynamic model chosen is the NRTL. Section 5 extends the application to additional thermodynamic models and other systems of interest. Our main conclusions are in Section 6.

Bold face lowercase letters are used to represent vectors, bold face capital letters for continuous domains, blackboard bold capital letters for discrete domains, and capital letters for matrices. Finite sets containing ι elements are compactly represented by $[\![\iota]\!] = \{1, \cdots, \iota\}$. The symbol "$^\intercal$" is used to indicate the vector/matrix transpose operation, $\mathrm{tr}(\bullet)$ stands for the trace of a matrix and $\mathrm{card}(\bullet)$ for the cardinality of a discrete set.

2. Preliminaries

This section establishes the nomenclature used in the representation of the models and the background required in the framework. In §2.1 we present the VLE model. In §2.2 we present the experimental design problems outlined above, and in §2.3 and Appendix A we establish the fundamentals of the mathematical programming tools used in the algorithm.

2.1. VLE model

We start by introducing the formalism used to describe the phase equilibra. Let $[\![c]\!] = \{1, 2\}$ be the set of components in the mixture, and $[\![h]\!] = \{L, V\}$ (L standing for liquid and V for vapor) the set of phases into which the original mixture separates. Let $\boldsymbol{z} \in \mathbb{R}^2$ be the vector containing the molar fractions in the initial mixture submitted to the ebulliometer, such that $\sum_{j=1}^{2} z_j = 1$. The vector $\boldsymbol{d} \equiv \{\boldsymbol{z}, P, T\}$ contains the possible control factors in the experiments, namely, initial composition of the mixture, pressure and temperature. We note that for isobaric experiments, P (here expressed in $\mathrm{mm\,Hg}$) is fixed and the control factor is T, and for isothermal experiments the opposite occurs. The composition of the initial mixture, limited by the corresponding compositions at the dew point and bubble point, is a variable to be chosen in the experimental plan in agreement with the temperature or pressure. Finally, the molar fraction of one of the components in the initial mixture is not independent, i.e., the choice of z_1 automatically sets z_2.

The molar fraction of the components in each phase is denoted by $y_{p,j}$ where the first subscript identifies the phase, $p \in [\![h]\!]$, and the second is for the component, $j \in [\![c]\!]$. Similarly, in each phase and ignoring any measurement er-

rors, the component fractions sum to 1, i.e. $\sum_{j=1}^{2} y_{p,j} = 1,\ \forall p \in [\![h]\!]$ and $0 \leq y_{p,j} \leq 1,\ \forall p \in [\![h]\!],\ j \in [\![c]\!]$. Consequently, only one of the components is measured in each phase because the molar fraction of the other is not independent. Without loss of generality we consider that component 1 is measured. The vector containing the molar fraction of both components in phase p is denoted by $\boldsymbol{y}_{p,.},\ p \in [\![h]\!]$. Practically, for the most volatile component $y_{L,j} \leq z_j \leq y_{V,j}$, and for the less volatile $y_{V,j} \leq z_j \leq y_{L,j}$. The mathematical representation of the VLE model is detailed in Appendix B.

The measurements of the response variables in model (B.2) include an error component, i.e.

$$y_{p,j}^{\text{obs}} = y_{p,j} + \epsilon_{p,j}, \quad p \in [\![h]\!] \quad \text{and} \quad j \in [\![c]\!], \tag{1}$$

where $y_{p,j}^{\text{obs}}$ is the measurement of the molar fraction of component j in the p^{th} phase and $\epsilon_{p,j}$ is the corresponding observational error. Herein, we consider the errors affecting each of the responses to be normally distributed with zero mean and standard deviation $\sigma_{p,j}$, i.e. $\epsilon_{p,j} \sim \mathcal{N}(0, \sigma_{p,j})$. Problems can arise when assuming that measurement errors are normally distributed because the molar fractions are constrained between 0 and 1. Here, these problems are mitigated because the standard deviation of the errors is small. Without loss of generality, the standard deviation of the observational errors is assumed to be known from the performance of the measurement device and experimental procedure considered.

2.2. Optimal design of experiments

The regression models addressed herein are described by $\boldsymbol{g}(\bullet) = 0$ and have a two-variate response with two control factors, corresponding to the initial mixture composition and pressure (in isothermal experiments) or temperature (in isobaric

experiments). It is noteworthy that the choice of the temperature or pressure influences the choice of the composition of the initial mixture as it must be within the compositions at dew and bubble points. Practically, the initial mixture is a control factor although it is dependent on the other control factor(s). The system response is (i) the molar fraction of one of the components in L phase; or (ii) the composition of one of the components in both L and in V phases. In the latter case, the mean response of the system at conditions $\boldsymbol{d}$ are denoted by $\mathbb{E}[\boldsymbol{y}_{V,\cdot}^{\text{obs}}, \boldsymbol{y}_{L,\cdot}^{\text{obs}} | \boldsymbol{d}, \boldsymbol{\theta}]$, where the expectations of $\boldsymbol{y}_{L,\cdot}^{\text{obs}}$ and $\boldsymbol{y}_{V,\cdot}^{\text{obs}}$ satisfy Equations (B.2–1) for a given vector $\boldsymbol{d}$. Here, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^{n_\theta}$ where $\boldsymbol{\Theta}$ is a compact domain containing the parameter values and $\mathbb{E}[\bullet]$ is the expectation operator with respect to the error distribution.

Because the model $\boldsymbol{g}(\bullet) = 0$ is nonlinear, the Fisher Information Matrix (FIM, Eq. (3)) depends on $\boldsymbol{\theta}$, the value of which is to be estimated from the experimental results [5]. The simplest approach to overcome the interdependence is fixing $\boldsymbol{\theta}$ to a set of values known *a priori*, construct the respective FIM and find the optimal design for such a vector which is commonly designated as a *locally optimal design* [22]. Another strategy is sequential optimal design; initially a locally optimal design is found for a given set of postulated parameters which are subsequently improved using the results of new iteratively designed experiments. We adopt the latter strategy. In the first iteration of the M-bSODE procedure, the *locally optimal* design is found with model parameters taken from previous studies or experiments, which are treated as known in the FIM construction. We denote the initial vector of parameters as $\boldsymbol{\theta}^{\text{loc},0}$ where the number in the superscript is used to identify the iteration.

To generate an initial optimal experimental plan we consider *continuous designs*, denoted by ξ^{cont}, which allocate a weight $w_i \in [0, 1]$ to the i^{th} candidate

experiment $\boldsymbol{d}_i \in \mathbb{X}^{n_q}$, with $\sum_{i=1}^{n_q} w_i = 1$. The weights w_i can also be interpreted as the relative effort at conditions $\boldsymbol{d}_i$ in the complete experimental plan (e.g., if w_1 is 0.1, the number of experiments carried out at conditions $\boldsymbol{d}_1$ is $10\,\%$ of all the experiments sought); many w_i's may be zero in the optimal design if the additional information provided by the corresponding conditions is low (or none). $\mathbb{X}^{n_q} := (\boldsymbol{d}_1^\intercal, \cdots, \boldsymbol{d}_{n_q}^\intercal)$ is the set containing n_q candidate treatments $\boldsymbol{d}_i$ with each element including (i) the initial mixture composition; (ii) pressure (fixed in isobaric setups); and (iii) temperature (fixed in isothermal setups) at which measurements can be taken. Specifically, $\mathbb{X}^{n_q}$ contains n_q discrete points of the continuous (and compact) design space $\mathbf{X} \equiv \mathbf{Z} \times [P^{\min}, P^{\max}] \times [T^{\min}, T^{\max}]$, where $\mathbf{Z}$ is the two-dimensional simplex domain satisfying $\mathbf{Z} \equiv \{z_j \in [0,1]^2 : \; z_1 + z_2 = 1\}$, $[P^{\min}, P^{\max}]$ is the domain for pressure (being a singleton value in isobaric experiments), and $[T^{\min}, T^{\max}]$ is the domain for temperature (being a singleton value in isothermal experiments). Here, $P^{\min}$ and $P^{\max}$ are the minimum and maximum values of pressure allowed by the ebulliometer, and $T^{\min}$ and $T^{\max}$ the minimum and maximum values of temperature. Note that, for each temperature and pressure in $\mathbf{X}$, the model (B.2) allows calculation of the composition of the initial mixture to be used, $\boldsymbol{z}$, as it depends on the composition of phases at equilibrium which in turn depend on T and P.

Consequently, continuous optimal designs are formed by a set of $n_s (\leq n_q)$ support points $\boldsymbol{d}_i$ (different experimental conditions) and respective weights $w_i (> 0)$, where $n_s \geq n_\theta$ is required to ensure non-singular FIMs. The advantages of working with continuous designs are many, and there is a unified framework for finding optimal continuous designs for M-bODE problems when the design criterion is a convex function on the set of all continuous designs [36]. In particular,

the optimal design problem can be formulated as a mathematical optimization program with convex properties. In addition, so-called "equivalence theorems" are available to check the optimality of obtained designs.

Exact designs are experimental plans where the weights w_i are ratios n_i/N satisfying the conditions: (i) all n_i's are integer (or null); and (ii) sum to N. In practice, implementable exact designs are obtained from continuous designs assuming a given number N of experiments in the plan. Herein, they are denoted as ξ_N^{exact}, and are implemented by taking roughly $N \times w_i$ replicates at conditions $\boldsymbol{d}_i,\ i \in [\![n_s]\!]$ after rounding $N \times w_i$ to an integer, subject to the constraint $N \times w_1 + \cdots + N \times w_{n_s} = N$. Rounding procedures for exact designs can be found in Pukelsheim and Rieder [58]. We identify the set Ξ^{cont} of all feasible continuous designs in $\mathbb{X}^{n_q}$ as the n_q-1-dimensional simplex in the space of weights satisfying $\Xi^{\text{cont}} \equiv \left\{\boldsymbol{w} \in [0,1]^{n_q} : \sum_{i=1}^{n_q} w_i = 1 | \mathbb{X}^{n_q}\right\}$. A continuous optimal experimental design is represented by a n_s-tuple

$$\xi^{\text{cont}} := \begin{pmatrix} \boldsymbol{d}_1^\intercal & \cdots & \boldsymbol{d}_{n_s}^\intercal \\ w_1 & \cdots & w_{n_s} \end{pmatrix},$$

where n_s one-point designs have $w_i > 0$. Here, the upper part of the matrix contains the conditions of the experiments, $\boldsymbol{d}_i,\ \forall i \in [\![n_s]\!]$, and the last line the respective weights in the experimental design.

Similarly, the set Ξ_N^{exact} contains all feasible exact designs of total size N in $\mathbb{X}^{n_q}$ satisfying $\Xi_N^{\text{exact}} \equiv \left\{\boldsymbol{n} \in \mathbb{N}_0^{n_q} : \sum_{i=1}^{n_q} n_i = N | \mathbb{X}^{n_q}\right\}$, where $\mathbb{N}_0$ is the set of non-negative integers. An exact experimental design is represented by

$$\xi_N^{\text{exact}} := \begin{pmatrix} \boldsymbol{d}_1^\intercal & \cdots & \boldsymbol{d}_{n_s}^\intercal \\ n_1 & \cdots & n_{n_s} \end{pmatrix}.$$

To systematize the nomenclature, we call the Fisher Information Matrix obtained for a measurement at a candidate point an *elemental* FIM, with a *global* FIM that resulting from combining *all* the elemental FIMs.

The log-likelihood function is [37, Ch. 1]

$$\mathcal{L}(\mathbb{X}^{n_q}, \boldsymbol{\theta}) = \sum_{p \in [\![h]\!]} \sum_{i=1}^{n_q} (\eta^{\text{obs}}_{p,1,i} - \eta_{p,1,i})\; S^{-1}\; (\eta^{\text{obs}}_{p,1,i} - \eta_{p,1,i})^{\intercal}, \tag{2}$$

where S is the (constant) variance-covariance matrix, $\eta^{\text{obs}}_{p,j,i}$ refers to measurements of $y_{p,1}$ at the i^{th} candidate point and $\eta_{p,1,i}$ stands for the respective prediction constructed using model (B.2).

The performance of the design ξ^{cont} is measured by a convex functional of its global FIM. The elements of the normalized FIM obtained after adjusting for the sample size are the negatives of the expectations of the second-order derivatives, with respect to the parameters, of the log-likelihood. For a single response, say the molar fraction of component 1 in phase p, given the set of candidate treatments $\mathbb{X}^{n_q}$, the global partial FIM is proportional to

$$\mathcal{M}(\xi^{\text{cont}}|\mathbb{X}^{n_q}, \boldsymbol{\theta}) = \sum_{i=1}^{n_q} w_i\; M(\boldsymbol{d}_i|\mathbb{X}^{n_q}, \boldsymbol{\theta}) = \sum_{i=1}^{n_q} w_i\; \left(\frac{\partial \eta_{p,1,i}}{\partial \boldsymbol{\theta}}\right)\; S^{-1}\; \left(\frac{\partial \eta_{p,1,i}}{\partial \boldsymbol{\theta}^{\intercal}}\right), \tag{3}$$

where the elemental FIM obtained from measurements at conditions $\boldsymbol{d}_i$ is $M(\boldsymbol{d}_i|\mathbb{X}^{n_q}, \boldsymbol{\theta})$. By assumption all the measurements have the same observational error variance, which yields a diagonal variance-covariance matrix of ones after normalization [28, 76].

When the observational errors are independent and identically distributed (iid), the volume of the confidence region for model parameters $\boldsymbol{\theta}$ is inversely proportional to $\det[\mathcal{M}^{1/2}(\xi^{\text{cont}}|\mathbb{X}^{n_q}, \boldsymbol{\theta})]$. Consequently, maximizing the determinant of

the FIM, by choice of design, leads to the most accurate estimates for the parameters. If interest is in finding the continuous locally D-optimal design, the optimization problem is

$$\xi_D^{\text{cont}} = \arg \max_{\xi^{\text{cont}} \in \Xi^{\text{cont}}} \{\det[\mathcal{M}(\xi^{\text{cont}} | \mathbb{X}^{n_q}, \boldsymbol{\theta})]\}^{1/n_\theta}. \tag{4}$$

To check the optimality of the D-optimal designs found we use an equivalence theorem (ET) derived from directional derivative [49, 74, 5]. For the D-optimality criterion the ET states that the scaled *sensitivity function* at point $\boldsymbol{d} \in \mathbb{X}^{n_q}$, representing the directional derivative, is given by Kiefer and Wolfowitz [49]

$$\Psi(\boldsymbol{d}, \xi^{\text{cont}}) = \left(\frac{\partial \boldsymbol{y}_{L,.}}{\partial \boldsymbol{\theta}} \quad \frac{\partial \boldsymbol{y}_{V,.}}{\partial \boldsymbol{\theta}} \right) \mathcal{M}^{-1}(\xi_D^{\text{cont}} | \mathbb{X}^{n_q}, \boldsymbol{\theta}) \begin{pmatrix} \frac{\partial \boldsymbol{y}_{L,.}}{\partial \boldsymbol{\theta}^\intercal} \\ \frac{\partial \boldsymbol{y}_{V,.}}{\partial \boldsymbol{\theta}^\intercal} \end{pmatrix} - n_\theta, \tag{5}$$

being limited from above by 0, and achieving the maximum at the support points of the design that correspond to $\boldsymbol{d}_i$'s with positive weight.

2.3. Mathematical Programming tools

In this paper semidefinite programming is employed to solve the locally optimal design problem for D-optimality over a given discrete domain $\mathbb{X}^{n_q}$. In turn, Nonlinear Programming (NLP) is employed to fit the data to the thermodynamic model and to solve the problem of finding the subsequent optimal experiment. In §A.1 of the Appendix we overview the fundamentals of SDP and its application in our context.

The problem of calculating a design for a pre-specified set of candidate experiments $\mathbb{X}^{n_q}$ with points $\boldsymbol{d}_i$, $\forall i \in [\![n_q]\!]$, is solved with the general formulation (A.2) complemented by the linear constraints on $\boldsymbol{w}$: (i) $\boldsymbol{w} \geq 0$, and (ii) $\mathbf{1}_{n_q}^\intercal \boldsymbol{w} = 1$, where $\mathbf{1}_{n_q}^\intercal$ is a unitary column vector with n_q rows. The problem (A.2) is the classic SDP problem which includes LMIs representing conic constraints.

3. Sequential optimal design of experiments

Here we introduce the framework proposed for helping in the construction of VLE models from data. First, the complete algorithm is discussed, then each of the steps is analyzed in depth, see §3.1–§3.9.

Figure 1 illustrates the basic sequence of steps of the proposed tool. Steps 1 to 4 find an initial locally optimal experimental design for a postulated vector of parameters, and Steps 5 to 9 provide the iteration of parameter estimation given the information available followed by a tool to find the next (optimal) experiment for the updated vector of parameters. The procedure ends when a given stopping criterion is attained (see Step 9).

[Figure 1 about here.]

This framework for optimal design of experiments is compatible with the traditional experimental VLE workflow described in the Introduction section. In this existing workflow, the new framework is intended to be applied in Steps (iii), for constructing an initial experimental plan, and (vii), for the incremental planning of new experiments, until a stopping condition is reached. Since the overall workflow is maintained, the role of the additional steps is also implicitly assumed. This is important due to the crucial role that other support tasks play, including the use of validation criteria and data consistency tests as is the usual practice in the field. A more detailed description of the individual steps is provided below.

3.1. Discretization of the design space

For isobaric experiments the set of temperatures in candidate experiments $\boldsymbol{d} \in \mathbb{X}^{n_q}$ is constructed from a discretization scheme, say with a constant step, where $T_1 = T^{\min}$ and the other points are determined recursively with a rule $T_j =$

$T_{j-1} + \Delta T_{j-1},\ j = 2, \cdots, n_q$. As the notation suggests, the discretization steps can be unequal and the size of $\mathbb{X}^{n_q}$ (i.e., n_q) can be large. Practically, the choice of n_q is a trade-off between the accuracy required and the size of the SDP problem to be solved in Step 3 which can be prohibitive for very large values. Fixed data points such as the ones corresponding to pure component information and the locations of azeotropes can also be included in this discretization grid as additional sampling locations in the description of the domain. This is especially relevant due to the particular information that these points provide, including sometimes their *a priori* availability with smaller uncertainties associated compared to the remaining measurements to be performed.

3.2. Construction of the global FIM

This section describes the procedure used to construct the global FIM for a candidate set of control factors $\mathbb{X}^{n_q}$ (Step 2). The model that is used to describe the VLE equilibria between phases is represented by a set of equations with the form $g_m(\boldsymbol{y}_{L,\cdot}, \boldsymbol{y}_{V,\cdot}|\boldsymbol{d}, \boldsymbol{\theta}) = 0$, where $\boldsymbol{\theta}$ is the parameter vector evaluated at candidate solutions $\boldsymbol{\theta}^{\text{loc},0}$; this is detailed in Appendix B. In the application of the framework, no particular assumptions relative to the form of $g(\cdot)$ are required, allowing this procedure to be applicable to alternative models for phase behavior description. The nonlinear algebraic equations in (B.2) are solved for each candidate treatment in $\mathbb{X}^{n_q}$ as a constrained nonlinear algebraic system (CNS), using for example the `CONOPT` NLP solver, a Generalized Reduced Gradient (GRG) algorithm [29]. With the examples considered, relative and absolute tolerances of 1×10^{-7} and 1×10^{-8} were used, respectively.

The chain rule for differentiation with respect to the parameters applied to this

set of equations leads to

$$\sum_{p\in[\![h]\!]} \frac{\partial g_m(\boldsymbol{y}_{L,.}, \boldsymbol{y}_{V,.}|\boldsymbol{d}_i, \boldsymbol{\theta})}{\partial y_{p,1}}\, s_{p,1,i} + \frac{\partial g_m(\boldsymbol{y}_{L,.}, \boldsymbol{y}_{V,.}|\boldsymbol{d}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0,\ \forall m \in [\![8]\!],$$
$$\boldsymbol{d}_i \in \mathbb{X}^{n_q} \tag{6}$$

where $s_{p,1,i}$ denotes the vector sensitivities of the response $y_{p,1}$ to $\boldsymbol{\theta}$ at $\boldsymbol{d}_i \in \mathbb{X}^{n_q}$. The partial derivatives $\partial g_m(\boldsymbol{y}_{L,.}, \boldsymbol{y}_{V,.}|\boldsymbol{d}_i, \boldsymbol{\theta})/\partial y_{p,1}$ and $\partial g_m(\boldsymbol{y}_{L,.}, \boldsymbol{y}_{V,.}|\boldsymbol{d}_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ are determined via automatic differentiation with `ADiMat` [12].

The set of algebraic equations (6) can be written as

$$F_{p,1,i}\ s_{p,1,i} + r_{p,1,i} = 0, \quad p \in [\![h]\!],\ i \in [\![n_q]\!] \tag{7}$$

where $F_{p,1,i}$ are $n_\theta \times n_\theta$ matrices formed by terms $\partial g_m(\boldsymbol{y}_{L,.}, \boldsymbol{y}_{V,.}|\boldsymbol{d}_i, \boldsymbol{\theta})/\partial y_{p,1}$, and $r_{p,1,i}$ column vectors of size n_θ containing the terms $\partial g_m(\boldsymbol{y}_{L,.}, \boldsymbol{y}_{V,.}|\boldsymbol{d}_i, \boldsymbol{\theta})/\partial \theta$. Consequently, (7) can be solved as

$$s_{p,1,i} = -F_{p,1,i}^{-1}\, r_{p,1,i}, \quad p \in [\![h]\!],\ i \in [\![n_q]\!], \tag{8}$$

provided that $F_{p,1,i}$ is invertible. Here, $s_{p,1,i}$ are column vectors of size n_θ corresponding to $\partial y_{p,1}/\partial \boldsymbol{\theta},\ p \in [\![h]\!]$ at the i^{th} candidate treatment. In this case, the $n_\theta \times n_\theta$ elemental FIMs can be generated using the relation

$$M(\boldsymbol{d}_i|\mathbb{X}^{n_q}, \boldsymbol{\theta}) = s_{p,1,i}\, S^{-1}\, s_{p,1,i}^{\intercal}, \quad \forall \boldsymbol{d}_i \in \mathbb{X}^{n_q},\ p \in [\![h]\!], \tag{9}$$

which allows the computation of $\mathcal{M}(\xi^{\text{cont}}|\mathbb{X}^{n_q}, \boldsymbol{\theta})$ using Eq. (3).

3.3. Finding locally D-optimal continuous designs

This section presents the formulation for finding locally D-optimal designs via SDP (Step 3 of the algorithm). We notice that the SDP can be computationally

challenging if the number of candidate experiments is large, although it assures that the global optimum is found. Methods based on the Wynn-Fedorov algorithm [77] and subsequent improvements [24] or in the KL algorithm [5] may be lighter, but they require successive reinitializations and the global optimality is not guaranteed.

We recall the continuous D-optimal design problem (4), and use the formulation proposed by Vandenberghe and Boyd [71] to handle the problem. That is, given the VLE model, the set of local parameters, $\boldsymbol{\theta}^{\text{loc},0}$, the set of candidate experiments $\mathbb{X}^{n_q}$ (already generated) and the elemental FIMs for each candidate experiment, the problem for finding continuous D-optimal designs on Ξ^{cont} may be formulated as:

$$\max_{\boldsymbol{w}\in[0,1]^{n_q}} \ \left\{\det\left[\mathcal{M}(\xi^{\text{cont}}|\mathbb{X}^{n_q},\boldsymbol{\theta}^{\text{loc},0})\right]\right\}^{1/n_\theta} \tag{10a}$$

$$\text{s.t. } \mathcal{M}(\xi^{\text{cont}}|\mathbb{X}^{n_q},\boldsymbol{\theta}^{\text{loc},0}) = \sum_{i=1}^{n_q} w_i\ M(\boldsymbol{d}_i|\mathbb{X}^{n_q},\boldsymbol{\theta}^{\text{loc},0}) \tag{10b}$$

$$\mathcal{M}(\xi^{\text{cont}}|\mathbb{X}^{n_q},\boldsymbol{\theta}^{\text{loc},0}) \in \mathbb{S}_+^{n_\theta} \tag{10c}$$

$$\mathbf{1}_{n_q}^\intercal\ \boldsymbol{w} = 1 \tag{10d}$$

Here, Equation (10a) is the objective function, (10b) generates the global FIM from elemental FIMs, (10c) is for the requirement of its semidefinite positiveness, $\mathbb{S}_+^{n_\theta}$ is the space of semipositive definite matrices, and (10d) guarantees that the weights of the one-point designs sum to 1. We note the decision variable, $\boldsymbol{w}$, is included in the design ξ^{cont} (cf. §2.2).

To handle the SDP problems (10), there are user-friendly interfaces, such as `cvx` [44] or PICOS [63], that automatically transform the semidefinite constraint (10c) and the objective function into a series of Linear Matrix Inequalities (see

Appendix A.1) before passing them to SDP solvers such as `SeDuMi` [67] or `Mosek` [2]. This is possible when the design criterion is SDr, which is true for D-optimality. In our work, we solved all SDP problems using the `cvx` environment combined with the solver `Mosek` that uses an efficient interior point algorithm [78]. The relative and absolute tolerances used to solve the SDP problem were set to 1×10^{-5}.

3.4. Building D-optimal exact designs

Here, we detail the procedure to generate exact D-optimal designs, ξ_N^{exact}, from continuous designs ξ^{cont} (Step 4).

We assume the number of experiments in the initial phase (Step 5) is equal to the number of support points of the continuous design, and use the rounding procedure proposed by Pukelsheim and Rieder [58] to generate the corresponding exact design. Note that the number of support points of the exact design may be lower than the number of support points of the continuous design when $n_i = 0$ for support point i. This implies that one or more other support points have increased weight compared to those of the continuous design. The FIMs for exact designs are updated with Equation (3). Specifically, the weights w_i are replaced by the ratios $n_i/N,\ n_i \in \mathbb{N}_0$. We call $\mathbb{C}^{n_s}(\subseteq \mathbb{X}^{n_q})$ the set of n_s design points forming the exact experimental plan prescribed in this phase.

3.5. Performing the experiments

During Step 5, the standing experimental sampling locations are evaluated. After this, the results are added to the complete data set available for model regression, and the framework proceeds at the next step.

3.6. Estimating the parameters

This section presents the strategy for estimating the model parameters from data collected from previous experiments, corresponding to Step 6 of the algorithm. In this step, we use all the information (data) previously obtained (i.e. from experiments of the initial design as well as from sequential experiments).

For simplification we consider the first iteration, i.e. $k = 1$ with k used to count the iterations. Specifically, after carrying out the experimental plan prescribed in Step 4 (comprising n_s experiments) where the responses are the molar fractions of component 1, $\eta^{\text{obs}}_{p,1,i},\ p \in [\![h]\!],\ i \in [\![n_s]\!]$, with the subscript i identifying the experiment, the data are used for fitting the model and estimating the parameters. Moreover, let $\eta^{\text{e}}_{L,1,i}$ and $\eta^{\text{e}}_{V,1,i},\ i \in [\![n_s]\!]$, be the model estimates of the equilibrium fractions in both phases at experimental conditions $\boldsymbol{d}_i \in \mathbb{C}^{n_s}$.

The objective function adopted to fit the data is the Ordinary Least Squares (OLS) criterion as we assume homoscedastic observational error. The model fitting problem falls into the class of implicit least squares with the estimated values of the responses being calculated from a model embedded in the mathematical program that maximizes the log-likelihood. The problem is formulated as

$$\min_{\boldsymbol{\theta}\in\Theta} \sum_{p\in[\![h]\!]} \sum_{i=1}^{n_s} \left(\eta^{\text{e}}_{p,1,i} - \eta^{\text{obs}}_{p,1,i}\right)^2 \tag{11a}$$

$$\text{s.t. } \eta^{\text{e}}_{V,j,i} - \eta^{\text{e}}_{L,j,i}\,\gamma_j(\boldsymbol{\eta}^{\text{e}}_{L,.,i}, T_i, \boldsymbol{\theta})\,\frac{P^v_j(T_i)}{P_i} = 0, \quad j \in [\![c]\!], \quad i \in [\![n_s]\!], \tag{11b}$$

$$\sum_{j=1}^{2} \eta^{\text{e}}_{p,j,i} = 1, \quad p \in [\![h]\!], \quad i \in [\![n_s]\!], \tag{11c}$$

$$P^v_j(T_i) = 10^{A_j - B_j/(T_i + C_j)}, \quad j \in [\![c]\!], \quad i \in [\![n_s]\!], \tag{11d}$$

$$(T_i,\ P_i) \subset \boldsymbol{d}_i \in \mathbb{C}^{n_s} \tag{11e}$$

where (11a) is the objective function, (11b–11d) are model equations and (11e) is the set of experimental conditions used. The problem (11) is in the NLP class and we again use CONOPT, with relative and absolute tolerances set to 10^{-7} in all problems.

3.7. Checking the need for further experiments

After finding the (optimal) parameter estimates by solving problem (11) for k^{th} iteration, Step 7 requires the decision of whether or not to carry out another experiment. Common criteria to use in this step are (i) the total number of experiments of the plan which may be limited by economic constraints; and (ii) whether the dissimilarity of parameter estimates in two consecutive iterations is lower than a tolerance, ϱ, previously imposed, i.e.

$$\left(\boldsymbol{\theta}^{\text{loc},k} - \boldsymbol{\theta}^{\text{loc},k-1}\right)^{\mathsf{T}} \left(\boldsymbol{\theta}^{\text{loc},k} - \boldsymbol{\theta}^{\text{loc},k-1}\right) \leq \varrho. \tag{12}$$

When the iterations in the cyclic part of the procedure exceed the maximum previously set, $N_{\text{it}}^{\max}$, or the condition (12) is satisfied, the procedure stops; otherwise the global FIM for the total set of experiments available is updated for parameters $\boldsymbol{\theta}^{\text{loc},k}$ (Step 8). Next, a new (optimal) experiment at conditions $\boldsymbol{d}_{n_s+k}$, corresponding to the k^{th} iteration, is determined (Step 9) and Steps 6–9 are repeated. When the first decision rule is adopted, the procedure stops when k reaches $N_{\text{it}}^{\max}$, previously set.

3.8. Update the elemental and global FIMs

Let $\boldsymbol{\theta}^{\text{loc},1}$ be the optimum set of parameters estimated from the solution of (11). Likewise, for stage k of the experimental plan, which involves data from $n_s + k - 1$ experiments, we obtain the parameter estimate $\boldsymbol{\theta}^{\text{loc},k}$. The sequence of

parameter vectors accumulated in the iteration procedure tends to the true values of the model, see Fedorov and Leonov [37, Chap. 8]. In our study, the true values of the parameters are those used for simulating the expected responses (without observational error), see §3.5. After finding the vector $\boldsymbol{\theta}^{\text{loc},1}$, the FIMs are updated using the same strategy as for Step 2. Practically, in this Step the FIMs are updated using the most up-to-date parameter estimates obtained in Step 6 and all the data obtained so far.

3.9. Find the next experiment

We now formulate the problem of finding the next experiment to be carried out in the plan. We consider the current estimate of parameters is $\boldsymbol{\theta}^{\text{loc},1}$,when the global FIM obtained in Step 7 is $\mathcal{M}(\xi_{n_s+1}^{\text{exact}}|\mathbb{C}^{n_s}, \boldsymbol{\theta}^{\text{loc},1})$. That is, the problem is solved for the most up-to-date parameter estimates.

We adopt the ideas of Box and Hunter [14] and Fedorov [35] who formulated the problem of finding the next experiment as the search for the point of the design space with the largest increment of the optimality criterion of interest. Such a point is the maximum of the directional derivative represented by the sensitivity function. For D-optimality, the problem corresponds to finding the maximum of (5), i.e.:

$$\max_{\boldsymbol{d}} \begin{pmatrix} \frac{\partial \boldsymbol{y}_{L,.}}{\partial \boldsymbol{\theta}^{\text{loc},1}} & \frac{\partial \boldsymbol{y}_{V,.}}{\partial \boldsymbol{\theta}^{\text{loc},1}} \end{pmatrix} \mathcal{M}^{-1}(\xi_{n_s+1}^{\text{exact}}|\mathbb{C}^{n_s}, \boldsymbol{\theta}^{\text{loc},1}) \begin{pmatrix} \frac{\partial \boldsymbol{y}_{L,.}}{\partial (\boldsymbol{\theta}^{\text{loc},1})^{\intercal}} \\ \frac{\partial \boldsymbol{y}_{V,.}}{\partial (\boldsymbol{\theta}^{\text{loc},1})^{\intercal}} \end{pmatrix} \tag{13a}$$

$$\text{s.t. } y_{V,j} - y_{L,j}\, \gamma_i(\boldsymbol{y}_{L,.}, T, \boldsymbol{\theta}^{\text{loc},1})\, \frac{P_j^v(T)}{P} = 0, \quad j \in [\![c]\!] \tag{13b}$$

$$\sum_{j=1}^{2} y_{p,j} = 1, \quad p \in [\![h]\!] \tag{13c}$$

$$P_j^v(T) = 10^{A_j - B_j/(T+C_j)}, \quad j \in [\![c]\!] \tag{13d}$$

$$z_1 - [\omega\, y_{V,1} + (1-\omega)\, y_{L,1}] = 0 \tag{13e}$$

$$\sum_{j=1}^{2} z_j - 1 = 0 \tag{13f}$$

$$\text{Equations (3) and } (6-9) \tag{13g}$$

We recall that the decision variable is the vector of experimental conditions $\boldsymbol{d} \in \mathbf{Z} \times \{P\} \times [T^{\min}, T^{\max}]$ where P is fixed. Equations (13b–13f) represent the model, and the set of equations in (13g) is to compute sensitivities and the global FIM.

Let $\boldsymbol{d}_{n_s+k}$ be the solution of problem (13) obtained from the k^{th} iteration. Afterwards, the matrix form of the experimental design is updated using the rule

$$\mathbb{C}^{n_s+k} := \begin{pmatrix} \mathbb{C}^{n_s+k-1} \\ \boldsymbol{d}_{n_s+k} \end{pmatrix}.$$

Problem (13) is a non convex NLP and we use a multistart heuristic algorithm-based solver `OQNLP` [70] to solve it. Similarly, `ADiMat` is employed to generate gradient and Hessian information required by the solver via automatic differentiation.

4. Results

In this section, we apply the formulations of §3 to find locally D-optimal designs for VLE characterization using experiments at constant pressure. Without loss of generality, we consider that only the fraction of component 1 in the L phase at equilibrium ($y_{L,1}$) is measured, which is the most common experimental setup.

For demonstration purposes we consider the system formed by methanol (also designated MET — component 1) and water (designated WAT — component 2).

The VLE of this binary system is well characterized and allows analyzing the details of the framework without the need of considering specificities that may increase the difficulty such as the occurrence of azeotropic points. The VLE of more complex binary systems is considered in §5.

4.1. Generation of test data

For a more precise evaluation of the effect of the introduction of optimal sampling policies, while avoiding potential simultaneous influences from other sources, synthetic test data that does not correspond to actual experimental measurements were generated.

For each physical system, this data was obtained through the simulation of the equilibrium curves at a number of discrete points, using the same equilibrium model that is later used for parameter regression. Thus, in the absence of any numerical errors, the values of the parameters obtained through regression of this sampled information would be identical to the parameters used during the simulation phase, and this would occur for any sampling plan that included a sufficient number of data points (that is, with a practically identifiable model).

To avoid this trivial behavior, Gaussian random noise was added to the phase compositions predicted by each model to provide us with "simulated" experimental measurements. The (P,T) values included in the data points correspond exactly to the values that were also considered for simulation (no uncertainty). Due to the setup used, no noise was added to the measurements corresponding to the pure component values, and no additional sources of systematic errors were also considered. Hence, this type of data can be characterized as a best case scenario, where only random errors are present and their magnitude is controlled, since less problematic data would only be achievable through further reductions in the

magnitude of the random noise considered.

Thus when simulating the procedure for illustrative purposes we solve the model (B.2) for $\boldsymbol{\theta}^{\text{true}}$ using the set of prescribed experiments, where $\boldsymbol{\theta}^{\text{true}}$ stands for the set of parameters that model the data. Of course in applications $\boldsymbol{\theta}^{\text{true}}$ is not known but is to be estimated from the experimental results.

To account for the observational error we add a stochastic component modeled by the normal distribution to the responses calculated using $\boldsymbol{\theta}^{\text{true}}$. Then (1) becomes

$$y_{p,1}^{\text{obs}} = y_{p,1}^{\text{sim}} + \epsilon_{p,1}, \quad p \in [\![h]\!] \tag{14}$$

where $y_{p,1}^{\text{obs}}$ stands for the measurement of $y_{p,1}$ and $y_{p,1}^{\text{sim}}$ for the simulated value excluding observational error. When a single response is measured, say $y_{L,1}$, Equation (14) applies only to the molar fraction of component 1 in the liquid phase. For illustration in this study, see §4, we consider the observational error normally distributed, and different standard deviations are tested, including (i) absence of noise; (ii) $\sigma_{L,1} = 0.001\,\text{mol/mol}$, compatible with recent ebulliometer readings [21]; and (iii) $\sigma_{L,1} = 0.002\,\text{mol/mol}$.

4.2. Test implementation

With this system, the NRTL model (Appendix C) is used for describing the thermodynamic equilibrium where the BIPs $\tau_{i,j},\ i,j \in [\![c]\!]$ are dependent on parameters $a_{i,j}$ and $b_{i,j}$ according to Eq. (B.1). The set of parameters to be estimated from the experimental data is $\boldsymbol{\theta} \equiv \{a_{1,2},\ a_{2,1},\ b_{1,2},\ b_{2,1}\}$. The Antoine constants for both components are in Table 1 and the true values of binary interaction parameters used for simulation are in Table 2. The parameters $\boldsymbol{\alpha}$ are chosen previously to model fitting and are not to be estimated from the experimental plan;

for simulation purposes we use the values of $\boldsymbol{\alpha}$ in Table 2. We assume the operation pressure, P, is $760\,\mathrm{mm\,Hg}$ and $\omega = 0.3$; consequently, the design space is $\mathbf{X} \equiv \mathbf{Z} \times \{760\} \times [323.15,\ 374.15]$.

[Table 1 about here.]

[Table 2 about here.]

Instead of distributing the grid of candidate temperatures uniformly over the design space, we adopt a different approach where the initial grid of candidate temperatures results from distributing the points such that the increments of $y_{L,1}$ in consecutive experiments in the grid are nearly equal. This technique focuses on the region of the design space where both phases coexist. We used increments of $y_{L,1}$ equal to 0.025 and determine the equilibrium conditions for each molar fraction; the complete set of candidate experiments contains 41 temperatures; consequently $\mathbb{X}^{n_q}$ includes 41 values ($n_q = 41$). These choices can be properly adapted to different systems, as required. Typically, the CPU time required by the SDP solver used for determining the initial continuous design increases as the grid becomes finer, but the solver can efficiently tackle problems including thousands of LMIs. In this experiment the CPU time required for finding the initial continuous locally optimal design is around $6\,\mathrm{s}$ in an Intel Core i7 machine running a 64 bits Windows 10 operating system with a $2.80\,\mathrm{GHz}$ processor.

For the sake of clarity, we call the BIPs in Table 2 the "true" parameters, $\boldsymbol{\theta}^{\text{true}}$. These values are used for simulation purposes. For mimicking the lack of knowledge at the beginning of the VLE characterization study, we assume the values of $a_{i,j}$ and $b_{i,j},\ i,j \in [\![c]\!]$, used for finding the continuous D-optimal design are $10\,\%$ lower than those in Table 2, except for $\alpha_{i,j},\ i,j \in [\![c]\!]$, which are equal

to the values used in the simulation. Practically, the temperature of the candidate experiments is determined for $\boldsymbol{\theta}^{\text{loc},0} = 0.9 \times \boldsymbol{\theta}^{\text{true}}$. Figure 2(a) shows the VLE envelope and the candidate experiments corresponding to all the temperatures as well as the initial composition.

[Figure 2 about here.]

The continuous optimal design obtained for the setup described above includes 4 support points and is listed in Table 3. The vectors $\boldsymbol{d}_i^{\intercal}$ appear in the first four lines (first the composition of the initial mixture, then the pressure, and finally the temperature). The fraction of MET in the liquid phase at equilibrium is in fifth line, its fraction in the vapor phase at equilibrium in sixth line, and the weights of the support points in seventh line. We note the lowest temperature of the experiments is $339.73\,\text{K}$ and the highest is $364.13\,\text{K}$. The representation of the optimal design in the VLE envelope appears in Figure 2(b) and we notice that (i) the design is minimally supported as it contains a number of support points, n_s, equal to the number of parameters ($n_\theta = 4$) given that a single measurement is obtained per experiment; and (ii) the design points closer the extremes of the design space have higher weights.

The exact optimal design, is obtained from ξ^{cont} using the rounding approach of §3.4. We set the number of experiments in this initial phase to n_s. The optimal design obtained is in the second half of Table 3. Because its structure is similar to ξ^{cont} we limit its presentation to first and last line, which now contains the number of experiments at each design point.

[Table 3 about here.]

Now we address the iterative procedure (Steps 5-9) of updating the parameter

estimates and accumulating new responses from locally optimal sequential experiments. The decision of carrying out (or not) the next experiment prescribed in Step 7 is taken based on the maximum number of experiments in the iteration phase, which is previously set, see Step 7 of the algorithm. To analyze the convergence of the parameter estimates from data obtained from previous experiments, $\boldsymbol{\theta}^{\mathrm{loc},k}$, to the "true" parameter values used for response generation, $\boldsymbol{\theta}^{\mathrm{true}}$, we set $N_{\mathrm{it}}^{\max} = 5$. Since the number of experiments of the initial D-optimal design in this example is 4, the complete plan will have a total of 9 experiments. We suppose we are in stage k where $k \in [\![5]\!]$.

First, we assume there is no observational error in $y_{L,1}$, that is $\sigma_{L,1} = 0\,\mathrm{mol/mol}$. Next, we address more realistic scenarios where noise is added to measurements and Equation (14) is employed to simulate experimental responses. Here, two scenarios are considered: (i) $\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$; and (ii) $\sigma_{L,1} = 0.002\,\mathrm{mol/mol}$. The values of the parameter estimates in successive iterations of the procedure are in Figure 3. Stage 0 corresponds to initial parameter guesses. We note that after the first iteration the estimates obtained for $\sigma_{L,1} = 0\,\mathrm{mol/mol}$ by fitting the data agree with the values used in the simulation, as they should in the absence of observational error. This result demonstrates the accuracy of our routines. A different behavior is observed for other scenarios where noise induces small fluctuations in the estimates of the parameters around the values used to generate the responses. As an aid to more exact analysis, the sequences of parameter estimates in Figure 3 are repeated in Table 4. The evolution of the estimates of $b_{1,2}$ and $b_{2,1}$ shows convergence to the "true" values. Increased accuracy in the estimation of $a_{i,j}$'s can be achieved using the $\mathrm{D_s}$-optimality criterion which allows determining a subset of parameters as precisely as possible

[5].

[Figure 3 about here.]

[Table 4 about here.]

Table 5 presents the experiments prescribed in the iterative phase of the procedure. New experiments complement the design of the initial phase, see Table 3 and include points that are nearly replicates. Because, in the initial design (Table 3) the weights of the support points were flattened by rounding, these replications increases the efficiency of the design towards that of the initial continuous design. The sequential optimal design is similar for all observation error distributions, except from the change of order for $k = 2$ and 3, indicating slight dependence of the procedure on noise at this level of magnitude. Figures 4(a) and 4(b) present the complete set of 9 experiments when $\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$, and $\sigma_{L,1} = 0.002\,\mathrm{mol/mol}$, respectively. The experiments of the initial design are represented by dashed lines and those of the iterative phase by solid lines.

[Table 5 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

The complete numerical experiment reported in this section requires about $70\,\mathrm{s}$ of CPU time where each iteration of Steps 5–9 requires on average $12\,\mathrm{s}$.

Finally we compare the performance of the optimal experimental plan prescribed with M-bSODE after 9 experiments with that resulting from distributing the experiments uniformly in the domain of z_1, i.e. $[0, 1]$. The parameters

were fitted and $95\,\%$ confidence regions were obtained for each value of $\sigma_{L,1}$, using the observed FIM employing standard asymptotic theory [65]. Plots of the confidence regions for pairs of parameter values showed that regions for the sequential experiment always lay within those for the uniform experimental design. We summarize these results in Table 6 as $95\,\%$ confidence intervals appear after each estimated value and the sign "$\pm$". We note the confidence intervals resulting from the M-bSODE algorithm are lower for both $\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$ and $\sigma_{L,1} = 0.002\,\mathrm{mol/mol}$, being similar (and approximately null) in the absence of noise. On average the confidence limits obtained with M-bSODE algorithm are $27.6\,\%$ tighter when $\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$ and $36.0\,\%$ tighter when $\sigma_{L,1} = 0.002\,\mathrm{mol/mol}$, demonstrating the advantages of using M-bSODE. We note the parameters confidence intervals can be used as an alternative criterion of stopping the experimentation in Step 9, i.e., the iterative sequence of experiments may end when the $100 \times (1-\alpha)\,\%$ confidence intervals of the parameters achieve a previously set value expressed in absolute or relative form. Our results show that a limited number of experiments optimally designed (i.e., 9) allows achieving a parameter accuracy similar to that resulting from more extensive experimental plans designed by distributing the experiments uniformly in the operating space. As can be observed, the model structure together with the information obtained in previous experiments guide the choice of the next experiment location, allowing relevant savings in time and other experimental resources.

[Table 6 about here.]

5. Application to other binary systems

This section applies the framework presented in §3 to binary systems with more complex VLE. To demonstrate the generalization of the algorithm proposed relative to the thermodynamic model both NRTL and UNIQUAC models are used for L phase. Additionally, the following four VLE systems were considered:

1. Ethanol - Water, L phase modeled with NRTL;
2. Water - 7-methyl-1,5,7-triazabicyclo[4.4.0]dec-5-ene (mTBD), L phase modeled with NRTL;
3. Methanol - Water, L phase modeled with UNIQUAC; and
4. Ethanol - Water, L phase modeled with UNIQUAC.

In all the tests, the observational noise affecting the measurements is normally distributed with $\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$ and the variable observed is the concentration of component 1 in L phase. As in the MET-WAT system, the initial grid of candidate experiments is formed by 41 points. For demonstration purposes, the number of experiments in the iterative phase is set to 5. However, the number of experiments in the initial design is prescribed by Steps 3 and 4 of the framework, and so varies from example to example.

5.1. System: Ethanol-Water; L phase model: NRTL

The first system considered for testing is Ethanol—Water (ETH-WAT), where Ethanol is component 1 and Water component 2. Because the binary ETH-WAT is azeotropic, additional thermodynamic insights are needed to generate candidate

experiments. Specifically, the azeotrope is first determined using the global optimization approach proposed by Bonilla-Petriciolet et al. [13]. Here, the orthogonal derivatives of the Gibbs free energy, the tangent plane criterion and the mass balances were used to find the homogeneous azeotropic point. After computing the azeotrope, the set of candidate experiments is generated using the methodology in §3.1. That requires generating a set $\boldsymbol{d} \in \mathbb{X}^{n_q}$ from a grid of candidate experimental temperatures. The equilibrium compositions and the initial mixture are determined using model (B.2) and the information of the azeotropic point. First, we use the NRTL model for describing the equilibrium; in §5.4 the UNIQUAC model is considered.

The Antoine constants for ETH and WAT were obtained from [56]. The extended model (B.1) is used for BIPs. The M-bSODE is determined for singleton parameter vector formed by $a_{1,2} = -0.801$, $a_{2,1} = 3.458$, $b_{1,2} = 246.2\,\mathrm{K}$, $b_{2,1} = -586.1\,\mathrm{K}$, and $\alpha_{1,2} = \alpha_{2,1} = 0.3$ [11, Appendix B]; i.e., $\boldsymbol{\theta}^{\mathrm{loc},0} = \{a_{1,2}, a_{2,1}, b_{1,2}, b_{2,1}, \alpha_{1,2}, \alpha_{2,1}\}$.

Figure 6(a) shows the VLE envelope and the candidate experiments corresponding to all the temperatures for $\boldsymbol{\theta}^{\mathrm{loc},0}$. The initial ODE prescribed is formed by 9 points, more than that initially prescribed for MET-WAT system (4), see Table 7. This increment is related to the behavior of the sensitivities on each side of the azeotropic composition. In Table 7 the horizontal line separates the experiments prescribed in the first stage (above the line) from those of the iterative procedure (below the line). This representation is adopted for similar tables from this point on. Figure 6(b) combines graphically the set of experimental conditions to be carried out in both phases. The prescribed experiments lie on both sides of the azeotropic point, see Figure 5(b).

[Figure 6 about here.]

[Table 7 about here.]

Table 8 contains the estimates of the BIPs and respective $95\,\%$ confidence intervals for ETH-WAT system. The M-bSODE approach is advantageous relative to uniform choice but the difference is relatively small. Specifically, the corresponding parametric variances are in average $6.03\,\%$ smaller. It should be noted that some of the experiments chosen by M-bSODE are in the vicinity of the azeotropic point.

[Table 8 about here.]

5.2. System: Water–7-Methyl-1,5,7-triazabicyclo[4.4.0]dec-5-ene; L phase model: NRTL

Here, we consider the system formed by Water and ionic liquid 7-Methyl-1,5,7-triazabicyclo[4.4.0]dec-5-ene; the first is component 1 (designated WAT), the second is component 2 (mTBD). This system is of large interest in solid-liquid and liquid-liquid extraction operations and is commonly used to extract valuable chemicals from raw materials [8]. The VLE characterization places additional challenges on experimental work as it requires operating the ebulliometer at relatively high temperature; M-bSODE might increase experimental efficiency.

The Antoine constants for water were obtained from [56], and those for mTBD were obtained regressing the data in Baird et al. [7]. The data were limited to $451.2\,\mathrm{K}$ which is also the limiting temperature in our simulation. The values of Antoine constants obtained for vapor pressure of mTBD expressed in $\mathrm{mm\,Hg}$ are $A = 7.548\,82$, $B = 2181.15\,^{\circ}\mathrm{C}$ and $C = 207.26\,^{\circ}\mathrm{C}$. The VLE is described

by the NRTL model, and the BIPs also follow the extended model (B.1). The optimal experimental design is to be determined for the parameter vector with $a_{1,2} = 0.352\,13$, $a_{2,1} = -1.219\,66$, $b_{1,2} = -857.380\,\mathrm{K}$, $b_{2,1} = 978.478\,\mathrm{K}$, and $\alpha_{1,2} = \alpha_{2,1} = 0.3$ [8].

Figure 7(a) shows the candidate experiments for $\boldsymbol{\theta}^{\mathrm{loc},0}$. The initial ODE prescribed is formed by 5 points, see Table 9 and Figure 7(b). Table 10 compares the estimates of parameters and the corresponding $95\,\%$ confidence intervals obtained with M-bSODE and uniform choice. The average estimates are similar but the size of the uncertainty region of the parameters obtained from M-bSODE is considerably tighter. The average reduction in parametric variance is $39.6\,\%$.

[Figure 7 about here.]

[Table 9 about here.]

[Table 10 about here.]

5.3. System: Methanol—Water; L phase model: UNIQUAC

To demonstrate the ability of the framework to handle different thermodynamic models we now consider the MET-WAT system with the UNIQUAC representation. The UNIQUAC model [1] is in Appendix C, and the design of experiments is sought to find the BIPs, $\tau_{i,j},\ i,j \in [\![c]\!]$, given by

$$\tau_{i,j} = \exp\left(a_{i,j} + \frac{b_{i,j}}{T}\right), \tag{15}$$

a structure similar to that of the NRTL model, except for the $\boldsymbol{\alpha}$'s which are omitted. The values of $\tau_{i,i},\ i \in [\![c]\!]$ are set to 1.0. The parameters to be fitted from the experiments are aggregated into $\boldsymbol{\theta} \equiv \{a_{1,2},\ a_{2,1},\ b_{1,2},\ b_{2,1}\}$. In the numerical

tests, the setup used for NRTL model is replicated. Similarly, the uncertainty of the model predictions of experiments prescribed with M-bSODE are compared to experiments prescribed using uniform choice.

The Antoine constants for both components are in Table 1. The locally D-optimal design is found for $\boldsymbol{\theta}^{\text{loc},0} = \{a_{1,2},\ a_{2,1},\ b_{1,2},\ b_{2,1}\}$ where $a_{1,2} = -1.0622$, $a_{2,1} = 0.6437$, $b_{1,2} = 432.8785\,\text{K}$ and $b_{2,1} = -322.1312\,\text{K}$ [3]. The design prescribed by SDP has 5 experiments (one more than that obtained for NRTL model, see Table 3). The complete set of experiments is in Table 11. The experiments concentrate in the vicinity of $341\,\text{K}$ and $360\,\text{K}$. Figure 8(a) displays the experiments in the phase diagram.

[Table 11 about here.]

Table 12 compares the estimates of parameters obtained with M-bSODE and uniform choice, and the M-bSODE framework also allows reducing the parametric uncertainty. The average parametric variance reduction is $20.4\,\%$

[Table 12 about here.]

5.4. System: Ethanol—Water; L phase model: UNIQUAC

Finally, in this section we test the framework for the ETH-WAT system using the UNIQUAC model. This binary mixture is more challenging because of the azeotropic point. The Antoine constants for both components are in §5.1. The locally D-optimal design is constructed for singleton $\boldsymbol{\theta}^{\text{loc},0}$ where $a_{1,2} = 2.0046$, $a_{2,1} = -2.4936$, $b_{1,2} = -728.9705\,\text{K}$ and $b_{2,1} = 756.9477\,\text{K}$ [3]. The initial design prescribed has 8 experiments, see Table 13. Similarly to the NRTL model, the experiments concentrate in the vicinity of the azeotropic point; see Figure 8(b).

[Table 13 about here.]

[Figure 8 about here.]

Table 14 compares the estimates of parameters; the M-bSODE framework reduces the parametric uncertainty after 13 experiments compared to experiments uniformly chosen. The average decrease in the parametric uncertainty is $70.9\,\%$.

[Table 14 about here.]

More significant reductions in the confidence intervals for the parameters of the UNIQUAC model were observed in this case, compared with the previous examples. An interpretation of this behavior can be found in the nature of the model used, which is based on the description of the excess Gibbs free energy of the mixture. Similarly to the NRTL model, the residual term is often obtained as a sum of contributions of approximately equal size but opposite sign, which cancel most of their magnitudes like the forces on the extremes of a cable under tension. In the present case, this becomes quite visible in the values presented in Table 14, where the corresponding parameters for each species almost cancel out. When this situation occurs, the presence of multiple local optima in the regression of these model equations are also quite frequent, and this means that the model displays a significant sensitivity to the data values considered. In this example this behavior is exploited by the M-bSODE framework to significantly reduce the uncertainty in the estimated parameters through the optimal choice of the sampling locations for this system.

6. Conclusions

This paper presented a general framework to assist experimenters interested in VLE studies, relatively to the task of deciding the optimal set of experiments to maximize the efficiency of their work, and accurately fit a particular thermodynamic model. The framework uses the tools of M-bODE to construct an initial design. This can be complemented by new experiments during the application of the iterative refining phase, limited by the available resources or terminated when the target accuracy for the parameter estimates is reached. Here we have focused on the NRTL and UNIQUAC models for the liquid phase, and the D-optimality criterion, but the framework allows its generalization to other models for liquid and vapor phases, regression criteria, and the measurement of different response variables. The approach is first demonstrated with the methanol-water system and subsequently tested with more complex systems. Both the absence of observational error and the presence of homoscedastic Gaussian observational error are considered.

The results clearly demonstrate the advantages associated with the M-bSODE, relatively to regularly spaced experiments throughout the domain. In all tests, after an equal number of experiments, the confidence regions for the parameters are tighter, corresponding to a reduced parametric variance of the estimates (between $6\,\%$ and about $70\,\%$). Consequently, the M-bSODE is more efficient on a *per observation* basis, i.e., more information is extracted on average from each experiment. This was expected since the framework iteratively locates the experimental conditions that can provide more information.

Fitting of thermodynamic models to VLE data can also be based on regression criteria other than OLS. Such methods require adaptation of the FIMs and are a

topic for future exploration. In the present paper we have also assumed that the errors are normally distributed and homoscedastic. Different assumptions will also impact the derivation of the FIMs, with a subsequent effect on the optimal design. If the measurement errors are not normally distributed a different criterion or an approach based on set-membership estimation, often known as *Guaranteed Parameter Estimation* (GPE), should be considered [6]. The optimal design of experiments for GPE of nonlinear models has been addressed by Mukkula and Paulen [54] among others, and can be adapted to our algorithm. In this case, the confidence regions might significantly differ from the ellipsoids considered during this work, resulting from the asymptotic theory assumptions made.

Acknowledgments

José Granjo acknowledges the financial support of CERENA under the contract programme UIBD/04028/2020.

References

[1] Abrams, D.S., Prausnitz, J.M., 1975. Statistical thermodynamics of liquid mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems. AIChE Journal 21, 116–128. doi:10.1002/aic.690210115.

[2] Andersen, E., Jensen, B., Jensen, J., Sandvik, R., Worsøe, U., 2009. MOSEK version 6. Technical Report. Technical Report TR–2009–3, MOSEK.

[3] Aspentech, Inc., 2016. Aspen Physical Property System, V9. Bedford, MA, USA. Software.

[4] Asprey, S., Macchietto, S., 2000. Statistical tools for optimal dynamic model building. Computers & Chemical Engineering 24, 1261 – 1267. doi:https://doi.org/10.1016/S0098-1354(00)00328-8.

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

[6] Bai, E.W., Tempo, R., Cho, H., 1995. Membership set estimators: size, optimal inputs, complexity and relations with least squares. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications 42, 266–277. doi:10.1109/81.386160.

[7] Baird, Z.S., Dahlberg, A., Uusi-Kyyny, P., Osmanbegovic, N., Witos, J., Helminen, J., Cederkrantz, D., Hyväri, P., Alopaeus, V., Kilpeläinen, I., Wiedmer, S.K., Sixta, H., 2019. Physical properties of 7-methyl-1,5,7-triazabicyclo[4.4.0]dec-5-ene (mTBD). Int J Thermophys 40, 1–23. doi:https://doi.org/10.1007/s10765-019-2540-2.

[8] Baird, Z.S., Uusi-Kyyny, P., Witos, J., Rantamäki, A.H., Sixta, H., Wiedmer, S.K., Alopaeus, V., 2020. Vapor–liquid equilibrium of ionic liquid 7-methyl-1,5,7-triazabicyclo[4.4.0]dec-5-enium acetate and its mixtures with water. Journal of Chemical & Engineering Data 65, 2405–2421. doi:10.1021/acs.jced.9b01039.

[9] Barz, T., López Cárdenas, D.C., Arellano-Garcia, H., Wozny, G., 2013. Experimental evaluation of an approach to online redesign of experiments for parameter determination. AIChE Journal 59, 1981–1995. doi:10.1002/aic.13957.

[10] Ben-Tal, A., Nemirovski, A.S., 2001. Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications. Society for Industrial and Applied Mathematics, Philadelphia.

[11] Beneke, D., Peters, M., Glasser, D., Hildebrandt, D., 2012. Understanding Distillation Using Column Profile Maps. John Wiley & Sons, Ltd, New York, USA. doi:10.1002/9781118477304.ch6.

[12] Bischof, C.H., Bücker, H.M., Lang, B., Rasch, A., Vehreschild, A., 2002. Combining source transformation and operator overloading techniques to compute derivatives for Matlab programs, in: Proceedings of the Second IEEE International Workshop on Source Code Analysis and Manipulation (SCAM 2002), IEEE Computer Society, Los Alamitos, CA, USA. pp. 65–72.

[13] Bonilla-Petriciolet, A., Iglesias-Silva, G.A., Hall, K.R., 2009. Calculation of homogeneous azeotropes in reactive and non-reactive mixtures using a stochastic optimization approach. Fluid Phase Equilibria 281, 22 – 31. doi:https://doi.org/10.1016/j.fluid.2009.03.009.

[14] Box, G.E.P., Hunter, W.G., 1965. Sequential design of experiments for nonlinear models, in: Kort, J.J. (Ed.), Proceedings of IBM Scientific Computing Symposium: Statistics, IBM, White Plains. pp. 113–137.

[15] Boyd, S., Vandenberghe, L., 2004. Convex Optimization. University Press, Cambridge.

[16] Brendel, M., Mhamdi, A., Bonvin, D., Marquardt, W., 2004. An incremental approach for the identification of reaction kinetics. IFAC Pro-

ceedings Volumes 37, 173 – 178. doi:https://doi.org/10.1016/S1474-6670(17)38727-X. 7th International Symposium on Advanced Control of Chemical Processes (ADCHEM 2003), Hong-Kong, 11-14 January 2004.

[17] Brouwer, T., Schuur, B., 2019. Model performances evaluated for infinite dilution activity coefficients prediction at 298.15 K. Industrial & Engineering Chemistry Research 58, 8903–8914. doi:10.1021/acs.iecr.9b00727.

[18] Buzzi-Ferraris, G., Forzatti, P., Canu, P., 1990. An improved version of a sequential design criterion for discriminating among rival multiresponse models. Chemical Engineering Science 45, 477 – 481. doi:https://doi.org/10.1016/0009-2509(90)87034-P.

[19] Buzzi-Ferraris, G., Forzatti, P., Emig, G., Hofmann, H., 1984. Sequential experimental design for model discrimination in the case of multiple responses. Chemical Engineering Science 39, 81 – 85. doi:https://doi.org/10.1016/0009-2509(84)80132-3.

[20] Carlson, E.C., 1996. Don't gamble with physical properties. Chemical Engineering Progress 92, 35–46.

[21] Carvalho, P.J., Khan, I., Morais, A., Granjo, J.F., Oliveira, N.M., Santos, L.M., Coutinho, J.A., 2013. A new microebulliometer for the measurement of the vapor–liquid equilibrium of ionic liquid systems. Fluid Phase Equilibria 354, 156 – 165. doi:https://doi.org/10.1016/j.fluid.2013.06.015.

[22] Chernoff, H., 1953. Locally optimal designs for estimating parameters. The Annals of Mathematical Statistics 24, 586–602.

[23] Chernoff, H., 1959. Sequential design of experiments. Ann. Math. Statist. 30, 755–770. doi:10.1214/aoms/1177706205.

[24] Cook, R., Nachtsheim, C., 1982. Model robust, linear-optimal designs. Technometrics 24, 49–54.

[25] Dechambre, D., Wolff, L., Pauls, C., Bardow, A., 2014. Optimal experimental design for the characterization of liquid–liquid equilibria. Industrial & Engineering Chemistry Research 53, 19620–19627. doi:10.1021/ie5035573.

[26] Dragalin, V., Fedorov, V., 2006. Adaptive designs for dose-finding based on efficacy-toxicity response. Journal of Statistical Planning and Inference 136, 1800 – 1823. doi:https://doi.org/10.1016/j.jspi.2005.08.005. adaptive Designs in Clinical Trials.

[27] Dragalin, V., Fedorov, V., Wu, Y., 2008. Adaptive designs for selecting drug combinations based on efficacy-toxicity response. Journal of Statistical Planning and Inference 138, 352 – 373. doi:https://doi.org/10.1016/j.jspi.2007.06.017. special Issue: Statistical Design and Analysis in the Health Sciences.

[28] Draper, N.R., Hunter, W.G., 1966. Design of experiments for parameter estimation in multiresponse situations. Biometrika 53, 525–533. doi:https://doi.org/10.2307/2333656.

[29] Drud, A., 1985. CONOPT: A GRG code for large sparse dynamic nonlinear optimization problems. Mathematical Programming 31, 153–191.

[30] Duarte, B.P.M., Atkinson, A.C., Granjo, J.F.O., Oliveira, N.M.C., 2019. Optimal design of experiments for liquid–liquid equilibria characterization via semidefinite programming. Processes 7, 834.

[31] Duarte, B.P.M., Wong, W.K., 2015. Finding Bayesian optimal designs for nonlinear models: A semidefinite programming-based approach. International Statistical Review 83, 239–262.

[32] Englezos, P., Kalogerakis, N., 2001. Applied Parameter Estimation for Chemical Engineers. CRC Press, Boca Raton.

[33] Espie, D., Macchietto, S., 1989. The optimal design of dynamic experiments. AIChE Journal 35, 223–229. doi:10.1002/aic.690350206.

[34] Fabries, J.F., Renon, H., 1975. Method of evaluation and reduction of vapor-liquid equilibrium data of binary mixtures. AIChE Journal 21, 735–743. doi:10.1002/aic.690210414.

[35] Fedorov, V.V., 1972. Theory of Optimal Experiments. Academic Press.

[36] Fedorov, V.V., 1980. Convex design theory. Math. Operationsforsch. Statist. Ser. Statist. 11, 403–413.

[37] Fedorov, V.V., Leonov, S.L., 2014. Optimal Design for Nonlinear Response Models. Chapman and Hall/CRC Press, Boca Raton.

[38] Franceschini, G., Macchietto, S., 2008. Novel anticorrelation criteria for model-based experiment design: Theory and formulations. AIChE Journal 54, 1009–1024. doi:10.1002/aic.11429.

[39] Frenkel, M., Chirico, R.D., Diky, V., Yan, X., Dong, Q., Muzny, C., 2005. ThermoData Engine (TDE): Software implementation of the dynamic data evaluation concept. Journal of Chemical Information and Modeling 45, 816–838. doi:10.1021/ci050067b.

[40] Galvanin, F., Barolo, M., Pannocchia, G., Bezzo, F., 2012. Online model-based redesign of experiments with erratic models: A disturbance estimation approach. Computers & Chemical Engineering 42, 138 – 151. doi:https://doi.org/10.1016/j.compchemeng.2011.11.014. european Symposium of Computer Aided Process Engineering - 21.

[41] Galvanin, F., Boschiero, A., Barolo, M., Bezzo, F., 2011. Model-based design of experiments in the presence of continuous measurement systems. Industrial & Engineering Chemistry Research 50, 2167–2175. doi:10.1021/ie1019062.

[42] Gau, C.Y., Brennecke, J.F., Stadtherr, M.A., 2000. Reliable nonlinear parameter estimation in VLE modeling. Fluid Phase Equilibria 168, 1–18. doi:10.1016/S0378-3812(99)00332-5.

[43] Goujot, D., Meyer, X., Courtois, F., 2012. Identification of a rice drying model with an improved sequential optimal design of experiments. Journal

of Process Control 22, 95 – 107. doi:https://doi.org/10.1016/j.jprocont.2011.10.003.

[44] Grant, M., Boyd, S., Ye, Y., 2012. cvx Users Guide for cvx version 1.22. CVX Research, Inc.. 1104 Claire Ave., Austin, TX 78703-2502.

[45] Helton, J.W., Vinnikov, V., 2007. Linear matrix inequality representation of sets. Comm. Pure Appl. Math. 60, 654–674.

[46] Hoang, M.D., Barz, T., Merchan, V.A., Biegler, L.T., Arellano-Garcia, H., 2013. Simultaneous solution approach to model-based experimental design. AIChE Journal 59, 4169–4183. doi:10.1002/aic.14145.

[47] Howat, C.S., Swift, G.W., 1980. A new correlation of propene-propane vapor-liquid equilibrium data and application of the correlation to determine optimum fractionator operating pressure in the manufacture of polymerization-grade propene. Industrial & Engineering Chemistry Process Design and Development 19, 318–323. doi:10.1021/i260074a021.

[48] Kato, M., Kodama, D., Sugiyama, K., 2006. Phase equilibrium measurements of fluid mixtures at high pressures. The Review of High Pressure Science and Technology 16, 251–259. doi:10.4131/jshpreview.16.251.

[49] Kiefer, J., Wolfowitz, J., 1960. The equivalence of two extremum problem. Canadian Journal of Mathematics 12, 363–366.

[50] Kontogeorgis, G., Folas, G., 2010. Thermodynamic models for industrial applications: from classical and advanced mixing rules to association theories. Wiley.

[51] Körkel, S., Kostina, E., Bock, H.G., Schlöder, J.P., 2004. Numerical methods for optimal control problems in design of robust optimal experiments for nonlinear dynamic processes. Optimization Methods and Software 19, 327–338. doi:10.1080/10556780410001683078.

[52] Leonov, S., Miller, S., 2009. An adaptive optimal design for the $e_{\max}$ model and its application in clinical trials. Journal of Biopharmaceutical Statistics 19, 360–385. doi:10.1080/10543400802677240.

[53] Michelsen, M., 1982. The isothermal flash problem .1. stability. Fluid Phase Equilibria 9, 1–19. doi:10.1016/0378-3812(82)85001-2.

[54] Mukkula, A.R.G., Paulen, R., 2017. Robust model-based design of experiments for guaranteed parameter estimation, in: Espuña, A., Graells, M., Puigjaner, L. (Eds.), 27th European Symposium on Computer Aided Process Engineering. Elsevier. volume 40 of *Computer Aided Chemical Engineering*, pp. 1639 – 1644. doi:https://doi.org/10.1016/B978-0-444-63965-3.50275-0.

[55] Nöh, K., Niedenführ, S., Beyß, M., Wiechert, W., 2018. A pareto approach to resolve the conflict between information gain and experimental costs: Multiple-criteria design of carbon labeling experiments. PLOS Computational Biology 14, 1–30. doi:10.1371/journal.pcbi.1006533.

[56] Poling, B.E., Prausnitz, J.M., O'Connel, J.P., 2001. The Properties of Gases and Liquids. 5th. ed., McGraw-Hill, New York.

[57] Pronzato, L., 2008. Optimal experimental design and some related control problems. Automatica 44, 303–325.

[58] Pukelsheim, F., Rieder, S., 1992. Efficient rounding of approximate designs. Biometrika 79, 763–770.

[59] Raal, J.D., Mühlbauer, A.L., 1997. Phase Equilibria: Measurement and Computation. Series in Chemical and Mechanical Engineering, Taylor & Francis, London, UK.

[60] Renon, H., Asselineau, L., Cohen, G., Raimbault, C., 1971. Calcul sur Ordinateurs des Équilibres Liquide-Vapeur et Liquide-Liquide. Institut Français du Pétrole.

[61] Renon, H., Prausnitz, J.M., 1968. Local compositions in thermodynamic excess functions for liquid mixtures. AIChE Journal 14, 135–144.

[62] Richon, D., de Loos, T., 2005. Vapour–liquid equilibrium at high pressure, in: Weir, R., Loos, T.D. (Eds.), Measurement of the Thermodynamic Properties of Multiple Phases. Elsevier. volume 7 of *Experimental Thermodynamics*, pp. 89 – 136. doi:https://doi.org/10.1016/S1874-5644(05)80008-6.

[63] Sagnol, G., 2012. PICOS, a Python interface to conic optimization solvers. Technical Report 12-48. ZIB.

[64] Sagnol, G., 2013. On the semidefinite representation of real functions applied to symmetric matrices. Linear Algebra and its Applications 439, 2829 – 2843.

[65] Seber, G.A.F., Wild, C.J., 2003. Nonlinear Regression. John Wiley & Sons, New York.

[66] Soepyan, F.B., Anderson-Cook, C.M., Morgan, J.C., Tong, C.H., Bhattacharyya, D., Omell, B.P., Matuszewski, M.S., Bhat, K.S., Zamarripa, M.A., Eslick, J.C., Kress, J.D., Gattiker, J.R., Russell, C.S., Ng, B., Ou, J.C., Miller, D.C., 2018. Sequential design of experiments to maximize learning from carbon capture pilot plant testing, in: Eden, M.R., Ierapetritou, M.G., Towler, G.P. (Eds.), 13th International Symposium on Process Systems Engineering (PSE 2018), Elsevier. pp. 283 – 288. doi:https://doi.org/10.1016/B978-0-444-64241-7.50042-2.

[67] Sturm, J., 1999. Using SeDuMi 1.02, a Matlab toolbox for optimization oversymmetric cones. Optimization Methods and Software 11, 625–653.

[68] Telen, D., Nimmegeers, P., Impe, J.V., 2018. Uncertainty in optimal experiment design: comparing an online versus offline approaches. IFAC-PapersOnLine 51, 771 – 776. doi:https://doi.org/10.1016/j.ifacol.2018.04.007. 9th Vienna International Conference on Mathematical Modelling.

[69] Thompson, D.E., McAuley, K.B., McLellan, P.J., 2010. Design of optimal sequential experiments to improve model predictions from a polyethylene molecular weight distribution model. Macromolecular Reaction Engineering 4, 73–85. doi:10.1002/mren.200900033.

[70] Ugray, Z., Lasdon, L., Plummer, J., Glover, F., Kelly, J., Martí, R., 2005. A multistart scatter search heuristic for smooth nlp and minlp problems, in: Metaheuristic Optimization via Memory and Evolution. Springer, pp. 25–51.

[71] Vandenberghe, L., Boyd, S., 1999. Applications of semidefinite programming. Applied Numerical Mathematics 29, 283–299.

[72] Vassiliadis, V.S., Sargent, R.W.H., Pantelides, C.C., 1994. Solution of a class of multistage dynamic optimization problems. 1. problems with path constraints. Industrial & Engineering Chemistry Research 33, 2111–2122.

[73] Wang, S.J., James Hung, H.M., O'Neill, R., 2012. Paradigms for adaptive statistical information designs: practical experiences and strategies. Statistics in Medicine 31, 3011–3023. doi:10.1002/sim.5410.

[74] Whittle, P., 1973. Some general points in the theory of optimal experimental design. Journal of the Royal Statistical Society, Ser. B 35, 123–130.

[75] Wisniak, J., Ortega, J., Fernández, L., 2017. A fresh look at the thermodynamic consistency of vapour-liquid equilibria data. The Journal of Chemical Thermodynamics 105, 385–395. doi:10.1016/j.jct.2016.10.038.

[76] Wong, W.K., Yin, Y., Zhou, J., 2019. Optimal designs for multi-response nonlinear regression models with several factors via semidefinite programming. Journal of Computational and Graphical Statistics 28, 61–73. doi:10.1080/10618600.2018.1476250.

[77] Wynn, H.P., 1972. Results in the theory and construction of $D-$optimum experimental designs. Journal of Royal Statistics Soc. - Ser. B 34, 133–147.

[78] Ye, Y., 1997. Interior Point Algorithms: Theory and Analysis. John Wiley & Sons, New York.

[79] Zarrop, M.B., 1979. Optimal Experimental Design for Dynamic System Identification: Lecture Notes in Control and Information Sciences 21. Springer-Verlag, New York.

[80] Zullo, L.C., 1991. Computer Aided Design of Experiments. An Engineering Approach. Ph.D. thesis. Imperial College of Science, Technology and Medicine. London.

Appendix A

A.1. Semidefinite Programming

Let $\mathbb{S}_+^{n_\theta}$ be the space of $n_\theta \times n_\theta$ symmetric positive semidefinite matrices, and $\mathbb{S}^{n_\theta}$ the space of $n_\theta \times n_\theta$ symmetric matrices. A convex set $\mathbf{S} \in \mathbb{R}^{m_1}$ is semidefinite representable (SDr) if and only if for each vector $\boldsymbol{\zeta} \in \mathbf{S}$ there exists a projection $\mathbf{P}$ on to a higher dimensional set that can be described by Linear Matrix Inequalities (LMIs). That is, $\mathbf{S}$ is SDr if and only if there exists some symmetric matrices $M_0, \cdots, M_{m_1}, M_{m_1+1}, \cdots, M_{m_1+m_2} \in \mathbb{S}^{n_\theta}$ such that [45]:

$$\boldsymbol{\zeta} \in \mathbf{S} \iff \exists \boldsymbol{v} \in \mathbb{R}^{m_2} : \; M_0 + \sum_{i=1}^{m_1} \zeta_i \, M_i + \sum_{j=1}^{m_2} v_j \, M_{m_1+j} \succeq 0. \tag{A.1}$$

Here $\succeq$ is the semidefinite operator, i.e., $A \succeq 0 \iff \langle A, \Omega \rangle \geq 0, \; \forall \Omega \in \mathbb{S}_+^{n_\theta}$, where $\langle ., . \rangle$ is the Frobenius inner product operator, $\boldsymbol{\zeta} \in \mathbb{R}^{m_1}$ is a point of the original set $\mathbf{S}$, $\boldsymbol{v}$ a point of the incremental space $\mathbb{R}^{m_2}$, and $\mathbf{P} : \boldsymbol{\zeta} \mapsto M_0 + \sum_{i=1}^{m_1} \zeta_i \, M_i + \sum_{j=1}^{m_2} v_j \, M_{m_1+j}$.

In turn, a convex (or concave) function $\varphi : \mathbb{R}^{m_1} \mapsto \mathbb{R}$ is SDr if and only if the epigraph of φ, $\{(t, \boldsymbol{\zeta}) : \varphi(\boldsymbol{\zeta}) \leq t\}$, or the hypograph, $\{(t, \boldsymbol{\zeta}) : \varphi(\boldsymbol{\zeta}) \geq t\}$, respectively, are SDr and can be formalized as a set of LMIs [10, 15]. The

problem of finding the optimal values, $\boldsymbol{\zeta}$, of SDr functions is then formulated as a *semidefinite program* of the form

$$\max_{\boldsymbol{\zeta}} \left\{ \boldsymbol{c}^{\intercal}\, \boldsymbol{\zeta}, \sum_{i=1}^{m_1} \zeta_i\, M_i - M_0 \succeq 0 \right\}. \tag{A.2}$$

In our design context, $\boldsymbol{c}$ is a vector of known constants that depends on the design problem, and matrices $M_i,\ i \in \{0, \cdots, m_1\}$ contain elemental FIMs and other matrices produced by the reformulation of the functions $\varphi(\boldsymbol{\zeta})$ into LMIs. The decision variables in vector $\boldsymbol{\zeta}$ are the weights $w_i,\ i \in [\![n_q]\!]$, of the optimal design and other auxiliary variables required.

Ben-Tal and Nemirovski [10] provide a list of SDr functions useful for solving continuous optimal design problems with SDP formulations, see Boyd and Vandenberghe [15, Sec. 7.3]. Recently, Sagnol [64] showed that each criterion in the Kiefer class of optimality criteria defined by

$$\Phi_\delta[\mathcal{M}(\xi^{\text{cont}}|\mathbb{X}^{n_q}, \boldsymbol{\theta})] = \left[\frac{1}{n_\theta}\ \text{tr}(\mathcal{M}(\xi^{\text{cont}}|\mathbb{X}^{n_q}, \boldsymbol{\theta})^\delta)\right]^{1/\delta} \tag{A.3}$$

is SDr for all rational values of $\delta \in (-\infty, -1]$ and general SDP formulations exist for them. This result also applies to the case when $\delta \to 0$; problem (4) falls into this class. Practically, the problem of finding locally optimal continuous experimental plans for the most common (convex) criteria can be formulated as a Semidefinite Programming problem falling into the general representation (A.1), see Vandenberghe and Boyd [71] or Duarte and Wong [31] among others.

Appendix B

In this section we present the VLE model used during the development of the framework. The equilibrium conditions are expressed through iso-fugacity

relations where, for the sake of the exposition, binary mixtures are considered at moderate pressures (i.e., below $10\,\mathrm{bar}$); electrolytic species, supercritical conditions, and self-associating components are excluded. In this situation the vapor phase can be approximated as an ideal gas while the fugacity of the liquid mixture is described with an excess Gibbs free energy model normalized to Raoult's law with a unitary Poynting correction factor [20]. The activity coefficient, here denoted by $\gamma(\boldsymbol{y}_{V,\cdot}, \boldsymbol{y}_{L,\cdot}, T, \boldsymbol{\theta})$, is expressed by the NRTL model [61], but without loss of generality, other models like UNIQUAC [1] can also be used. The vector of parameters in the model, $\boldsymbol{\theta}$, includes the dimensionless pair of binary interaction parameters (BIPs), $\boldsymbol{\tau}$, in turn related to the interaction energy parameters, and the non-randomness parameters, $\boldsymbol{\alpha}$. Herein, we consider the general extended form of the NRTL model implemented in the `aspenONE`® simulator [3] where $\boldsymbol{\tau}$ is given by

$$\tau_{i,j} = a_{i,j} + \frac{b_{i,j}}{T}, \tag{B.1}$$

$a_{i,j}$ and $b_{i,j}$ are BIPs between the i^{th} and j^{th} components of the mixture and T is the absolute temperature (in K). Consequently, in this study we have $\boldsymbol{\theta} \equiv \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\alpha}\}$. For binary systems, $\boldsymbol{\theta} \equiv \{a_{1,2},\ a_{2,1},\ b_{1,2},\ b_{2,1},\ \alpha_{1,2},\ \alpha_{2,1}\}$, where $\boldsymbol{\alpha}$ is fixed within the interval $[0.0,\ 0.3]$ and only the determination of the parameters $\boldsymbol{\theta} \equiv \{\boldsymbol{a}, \boldsymbol{b}\}$ is sought. Since the interaction energy of each species with itself is null, $a_{i,i} = 0$ and $b_{i,i} = 0,\ \forall i \in [\![c]\!]$, these quantities do not need to be estimated from experimental data. We note that values of $\boldsymbol{\alpha}$ are symmetrical and prescribed for pairs of components given their properties; typically they are set before model fitting [60]. We designate n_θ as the number of parameters to be estimated from the experimental data. Here, given the assumptions listed above, $n_\theta = 4$, but can be generalized to any other number according to the thermodynamic model to be

fitted.

The VLE model for this case can be compactly written in the functional form $\boldsymbol{g}(\boldsymbol{y}_{L,.}, \boldsymbol{y}_{V,.}|\boldsymbol{d}, \boldsymbol{\theta})$ where $\boldsymbol{g}(\bullet) = 0$ is a set of nonlinear algebraic equations [61]:

$$y_{V,j} - y_{L,j}\ \gamma_j(\boldsymbol{y}_{L,.}, T, \boldsymbol{\theta})\ \frac{P_j^v(T)}{P} = 0, \quad j \in [\![c]\!], \tag{B.2a}$$

$$\sum_{j=1}^{2} y_{p,j} - 1 = 0, \quad p \in [\![h]\!] \tag{B.2b}$$

$$P_j^v(T) - 10^{A_j - B_j/(T+C_j)} = 0, \quad j \in [\![c]\!] \tag{B.2c}$$

$$z_1 - [\omega\ y_{V,1} + (1-\omega)\ y_{L,1}] = 0 \tag{B.2d}$$

$$\sum_{j=1}^{2} z_j - 1 = 0 \tag{B.2e}$$

Here P_i^v stands for the saturation pressure of component i in the mixture ($i \in [\![c]\!]$), estimated employing the Antoine equation (B.2c) and expressed in $\mathrm{mm\ Hg}$. Equations (B.2d–B.2e) allow the choice of the composition of the initial mixture; ω is the fraction of the initial mixture vaporized at equilibrium which we assume fixed. Alternative descriptions for the liquid activity coefficient model $\gamma_j(\boldsymbol{y}_{L,.}, T, \boldsymbol{\theta})$ from the ones considered here can also be used in (B.2a). Similarly, this equation can be generalized to the treatment of non-ideal vapour phases if $y_{V,j}$ is replaced by the corresponding fugacity of component j in the vapour phase, adding to the set of equations (B.2d–B.2e) a suitable model for its calculation.

Appendix C

In this section we present the thermodynamic models for activity coefficients. The NRTL activity model is [61]

$$\gamma_i = \exp\left[\frac{\sum_{j=1}^{n} x_j\ \tau_{j,i}\ G_{j,i}}{\sum_{k=1}^{n} x_k\ G_{k,i}} + \right.$$

$$+\sum_{j=1}^{n}\frac{x_j\,G_{i,j}}{\sum_{k=1}^{n}x_k\,G_{k,j}}\left(\tau_{i,j}-\frac{\sum_{m=1}^{n}x_m\,\tau_{m,j}\,G_{m,j}}{\sum_{k=1}^{n}x_k\,G_{k,j}}\right)\Bigg]$$

$$G_{i,j}=\exp\left(-\alpha_{i,j}\,\tau_{i,j}\right)$$

$$\tau_{i,j}=a_{i,j}+\frac{b_{i,j}}{T}$$

The UNIQUAC model is as follows [1]:

$$\gamma_i=\gamma_i^{\text{comb}}\times\gamma_i^{\text{resid}}$$

$$\log(\gamma_i^{\text{comb}})=\log(\Phi_i)+1-\Phi_i-\frac{\vartheta}{2}\,q_i\,\left[\log(\Phi_i)-\log(\Psi_i)+1-\frac{\Phi_i}{\Psi_i}\right]$$

$$\log(\gamma_i^{\text{resid}})=q_i\,\left[1-\log\left(\sum_{j=1}^{c}\Psi_j\,x_j\,\tau_{j,i}\right)-\sum_{j=1}^{c}\frac{\Psi_j\,x_j\,\tau_{i,j}}{\sum_{k=1}^{c}\Psi_k\,x_k\,\tau_{k,j}}\right]$$

$$\Phi_i=\frac{r_i}{\sum_{j=1}^{c}x_j\,r_j}$$

$$\Psi_i=\frac{q_i}{\sum_{j=1}^{c}x_j\,q_j}$$

$$\tau_{i,j}=\exp\left(a_{i,j}+\frac{b_{i,j}}{T}\right)$$

$$r_i=\sum_{k=1}^{g}\nu_k\,R_k,\quad q_i=\sum_{k=1}^{g}\nu_k\,Q_k,\quad \vartheta=10.$$

The group volume and area parameters, R_k and Q_k, respectively, are obtained from [3] or calculated from the pure component data.

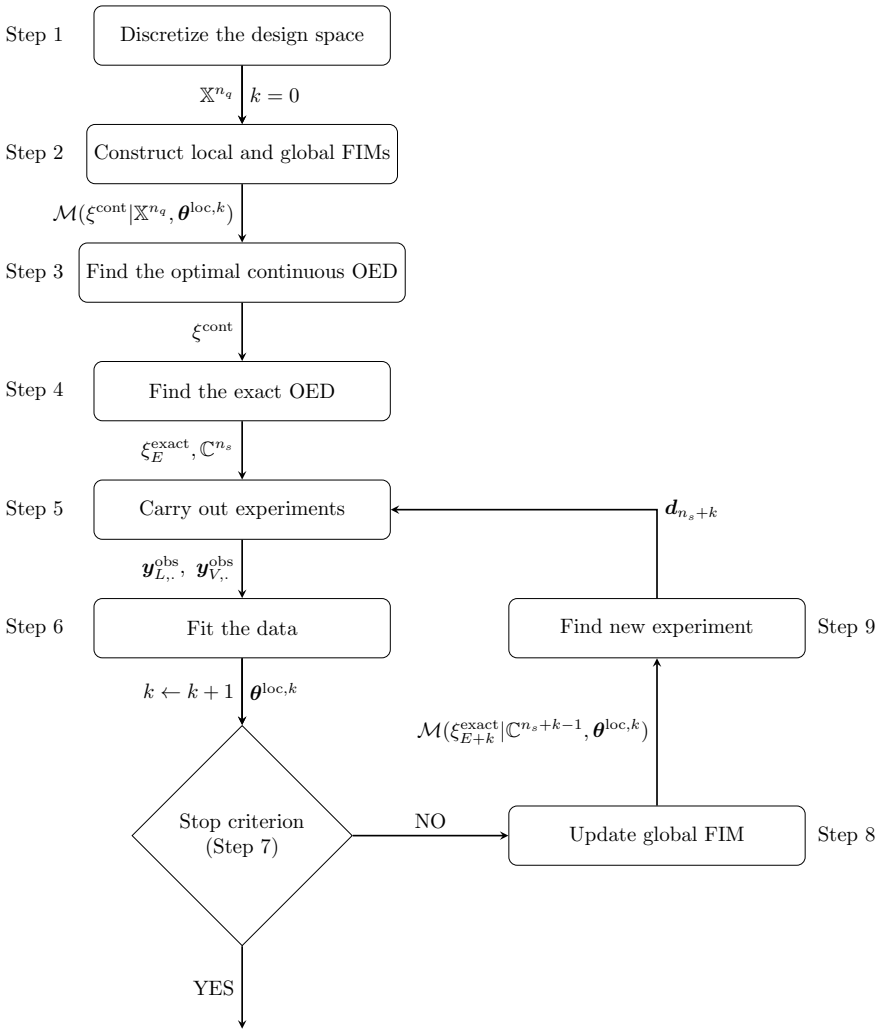
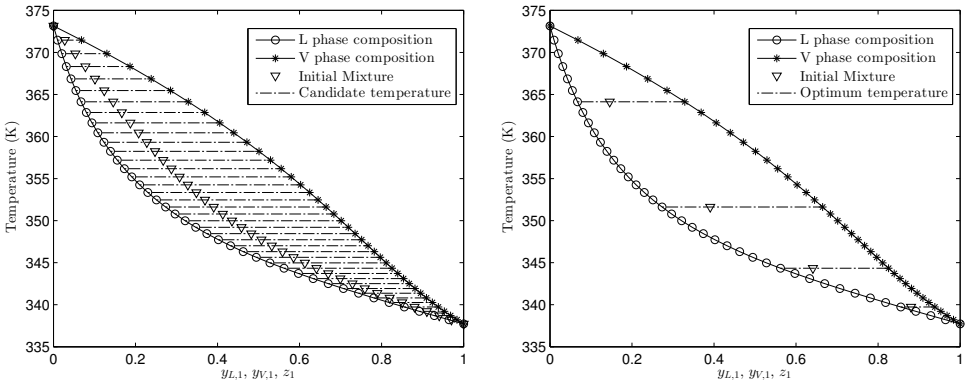


Figure 1: Algorithm for the sequential optimal design of experiments.



(a) (b)

Figure 2: Methanol-water VLE envelope for parameter value $\boldsymbol{\theta}^{\mathrm{loc},0}$: (a) candidate design points for the initial design; (b) continuous initial D-optimal design (based on 4 support points).

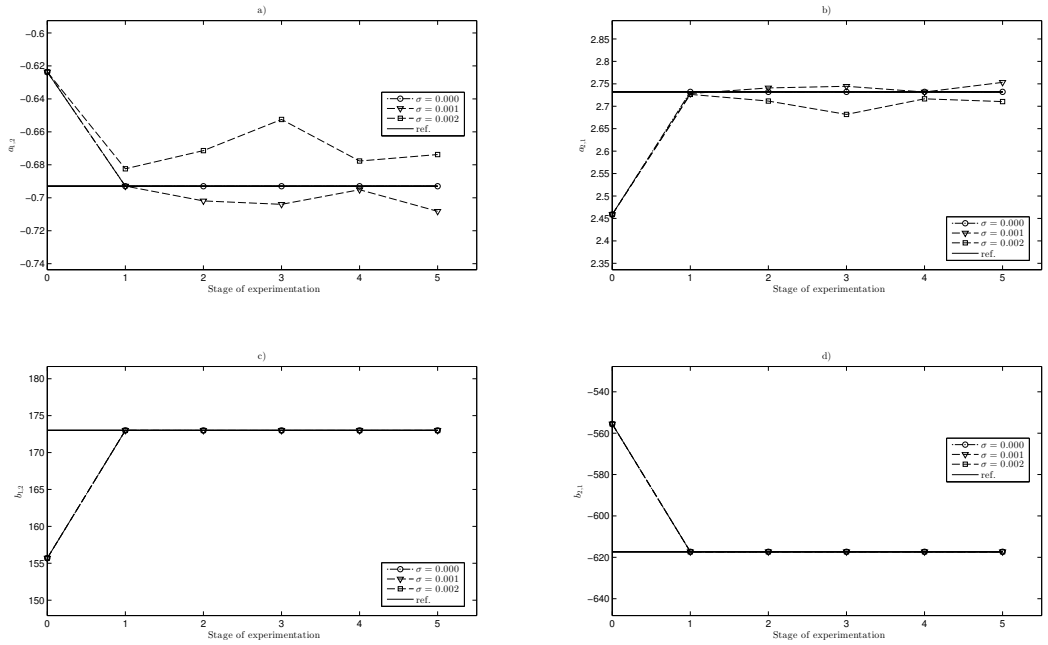


Figure 3: Parameter estimates in successive experimentation stages: a) $a_{1,2}$, b) $a_{2,1}$, c) $b_{1,2}$, d) $b_{2,1}$.

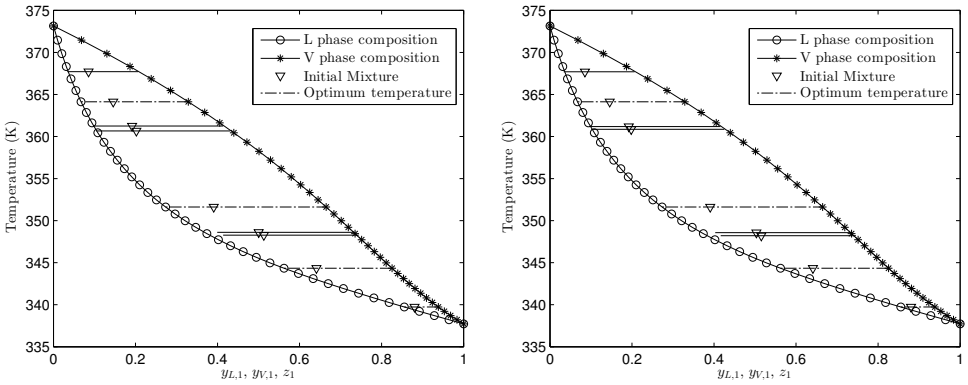
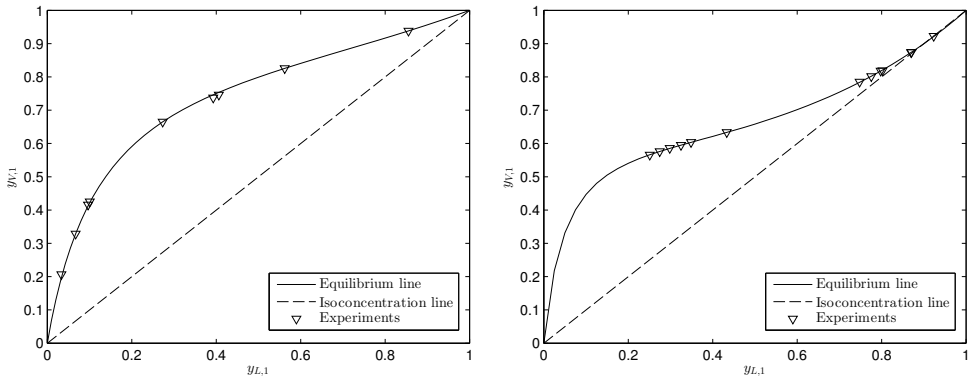
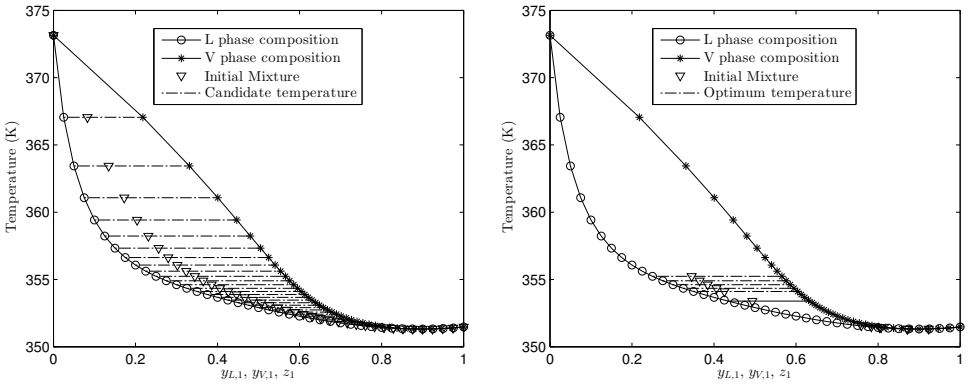


Figure 4: Methanol-water VLE envelope including the accumulated experiments for measurement noise modeled by $\mathcal{N}(0, \sigma_{L,1})$: (a) $\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$; (b) $\sigma_{L,1} = 0.002\,\mathrm{mol/mol}$. The experiments of the initial optimal design correspond to dashed lines and the experiments prescribed in the iterative phase correspond to solid lines.



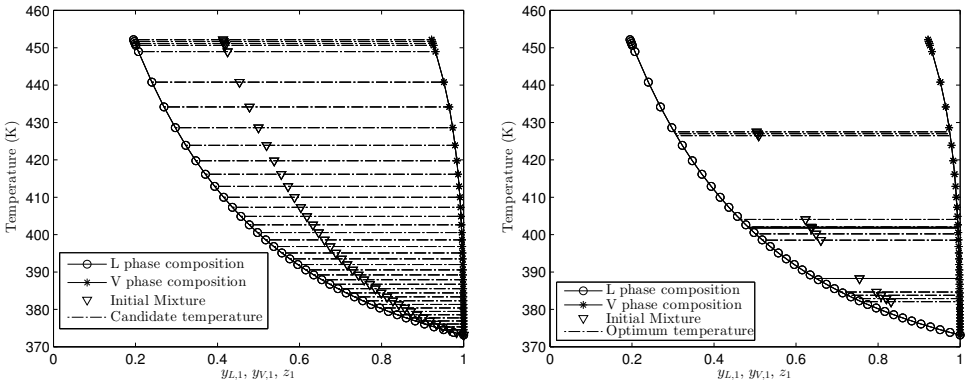
(a) (b)

Figure 5: Phase diagram including the accumulated experiments for measurement noise modeled by $\mathcal{N}(0, \sigma_{L,1})$ ($\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$) for: (a) Methanol-Water system; (b) Ethanol-Water system. The equilibrium attained in experiments correspond to triangles.



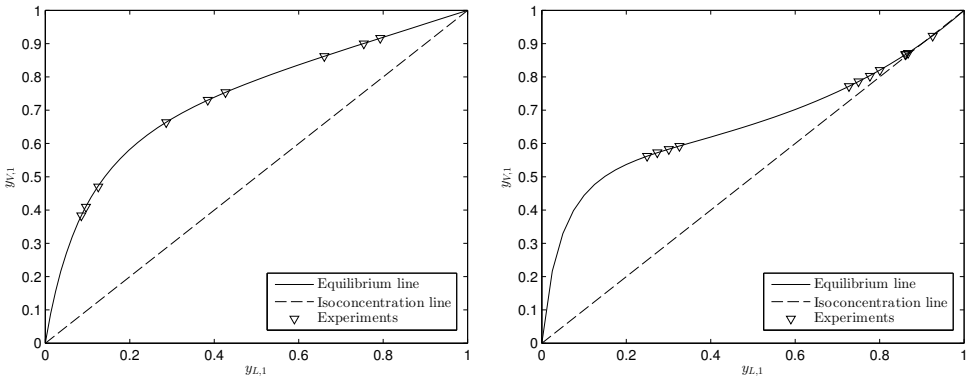
(a) (b)

Figure 6: Ethanol-water VLE envelope for parameter vector $\boldsymbol{\theta}^{\text{loc},0}$ (NRTL model): (a) candidate design points for the initial design; (b) continuous D-optimal design (based on 9 initial support points plus 5 additional experiments). The experiments of the initial optimal design correspond to dashed lines and the experiments prescribed in the iterative phase correspond to solid lines.



(a) (b)

Figure 7: Water-mTBD VLE envelope for parameter vector $\boldsymbol{\theta}^{\text{loc},0}$ (NRTL model): (a) candidate design points for the initial design; (b) continuous D-optimal design (based on 5 initial support points plus 5 additional experiments). The experiments of the initial optimal design correspond to dashed lines and the experiments prescribed in the iterative phase correspond to solid lines.



(a) (b)

Figure 8: Phase diagram including the accumulated experiments ($\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$) for: (a) Methanol-Water system (UNIQUAC model); (b) Ethanol-Water system (UNIQUAC model). The equilibrium attained in experiments correspond to triangles.

Table 1: Antoine constants for system MET-WAT [56].

Component	A	B	C
MET (1)	8.08097	1582.27	−34.450
WAT (2)	8.07131	1730.63	−39.724

Table 2: Binary interaction parameters for vapor-liquid equilibrium characterization via NRTL model for system MET-WAT [3].

	Parameters $a_{i,j}$		Parameters $b_{i,j}$		Parameters $\alpha^{\dagger}_{i,j}$	
	Component		Component		Component	
Component	MET (1)	WAT (2)	MET (1)	WAT (2)	MET (1)	WAT (2)
MET (1)	0.000	−0.693	0.0	173.0	0.0	0.3
WAT (2)	2.732	0.000	−617.3	0.0	0.3	0.0

† Fixed.

Table 3: D-optimal designs for VLE model using NRTL, $\boldsymbol{\theta}^{\text{loc},0} = 0.9 \times \boldsymbol{\theta}^{\text{true}}$ except for $\alpha_{i,j},\ i, j \in [\![c]\!]$, which is set to the value used in the simulation: $T \in [323.15,\ 374.15]$ K.

Designation	Optimal design
ξ^{cont}	$\begin{pmatrix} 0.8362 & 0.4670 & 0.4347 & 0.1760 \\ 0.1638 & 0.5330 & 0.5653 & 0.8240 \\ 760 & 760 & 760 & 760 \\ 341.0000 & 349.5905 & 350.5981 & 362.1546 \\ 0.7925 & 0.3613 & 0.3253 & 0.0856 \\ 0.9383 & 0.7137 & 0.6901 & 0.3868 \\ 0.4390 & 0.1071 & 0.1978 & 0.2571 \end{pmatrix}$
$\xi_{n_s}^{\text{exact}}$	$\begin{pmatrix} 0.8362 & 0.4670 & 0.4347 & 0.1760 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & 1 & 1 \end{pmatrix}$

Table 4: Parameter values fitted in iterative phase of the M-bSODE procedure.

	$\sigma_{L,1} = 0\,\mathrm{mol/mol}$				$\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$			
Iteration	$a_{1,2}$	$a_{2,1}$	$b_{1,2}$	$b_{2,1}$	$a_{1,2}$	$a_{2,1}$	$b_{1,2}$	$b_{2,1}$
0	−0.6237	2.4588	155.7000	−555.5700	−0.6237	2.4588	155.7000	−555.5700
1	−0.6930	2.7320	173.0002	−617.3002	−0.6929	2.7283	173.0002	−617.3002
2	−0.6930	2.7320	172.9942	−617.3016	−0.7019	2.7409	172.9931	−617.3016
3	−0.6930	2.7320	172.9920	−617.3019	−0.7040	2.7444	172.9937	−617.3019
4	−0.6930	2.7320	172.9964	−617.3020	−0.6951	2.7317	172.9965	−617.3021
5	−0.6930	2.7320	172.9929	−617.3029	−0.7082	2.7532	172.9991	−617.3023

	$\sigma_{L,1} = 0.002\,\mathrm{mol/mol}$			
Iteration	$a_{1,2}$	$a_{2,1}$	$b_{1,2}$	$b_{2,1}$
0	−0.6237	2.4588	155.7000	−555.5700
1	−0.6824	2.7265	173.0002	−617.3002
2	−0.6714	2.7115	172.9939	−617.3013
3	−0.6525	2.6818	172.9968	−617.3017
4	−0.6777	2.7165	172.9944	−617.3022
5	−0.6738	2.7104	172.9942	−617.3028

Table 5: Temperature of the prescribed experiments in the iterative stage of the procedure.

	$\sigma_{L,1}$		
Experiment	$0\,\mathrm{mol/mol}$	$0.001\,\mathrm{mol/mol}$	$0.002\,\mathrm{mol/mol}$
1	348.6134	348.5738	348.6427
2	348.2694	367.7045	367.7498
3	367.7102	348.1993	348.2941
4	361.2453	361.1808	361.3240
5	360.6602	360.8580	360.9504

Table 6: Parameter predictions and respective $95\,\%$ confidence intervals after 9 experiments.

	prescribed with M-bSODE		
Parameter	$\sigma_{L,1} = 0\,\mathrm{mol/mol}$	$\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$	$\sigma_{L,1} = 0.002\,\mathrm{mol/mol}$
$a_{1,2}$	-0.6930 ± 0.0000	-0.6846 ± 0.0428	-0.6648 ± 0.0541
$a_{2,1}$	2.7320 ± 0.0000	2.7186 ± 0.0592	2.6765 ± 0.0753
$b_{1,2}$	172.9929 ± 0.0002	172.9977 ± 6.9421	172.9993 ± 8.6112
$b_{2,1}$	-617.3029 ± 0.0002	-617.3024 ± 10.1207	-617.3025 ± 12.2620
	prescribed by uniform choice		
Parameter	$\sigma_{L,1} = 0\,\mathrm{mol/mol}$	$\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$	$\sigma_{L,1} = 0.002\,\mathrm{mol/mol}$
$a_{1,2}$	-0.6930 ± 0.0000	-0.6772 ± 0.0619	-0.6811 ± 0.0870
$a_{2,1}$	2.7320 ± 0.0000	2.7028 ± 0.0856	2.7079 ± 0.1206
$b_{1,2}$	172.9935 ± 0.0001	172.9940 ± 10.1171	172.9938 ± 14.2303
$b_{2,1}$	-617.3048 ± 0.0002	-617.3048 ± 12.2472	-617.3048 ± 17.2524

Table 7: Temperature, initial mixture and equilibrium conditions of the experiments prescribed for ETH-WAT system ($\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$, NRTL model).

	Initial and equilibrium conditions			
Experiment	T (K)	z_1 (mol/mol)	$y_{L,1}$ (mol/mol)	$y_{V,1}$ (mol/mol)
1	351.3273	0.9236	0.9256	0.9227
2	351.4337	0.8137	0.8008	0.8192
3	351.4989	0.7934	0.7741	0.8017
4	351.5769	0.7746	0.7502	0.7851
5	354.1048	0.5281	0.3516	0.6038
6	354.3466	0.5140	0.3251	0.5949
7	354.6101	0.4999	0.2997	0.5858
8	354.9023	0.4859	0.2755	0.5761
9	355.2333	0.4708	0.2497	0.5656
10	353.3995	0.5734	0.4339	0.6332
11	353.3991	0.5733	0.4334	0.6333
12	351.3274	0.8706	0.8678	0.8718
13	355.2257	0.4714	0.2510	0.5658
14	353.4351	0.5712	0.4303	0.6316

Table 8: Parameter predictions and respective $95\,\%$ confidence intervals for system ETH-WAT after 14 experiments ($\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$, NRTL model).

Parameter	prescribed with M-bSODE	prescribed by uniform choice
$a_{1,2}$	-0.8023 ± 0.0158	-0.8049 ± 0.0168
$a_{2,1}$	3.4524 ± 0.0309	3.4599 ± 0.0310
$b_{1,2}$	246.2000 ± 4.6492	246.200 ± 4.7259
$b_{2,1}$	-586.1000 ± 7.9301	-586.100 ± 8.2996

Table 9: Temperature, initial mixture and equilibrium conditions of the experiments prescribed for WAT-mTBD system ($\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$, NRTL model).

	Initial and equilibrium conditions			
Experiment	T (K)	z_1 (mol/mol)	$y_{L,1}$ (mol/mol)	$y_{V,1}$ (mol/mol)
1	424.7316	0.7800	0.3176	0.9782
2	403.0215	0.8384	0.4747	0.9943
3	401.1153	0.8443	0.4924	0.9951
4	384.4713	0.9134	0.7135	0.9990
5	383.6172	0.9181	0.7290	0.9992
6	389.8416	0.8863	0.6252	0.9982
7	408.0926	0.8231	0.4293	0.9919
8	382.2179	0.9266	0.7568	0.9993
9	381.1403	0.9330	0.7779	0.9994
10	380.2899	0.9392	0.7985	0.9995

Table 10: Parameter predictions and respective $95\,\%$ confidence intervals for WAT-mTBD system after 10 experiments ($\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$, NRTL model).

Parameter	prescribed with M-bSODE	prescribed by uniform choice
$a_{1,2}$	0.3886 ± 0.0660	0.2977 ± 0.0897
$a_{2,1}$	-1.3079 ± 0.0899	-1.1052 ± 0.1151
$b_{1,2}$	-857.3802 ± 15.5512	-857.3808 ± 20.3720
$b_{2,1}$	978.4779 ± 17.8381	978.4776 ± 21.5841

Table 11: Temperature, initial mixture and equilibrium conditions of the experiments prescribed for MET-WAT system ($\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$, UNIQUAC model).

	Initial and equilibrium conditions			
Experiment	T (K)	z_1 (mol/mol)	$y_{L,1}$ (mol/mol)	$y_{V,1}$ (mol/mol)
1	340.9419	0.8793	0.7931	0.9163
2	347.8379	0.6556	0.4268	0.7536
3	348.8431	0.6269	0.3849	0.7306
4	361.4484	0.3153	0.0962	0.4092
5	362.3175	0.2938	0.0852	0.3832
6	359.3193	0.3667	0.1254	0.4701
7	351.7348	0.5503	0.2864	0.6634
8	343.1779	0.8015	0.6608	0.8618
9	341.6018	0.8563	0.7544	0.9000
10	341.5970	0.8563	0.7542	0.9001

Table 12: Parameter predictions and respective $95\,\%$ confidence intervals for MET-WAT system after 10 experiments ($\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$, UNIQUAC model).

Parameter	prescribed with M-bSODE	prescribed by uniform choice
$a_{1,2}$	-1.0810 ± 0.0418	-1.0513 ± 0.0478
$a_{2,1}$	0.6636 ± 0.0502	0.6249 ± 0.0584
$b_{1,2}$	432.8785 ± 5.5746	432.8785 ± 6.6552
$b_{2,1}$	-322.1312 ± 8.6892	-322.1312 ± 8.7775

Table 13: Temperature, initial mixture and equilibrium conditions of the experiments prescribed for ETH-WAT system ($\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$, UNIQUAC model).

	Initial and equilibrium conditions			
Experiment	T (K)	z_1 (mol/mol)	$y_{L,1}$ (mol/mol)	$y_{V,1}$ (mol/mol)
1	355.3760	0.4683	0.2496	0.5620
2	355.0404	0.4830	0.2736	0.5727
3	354.7420	0.4981	0.3009	0.5826
4	354.4713	0.5122	0.3259	0.5921
5	351.6078	0.7752	0.7499	0.7860
6	351.5267	0.7950	0.7771	0.8026
7	351.4584	0.8142	0.8006	0.8201
8	351.3365	0.9238	0.9256	0.9230
9	351.6917	0.7583	0.7274	0.7716
10	351.3448	0.8703	0.8674	0.8715
11	351.3498	0.8653	0.8608	0.8673
12	351.3488	0.8664	0.8624	0.8681
13	351.3492	0.8658	0.8611	0.8678

Table 14: Parameter predictions and respective $95\,\%$ confidence intervals for ETH-WAT system after 13 experiments ($\sigma_{L,1} = 0.001\,\mathrm{mol/mol}$, UNIQUAC model).

Parameter	prescribed with M-bSODE	prescribed by uniform choice
$a_{1,2}$	2.0032 ± 0.0419	2.1573 ± 0.0868
$a_{2,1}$	-2.4917 ± 0.0518	-2.3699 ± 0.1221
$b_{1,2}$	-728.9705 ± 12.2946	-728.9705 ± 19.4923
$b_{2,1}$	756.9477 ± 14.7917	756.9477 ± 24.8235