Chemical Engineering Science 299 (2024) 120461

Contents lists available at ScienceDirect

Chemical Engineering Science

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

Data-efficient surrogate modeling of thermodynamic equilibria using Sobolev training, data augmentation and adaptive sampling

Joschka Winz *, Sebastian Engell

TU Dortmund university, Emil-Figge Str. 70, Dortmund, 44227, Germany

ARTICLE INFO

Keywords:
Grax-box modeling
Machine-learning
Parameter-estimation
Process optimization
Phase equilibria
PC-SAFT

ABSTRACT

Modern thermodynamic models, such as the PC-SAFT equation of state, are very accurate but also computationally intensive, which limits their applicability to process design optimization, for example. Surrogate models, which can be evaluated quickly, can be used to approximate the thermodynamic equilibria. However, this requires many data points from the flash calculation routine. In this paper, we investigate three approaches to reduce the number of samples and thus the effort needed to train the surrogate models. First, Sobolev training is used, where the surrogate model is trained not only on the output values, but also on derivative information. Second, data augmentation along the tie lines in LLE systems is proposed to generate samples without additional flash calculations. Third, adaptive sampling is revisited with a novel quality criterion. It is shown that the combination of these techniques can be used to significantly reduce the number of samples required.

1. Introduction

Accurate phase equilibrium predictions are crucial for the design, analysis, and optimization of many processes in the chemical industry. For instance, the assessment of liquid-liquid equilibria (LLE) is fundamental for operations involving separation, mixing, and reaction in multiphase systems. Flash calculations have been employed extensively to determine the stability of a given composition or the phase distribution at equilibrium. When using sophisticated thermodynamic models as e.g. the PC-SAFT (Perturbed-Chain Statistical Associating Fluid Theory) equation of state (Gross and Sadowski, 2001), these compositions can be predicted accurately but to evaluate the model is computationally intensive, posing challenges for real-time applications or optimization tasks that require numerous evaluations. In fact, researchers report that 70 to 75% of computation cost in modeling separation units is spent on the evaluation of phase equilibria (Leesley and Heyen, 1977; Wang and Stenby, 1994).

The recent surge in the development and application of machine learning techniques offers a promising approach to overcome this computational bottleneck. Specifically, surrogate models have been identified as powerful tools in approximating computationally expensive models, providing quick and accurate predictions. Reviews of the applications of surrogate modeling to chemical engineering problems can be found in McBride and Sundmacher (2019); Bhosekar and Ierapetritou (2018).

Surrogate models have also been used to approximate phase equilibrium calculations. Early work in this area used local models to fit simplified thermodynamic models around a point of interest. These can be evaluated much faster than their rigorous counterpart, but can only be used in a limited region of validity. Leesley and Heyen (1977) proposed such an approach using an approximation based only on temperature and pressure, but independent of composition. Chimowitz et al. (1983) extended this approach to a library of local model formulations based on either relative volatilities or equilibrium ratios. An application of this approach to process design can be found in Chimowitz et al. (1984).

Such surrogate models can also be fitted to have a larger range of validity. Over the years much work has considered approximating pT-flash and bubble point calculations of binary vapor-liquid equilibria (VLE) systems over larger input domains (Guimarãaes and McGreavy, 1995; Mohanty, 2005; Vaferi et al., 2013; Farzi and Tarjomannejad, 2015). More recently, the approximation of many different flash routines for such systems has been discussed (Poort et al., 2019).

For ternary systems, Schmitz et al. (2006) used an artificial neural network (ANN) and a type of radial basis function (RBF) networks for classifying between five different possible types of phase states for the multi-component system ethanol, ethyl acetate and water. They found that feed forward neural networks are more accurate than the tested RBF-networks and that networks with more layers, i.e. deeper networks, show a better performance.

* Corresponding author.
E-mail address: joschka.winz@tu-dortmund.de (J. Winz).

https://doi.org/10.1016/j.ces.2024.120461
Received 17 March 2024; Received in revised form 30 June 2024; Accepted 2 July 2024
Available online 8 July 2024

In a similar way Gaganis and Varotsis (2012) used support vector machines to predict phase stability of different mixtures. Kashinath et al. (2018) extended this work by using relevance vector machines, a kind of support vector machine architecture, to not only predict the number of coexisting phases but also the uncertainty in the prediction. This was implemented in an algorithm that uses the predicted phase state only if it is expected to be reliable. They also used an artificial neural network as a phase split surrogate model to predict K-values and correct the prediction if the isofugacity criterion is not fulfilled to the desired accuracy.

For phase split calculations Ihunde and Olorode (2022) investigated surrogate modeling of a pT-flash routine with physics-informed neural networks that penalize violations of the mass balance.

Lopez-Zamora et al. (2021) studied vapor-liquid and vapor-liquid-liquid equilibria. In their work they find that conventional process simulation software is not always reliable in predicting the correct number of phases. Thus, they use a machine learning approach for classification based on real world experimental data. By testing a broad range of ML-model types they found a k-nearest neighbor approach to work the best to model the phase regions.

Nentwich and Engell (2019) approximated the pT-flash calculations using the PC-SAFT equation of state (Gross and Sadowski, 2001), to model a simplified ternary LLE in a hydroformylation process. For the prediction of the phase stability support vector machines are used, while the phase split calculations are approximated by Kriging models. The focus in this work was on presenting a novel adaptive sequential sampling method that is used to determine samples that are most promising for improving global surrogate model accuracy. The full solvent system of the hydroformylation process was approximated using adaptive sampling in our previous work (Winz et al., 2021), where also the optimization of the process based on this model is discussed.

This multi-component system was investigated by McBride et al. (2017) as well for an integrated reaction-extraction process also using Kriging surrogate models for approximating the phase split calculations. A similar process but with a more complex reaction system was studied by Kaiser and Engell (2023). They used surrogate models for many different thermodynamic quantities: syn-gas solubilities in a reactor, LLE behavior in a decanter and the fugacity in a membrane unit. All surrogate models are realized by artificial neural networks.

In other work by Nentwich et al. (2019) a different approach was developed to approximate fugacity coefficients using surrogate models for phase split calculations. This approach leads to surrogate models that still require pT-flash calculations to be performed, but this can be done much more efficiently. Thus, these surrogate models implicitly describe the equilibrium.

This is also the case for the models developed in Kunde et al. (2019). They use a reformulation of the input space. Instead of modeling equilibrium conditions based on the molar fractions of the feed and on temperature, they use the position on the binodal line (or binodal surface in higher dimensions). Given a specific input and temperature this position is not known a priori and has to be solved iteratively, making this method implicit. For data generation a continuation method is used that generates good initial values for the flash calculations.

Another interesting approach to approximate phase equilibria was presented by Iftakher et al. (2022). They construct both an under- and an overestimator of the function that describes the system of equations that has to be solved, for example phase equilibrium conditions. Equilibrium points are then found solving a mixed integer linear program (MILP) to identify points with close upper and lower bounds. This method provides accurate predictions even with few samples but the model again is implicit and the computation time scales exponentially with the number of data points used.

Ma et al. (2022) used surrogate models to model not just the phase equilibrium but the input output behavior of a complete extractive distillation unit. They used ALAMO and ANN surrogate models.

The literature discussed above does not provide a unified and efficient method for generating explicit surrogate models that accurately approximate pT-flash calculations without requiring a large number of samples, i.e. calls of the full thermodynamic model. The objective of this work is to reduce the number of samples that are required to construct accurate and explicit surrogate models for phase stability and phase split calculations. Three approaches are presented to achieve this goal. Firstly, Sobolev training is introduced as a new methodology for pT-flash surrogate models. Sobolev training uses both input-output data and input-output sensitivities to train the surrogate models more efficiently. Secondly, a new methodology for data augmentation along tie lines is described which generates additional samples without evaluating the thermodynamic model. Finally, we improve adaptive sampling and propose a new criterion for sample quality, resulting in an increased number of samples within the miscibility gap.

In this paper, phase stability and phase split surrogate models are exemplary developed for two multi-component systems of different size that exhibit a miscibility gap. The LLEs can be accurately described using PC-SAFT, but this is computationally expensive, hence the need for data-efficient surrogate modeling.

The case studies with these component systems are presented in section 2. The input-output structure of the surrogate models and an extensive hyperparameter study are shown in section 3.

Sections 4, 5 and 6 present the details of our approach and the results of applying Sobolev training, data augmentation, and adaptive sampling. We demonstrate that the phase split surrogate models that were trained using the three approaches are more accurate than models that were trained using a standard approach, even when the latter were trained using 20-30 times more pT-flash computations.

Finally a summary of the findings and further research directions are given in section 7.

2. Case studies

This work considers two case studies: the mixtures that are present in the hydroformylation of 1-dodecene and in the hydroaminomethylation of 1-decene in thermomorphic solvent systems (TMS). TMS enable an innovative way of catalyst recycling for homogeneously catalyzed reactions. These homogeneously catalyzed production processes were discussed in detail in Brunsch and Behr (2013); Hentschel et al. (2015); Hernandez et al. (2018); Kraume et al. (2022). Precise descriptions of the LLE are necessary for the optimization of the design and operating parameters of the two processes, where the LLE models have to be called numerous times, hence the interest in fast and accurate surrogate models. The resulting LLEs differ in the number of components and in the species that are present in the mixture. The number of components significantly affects the difficulty of surrogate modeling. The phase split of the liquid phase has a significant impact on the process design because after the reaction stage catalyst recycling is realized using a decanter and recycle of the catalyst-rich phase to the reactor. Surrogate modeling approaches are several orders of magnitude more efficient numerically than evaluating the phase stability and phase split by pT-flash calculations using precise but computationally expensive models as e.g. PC-SAFT.

These two examples are representatives of the broader class of production processes that include a unit the output of which is strongly dependent on a liquid-liquid equilibrium, such as a decanter. The use of accurate pT-flash surrogate models, which faithfully approximate a computationally expensive predictive thermodynamic model can enable applications such as advanced process control and optimization.

The hydroformylation case study that involves a smaller number of components than the HAM case is used for detailed investigations and hyperparameter screenings. The second, more complex case study is used thereafter to validate the approach which we propose to determine classification and regression surrogate models of pT-flash calculations.

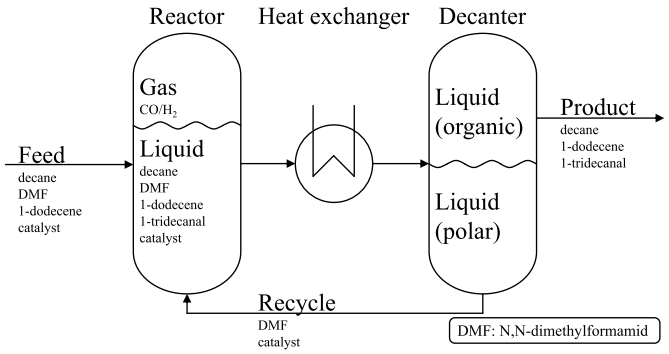


Fig. 1. Flowsheet of the process for the hydroformylation of 1-dodecene in a TMS system.

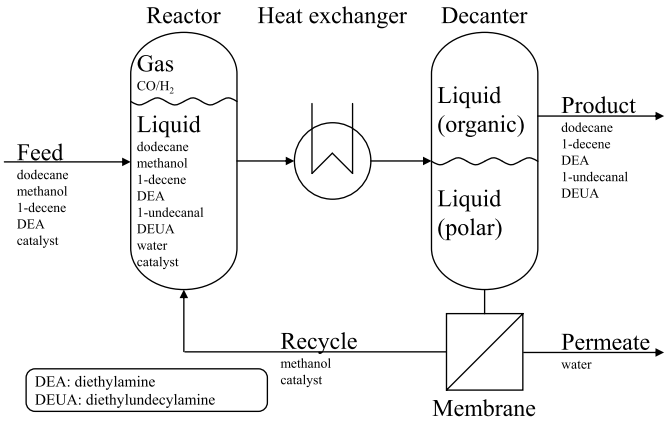


Fig. 2. Flowsheet of the process for the hydroaminomethylation of 1-decene in a TMS system.

2.1. Hydroformylation of 1-dodecene

The first case study is the hydroformylation (HyFo) of 1-dodecene to the desired product 1-tridecanal, which is used e.g. as an intermediate for fragrances (Schäfer et al., 2012). Due to the presence of side reactions, a catalyst with a high selectivity is needed. A homogeneous catalyst based on rhodium leads to a favorable product distribution, but due to the high price of rhodium, special care has to be taken about the recovery of the catalyst. A thermomorphic solvent system can be used to minimize the loss of catalyst, because the catalyst system is polar while the product is organic (Kiedorf et al., 2014). The TMS system consists of an organic and a polar solvent, here decane and dimethylformamid (DMF), and is designed to achieve a temperature dependent liquid-liquid miscibility gap. The reaction step is performed at an elevated temperature in a single-phase regime. The basic flowsheet of this process is shown in Fig. 1.

Subsequently the reaction mixture is cooled down leading to a phase separation in a decanter. The resulting organic stream contains the product 1-tridecanal, the organic solvent decane, and the reactant 1-dodecene. The polar stream containing the polar solvent DMF and the catalyst is recycled back to the reactor leading to a drastically reduced loss of the catalyst.

To accurately simulate and to optimize this process, an accurate model of the liquid-liquid equilibrium that is realized in the decanter is necessary. Prior studies have shown that the equation of state PC-SAFT can be used to this end (Schäfer et al., 2012; Merchan and Wozny, 2016). The mixture in the decanter contains the components decane, 1-dodecene, 1-tridecanal and DMF in significant quantities.

2.2. Hydroaminomethylation of 1-decene

The second case study considered in this work is the hydroaminomethylation (HAM) of 1-decene. This reaction is a combination of a hydroformylation step where 1-decene is converted to 1-undecanal with synthesis gas, and a reductive amination step in which 1-undecanal reacts with diethylamine (DEA) and hydrogen to diethylundecylamine (DEUA). Similar to the first case, a polar rhodium based catalyst system provides the desired selectivity and speed of reaction for this reaction system, and a TMS system was designed to recycle the catalyst efficiently. In this case dodecane and methanol were chosen as the organic and polar solvents (Bianga et al., 2020).

In the reductive amination step, water is formed which has to be separated from the polar recycle stream using a membrane to ensure that the liquid reactor phase is stable (Schlüter et al., 2021). The flowsheet of the hydroaminomethylation (HAM) process is shown in Fig. 2.

Except for the membrane, the process flowsheet of the HAM is similar to the flowsheet for the hydroformylation shown Fig. 1 as the catalyst recycling strategy is identical. An accurate description of the liquid-liquid equilibrium in the decanter also in this case study is crucial for process design and optimization, and is provided by PC-SAFT (Huxoll et al., 2021). The iterative computations by PC-SAFT however make the direct use of this model for process optimization unattractive. The system in the decanter contains the seven components, dodecane, 1-decene, DEA, 1-undecanal, DEUA, water, and methanol.

2.3. Description of liquid-liquid equilibria using PC-SAFT

Two phases I and II are in thermodynamic equilibrium when they have the same temperature T, pressure p and chemical potential μ_j of each of the n_c components:

$$T^I = T^{II} \tag{1}$$

$$p^I = p^{II} \tag{2}$$

$$\mu_j^I = \mu_j^{II} \quad \forall j = 1 \dots n_c. \tag{3}$$

The condition of equal chemical potential can be reformulated in terms of the fugacity coefficients φ, leading to the isofugacity condition:

$$x_j^I \varphi_j^I = x_j^{II} \varphi_j^{II} \quad \forall j = 1 \dots n_c. \tag{4}$$

The PC-SAFT equation of state can be used to model the fugacity coefficient as a function of the molar composition, the temperature and the pressure (Gross and Sadowski, 2001). The equation of state is formulated in terms of the residual Helmholtz energy a^{res}. This quantity consists of the summation of contributions of different interactive effects as shown in equation (5). a^{res} is a nonlinear function of the composition $\boldsymbol{x}$, the temperature T and the pressure p.

$$a^{res} = a^{hc} + a^{disp} + a^{dipole} + a^{assoc} = f(\boldsymbol{x}, T, p) \tag{5}$$

Here, a^{hc} describes the repulsive hard chain contribution, a^{disp} represents dispersive, a^{dipole} polar and a^{assoc} associative interactions. From a^{res} every thermodynamic property can be obtained. For instance the fugacity coefficients result from

$$\ln(\varphi_j) = a^{res} + \left(\frac{\partial a^{res}}{\partial x_j}\right) - \sum_{k=1}^{n_c} x_k \left(\frac{\partial a^{res}}{\partial x_k}\right) + Z - 1 - \ln(Z). \tag{6}$$

Z denotes the compressibility factor which is implicitly defined as shown in equation (7):

$$Z = 1 + \rho \left(\frac{\partial a^{res}}{\partial \rho}\right)$$
$$p = \rho Z k T, \tag{7}$$

where ρ denotes the number density and k denotes the Boltzmann constant. For each evaluation of the fugacity coefficient φ_i the system of equations (7) has to be solved.

For computing the phase split in the decanter in the two case studies, the composition of the organic and polar stream $\boldsymbol{x}^I$ and $\boldsymbol{x}^{II}$ are calculated for a given composition $\boldsymbol{x}^F$ of the feed to the decanter and decanter pressure p and temperature T. Thus, a pT-flash calculation has to be performed to find solutions of the equations (4). Additionally the molar balance around the phase separation has to be fulfilled, i.e.

$$x_j^F - x_j^I = \beta(x_j^{II} - x_j^I) \quad \forall j = 1 \dots n_c \tag{8}$$

has to hold for a single value of $\beta \in (0, 1)$.

A pT-flash algorithm starts with an estimate of the phase compositions of both phases and iteratively improves these estimates until the equilibrium conditions are approximately fulfilled. It should be noted that the conditions stated above are only necessary, not sufficient, conditions for a phase equilibrium. For more details on the conditions of equilibrium and pT-flash algorithms the reader is referred to Michelsen and Mollerup (2007).

In process simulation and optimization, the system of equations (7) has to be solved many times which makes simulation and optimization using PC-SAFT calls computationally expensive. This motivates to apply surrogate models to accelerate the phase equilibrium calculations.

3. Surrogate modeling of thermodynamic equilibria

3.1. Input-output structure of surrogate models for pT-flash approximations

The solutions of the pT-flash calculations are approximated using surrogate models. The system temperature T and the composition of the feed $\boldsymbol{x}^F$ are considered as the set of input variables $\underline{x}$. Due to the fact that all molar fractions add to one, the last component is not explicitly considered in this work. The pressure p is technically also an input variable but due to its small influence on the LLE-behavior it is disregarded here, so

$$\underline{x} = \left[T, x_1^F, \dots x_{n_c-1}^F\right]. \tag{9}$$

If the mixture at the feed composition is unstable at the given temperature and pressure, the outputs of the pT-flash are the two compositions $\boldsymbol{x}^I$ and $\boldsymbol{x}^{II}$ of the two phases. Much of the work on surrogate modeling has reformulated the set of output variables to the phase distribution coefficients κ_j (Nentwich and Engell, 2019; Winz et al., 2021; Kaiser and Engell, 2023)

$$\kappa_j = \frac{\dot{n}_j^I}{\dot{n}_j^F} = \frac{x_j^I}{x_j^F} \frac{x_j^F - x_j^{II}}{x_j^I - x_j^{II}}. \tag{10}$$

$\dot{n}_j^i$ denotes the molar flow rate of component j in phase i. The phase distribution coefficients quantify how much of a component ends up in phase I and in phase II after the phase split. If the feed composition is equal to one of the two phase compositions, the phase distribution coefficient tends to 0 or 1:

$$\begin{aligned} \lim_{\boldsymbol{x}^F \to \boldsymbol{x}^I} \kappa_j &= 1 \\ \lim_{\boldsymbol{x}^F \to \boldsymbol{x}^{II}} \kappa_j &= 0. \end{aligned} \tag{11}$$

From these limit values one can derive that in the vicinity of the critical point of the miscibility gap, in which $\boldsymbol{x}^I$ tends to $\boldsymbol{x}^{II}$, the phase distribution coefficients $\boldsymbol{\kappa}$ rapidly change from 1 to 0 along the tie line. This leads to gradients that tend to infinity at the critical point.

Such a response surface is difficult to approximate using a surrogate model, especially with Sobolev training, in which the deviation of the slope of the surrogate model from the slope of the original function is also included in the cost function. For this reason in this work we consider the molar fractions in the two phases for all but one component as the outputs:

$$\underline{y} = \left[x_1^I, \dots, x_{n_c-1}^I, x_1^{II}, \dots, x_{n_c-1}^{II}\right]. \tag{12}$$

Table 1
Input space for the hydroformylation case study.

	Lower bound	Upper bound
T [K]	278.15	303.15
x_{decane} [-]	0	1
x_{DMF} [-]	0	1
$x_{1-dodecene}$ [-]	0	1
$x_{1-tridecanal}$ [-]	0	1

Table 2
Input space for the hydroaminomethylation case study.

	Lower bound	Upper bound
T [K]	263.15	398.15
$x_{dodecane}$ [-]	0	1
$x_{methanol}$ [-]	0	1
$x_{1-decene}$ [-]	0	1
x_{DEA} [-]	0	1
$x_{1-undecanal}$ [-]	0	1
x_{DEUA} [-]	0	1
x_{water} [-]	0	1

Therefore, a regression surrogate model of the phase split, $\hat{f}_{split}$, represents a mapping $\hat{f}_{split} : \mathbb{R}^{n_c} \to \mathbb{R}^{2(n_c-1)}$. As discussed, this surrogate must only be used for inputs that correspond to conditions for which the feed is unstable. Thus the distribution model is combined with a surrogate model for classification that predicts the stability of the mixture at the feed composition: $\hat{f}_{stability} : \mathbb{R}^{n_c} \to [0, 1]$.

It should be emphasized that the surrogate model that predicts the phase stability $\hat{f}_{stability}$ not only is fast to evaluate but also differentiable. This can be used in process optimization to determine how to change the process conditions to achieve a stable composition in the reactor or an unstable composition in the decanter.

Both surrogate models are machine learning models and therefore have many hyperparameters. The next section describes how these hyperparameters are chosen in this work.

3.2. Hyperparameter study

In this work ANNs are used as surrogate models due to the fact that they can be applied to large data sets and give accurate predictions both for regression and for classification problems. Optimal hyperparameters should be used for each network to make a fair comparison between the different approaches. To this end, first a hyperparameter screening was performed in which a large number of different ANN structures was tested for training sets of different sizes. The training sets were created by randomly distributing points in the input space. The input spaces for the two case studies are shown in Table 1 and 2. For the HAM case study not only is the input space of higher dimension due to the larger number of components, but also the considered temperature range is much larger with 135 K compared to the range in the HyFo case study of 25 K.

Two different variants of the Levenberg-Marquardt (LM) training algorithm were investigated. In the first variant, the LM algorithm is used without regularization and in the second using an adaptive regularization technique that does not require an additional hyperparameter, referred to as Bayesian regularization (Dan Foresee and Hagan, 1997; MacKay, 1992). The LM algorithm is a quasi-second order method that leads to fast convergence, especially in noiseless settings. Also, since the learning rate is adjusted automatically, there are no hyperparameters for the training algorithm except for the termination criterion. In this work the LM algorithms were run for a maximum of 500 epochs and the training was stopped when the error on a validation set, taken as 10% of the total training set, was not decreasing over 6 epochs. The algorithm is taken from MATLABs Deep Learning Toolbox (Beale et al.,

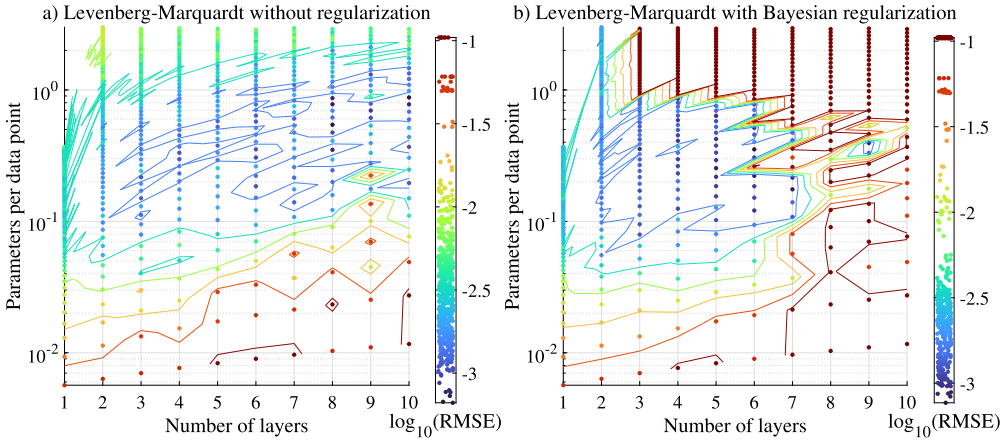


Fig. 3. Hyperparameter screening for the multi-component system from the hydroformylation process using a set of 500 randomly sampled points for the phase split surrogate model.

2018) and was reimplemented to allow for GPU-support and Sobolev training using pytorch (Paszke et al., 2019).

Besides the training algorithm there are further degrees of freedom that define the structure of the network. Here we consider only networks with fully connected layers having the same number of nodes in each hidden layer. The results of the screening of both the number of hidden layers and the number of nodes in each of these are shown in Fig. 3 for a data set of 500 samples of phase split data from the hydroformylation case study (see section 2.1). In this work we refer to a sample or sample point as the set of all output values for one set of input values, while a data point refers to one single output value. The error values are quantified as the root mean squared error (RMSE) for the phase split and the fraction of misclassified points (MisC) for phase stability models. The error metrics are defined in (13) and are always computed for large test sets that were not used in training or validation, consisting of 55000 samples for the HyFo system and 30000 samples for the HAM system.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N n_y}\sum_{i=1}^{N}\sum_{j=1}^{n_y}\left((\underline{y}_j^{\mathrm{split}})_i - (\underline{\hat{y}}_j^{\mathrm{split}})_i\right)^2}$$
$$\mathrm{MisC} = \frac{1}{N}\sum_{i=1}^{N}\left|(\underline{y}^{\mathrm{stability}})_i - (\underline{\hat{y}}^{\mathrm{stability}})_i\right| \tag{13}$$

Here, $(\cdot)_i$ denotes the evaluation at the i^{th} data point, $\underline{y}_j^{\mathrm{split}}$ is the j^{th} of n_y phase split outputs and $\underline{y}^{\mathrm{stability}}$ is the phase stability output. $\underline{y}$ and $\underline{\hat{y}}$ refer to the test set values and the model predictions. N denotes the number of samples in the test set. In this formulation the model predictions of the phase stability $\underline{\hat{y}}^{\mathrm{stability}}$ are rounded to either 0 or 1.

In Fig. 3 the results of the hyperparameter screening are shown. The color bar displays the markers from the diagrams, providing a quantification of errors and insights into their distribution.

The vertical axis in these diagrams is chosen as the number of parameters per data point, such that going horizontally from left to right relates to comparing network structures with the same number of parameters that are distributed in progressively deeper structures.

The maximum number of parameters per data point was set to three and the maximum number of nodes in each layer to 100.

From these results overfitting can be observed without regularization because the error increases for networks with two layers with a high number of parameters. The applied regularization here suppresses overfitting because the error does not increase with regularization.

Overregularization can also be observed, which occurs for a high or a very low number of parameters especially for deeper networks. This leads to networks that show approximately constant predictions for the inputs in the relevant range. The corresponding error values are slightly higher than 10^{-1}.

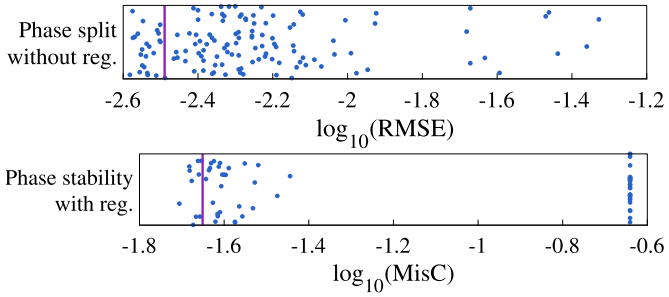


Fig. 4. Test set errors for the phase split and stability surrogate models for the system from hydroaminomethylation. Purple lines: Interpolated error values with the hyperparameters from the hydroformylation case.

The combinations of hyperparameters that yield low errors are different between the two methods, but the resulting error is similar. To investigate how sensitive the results are to the number of data points in the data set, the same screening was performed for data sets of 100, 300, 700 and 900 samples. The best parameter settings are found by minimizing the mean of the $\log_{10}$(RMSE) over the five different data sets for both methods of regularization. The same procedure was conducted for the phase stability surrogate model with data sets of 100, 500, 1000 and 2000 samples. The final results are shown in Table 3. All additional results from the screening are shown in the supplementary material.

It can be seen that the best error values do not change significantly when using regularization. For the phase split surrogate model, the unregularized training gives slightly better results.

To investigate whether the chosen hyperparameters can generally be used in the approximation of pT-flash calculations of systems that contain polar and nonpolar solvents and long-chain organic molecules, a hyperparameter screening was conducted for the larger system from the HAM case study. In Fig. 4, the results are shown and are compared with the results when using the hyperparameters that were found to be optimal for the hydroformylation case.

It can be seen that the hyperparameters that were determined for the mixture from the hydroformylation process also lead to good results for the mixture of the HAM process. Therefore it is reasonable to use the same hyperparameters for both case studies.

4. Sobolev training

Sobolev training denotes the approach to train a neural network not only using the input ($\underline{x}$) and output ($\underline{y}$) data, but also using data of

Table 3
Optimal values for the number of layers and parameters per data point for nominal neural network training, best parameter settings in bold print.

	Number of layers	Number of parameters per data point	mean $\log_{10}$(RMSE)
$\hat{f}_{split}$ without reg.	**5**	**0.422**	**-2.704**
$\hat{f}_{split}$ with reg.	4	0.292	-2.686
$\hat{f}_{stability}$ without reg.	2	0.192	-1.605
$\hat{f}_{stability}$ with reg.	**2**	**0.269**	**-1.691**

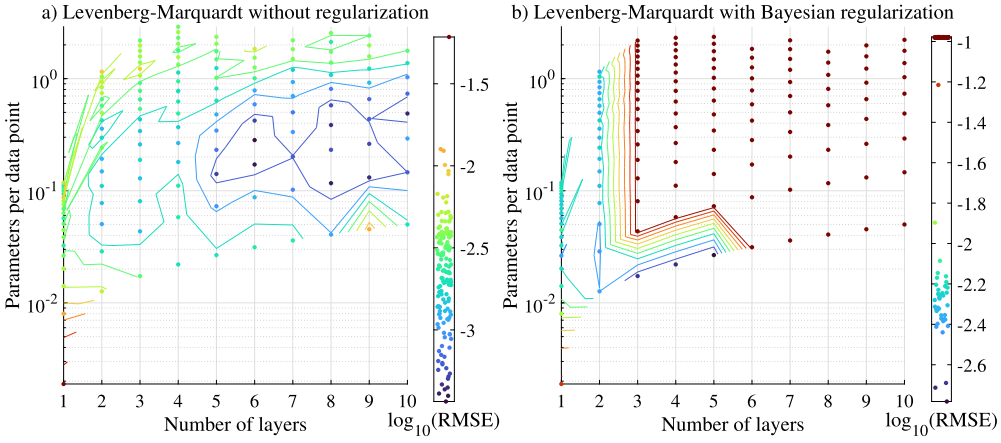


Fig. 5. Hyperparameter screening for the multi-component system from the hydroformylation process using a set of 300 randomly sampled points using Sobolev training with $\alpha = 0.01$.

the derivative of the output with respect to the input (Czarnecki et al., 2017), i.e. the Jacobian, in our case

$$\begin{pmatrix} \frac{dx_1^I}{dT} & \frac{dx_1^I}{dx_1^F} & \cdots & \frac{dx_1^I}{dx_{n_c-1}^F} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{dx_{n_c-1}^I}{dT} & \frac{dx_{n_c-1}^I}{dx_1^F} & \cdots & \frac{dx_{n_c-1}^I}{dx_{n_c-1}^F} \\ \frac{dx_1^{II}}{dT} & \frac{dx_1^{II}}{dx_1^F} & \cdots & \frac{dx_1^{II}}{dx_{n_c-1}^F} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{dx_{n_c-1}^{II}}{dT} & \frac{dx_{n_c-1}^{II}}{dx_1^F} & \cdots & \frac{dx_{n_c-1}^{II}}{dx_{n_c-1}^F} \end{pmatrix} \in \mathbb{R}^{2(n_c-1)\times n_c}. \tag{14}$$

The values of the Jacobian matrix were derived analytically. This is done by considering the set of nonlinear equations that constitute the necessary conditions for an equilibrium and applying the implicit function theorem. Details on this can be found in the supplementary material.

The resulting cost function for the training of the neural network is

$$L = \frac{1}{N n_y (n_x+1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{n_y} \left((\underline{y}_j^{\text{split}})_i - (\underline{\hat{y}}_j^{\text{split}})_i \right)^2 + \alpha^2 \sum_{i=1}^{N} \sum_{j=1}^{n_y} \sum_{k=1}^{n_x} \left(\left(\frac{d\underline{y}_j^{\text{split}}}{d\underline{x}_k} \right)_i - \left(\frac{d\underline{\hat{y}}_j^{\text{split}}}{d\underline{x}_k} \right)_i \right)^2 \right). \tag{15}$$

Here, $\underline{x}_k$ is the k^{th} element of the input $\underline{x}$. α is the scaling factor of the residuals of the Sobolev loss term. Note, that Sobolev training can only be applied to regression problems, in this case to the phase split approximation problem.

The hyperparameters that were obtained by the procedure discussed in section 3.2 are not necessarily optimal for this different neural network training problem. Therefore another hyperparameter screening was conducted with $\alpha = 0.01$ for 300 random samples from the hydroformylation multi-component system. The results are shown in Fig. 5.

Table 4
Optimal hyperparameters for Sobolev neural network training.

	Number of layers	Number of parameters per data point	mean $\log_{10}$(RMSE)
$\hat{f}_{split}$ without reg.	**9**	**0.393**	**-3.174**
$\hat{f}_{split}$ with reg.	2	0.082	-1.914

From these results it is apparent that the Bayesian regularization significantly increases the error. This can be explained by the derivative data itself having a regularizing effect. Therefore if weight decay is applied the result can be a too strong regularization leading to a low accuracy. The same effect was observed in work on approximating the solution of an optimization based controller (Lüken et al., 2023). Without regularization, deep networks perform significantly better with the same number of parameters compared to shallow networks.

In a similar fashion as before, screening runs were conducted for data sets of 100 and 500 samples. The optimal hyperparameters are shown in Table 4.

This data underlines the observations discussed before. The best found network structure for Sobolev training has a large number of layers with a similar number of parameters as the nominal networks. Networks trained with regularization lead to poor results in this case.

The effect of the Sobolev weighting parameter α was then investigated for the determined best network structure. To this end, a range of values of α was applied to data sets of different sizes. In Fig. 6 results are shown for 100, 300, 500 and 1000 sample points, with five repetitions each. On the same data sets, neural networks were trained in the nominal way for comparison. For nominal networks no derivative information is used and the network structure is determined from the hyperparameter screening in section 3.2.

In this figure it can be seen that for all tested data sets Sobolev training yields more accurate surrogate models than networks trained only with input-output data for a certain range of values of α. Also it becomes visible that with an increasing number of samples, the errors of

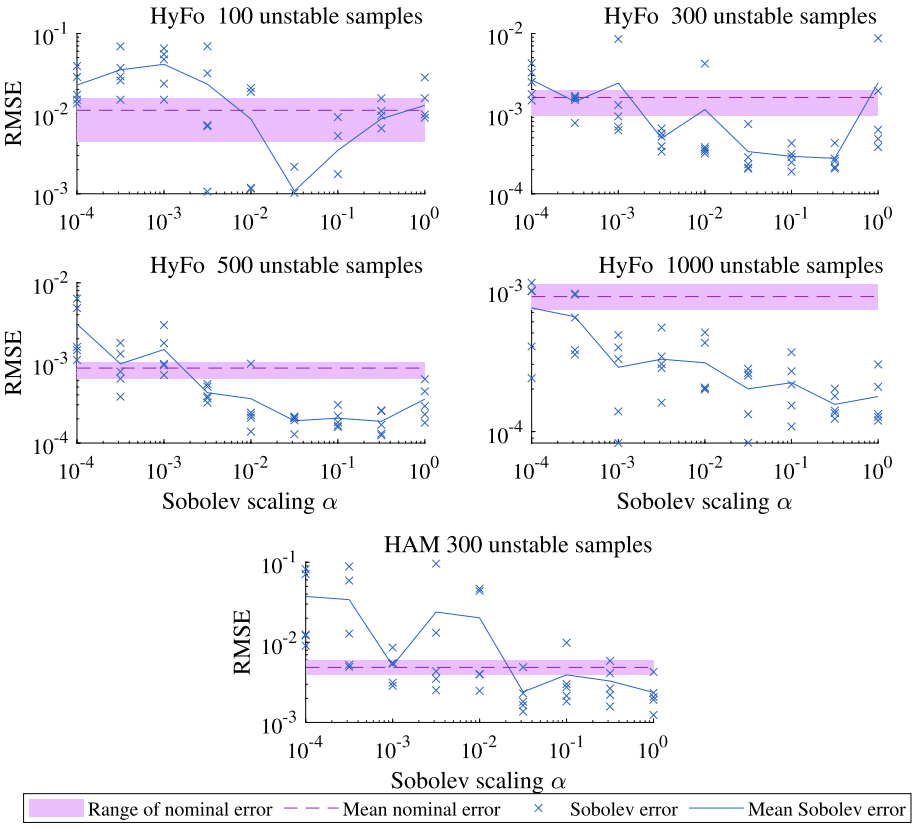


Fig. 6. Results of Sobolev training for data sets of different sizes for the two multi-component systems and for a range of values of the Sobolev scaling factor α.

the networks with Sobolev training decrease faster than the errors in the nominal case. For the data set of 1000 samples, all tested values of α yield more accurate surrogate models compared to training on input-output data only. The optimal value of α is $3 \cdot 10^{-2}$ for 100 samples, approximately 10^{-1} for 300 and 500 samples and $3 \cdot 10^{-1}$ for 1000 samples. From this one can conclude that if the number of samples is low, using the derivative data does not provide significant improvements, because the optimal weight α is low. This might be due to the fact that if at a point there is a large slope and there are no data points close to it, this can lead to larger extrapolation errors if the network is aggressively fitted to this data.

Comparing the results of the two systems, the same range of values of α was observed to be efficient but the advantage of using Sobolev training is smaller for the HAM system. This coincides with the observation that Sobolev training shows a bigger advantage on larger data sets with denser samples. In the system from the HAM process, the density of samples is lower compared to the system from the HyFo, as the input space is much larger.

All in all, Sobolev training is a promising approach for increasing the accuracy of pT-flash surrogate models. A value of $\alpha = 3 \cdot 10^{-2}$ yields improvements in all tested cases and is further used throughout this work.

5. Data augmentation

If for a given temperature T and pressure p a mixture of the composition $\boldsymbol{x}_0^F$ is unstable it will split into two or possibly more stable phases $\boldsymbol{x}^I$, $\boldsymbol{x}^{II}$. From thermodynamics it is known that each mixture of composition $\boldsymbol{x}_i^F$ splits into the same two phases $\boldsymbol{x}^I$, $\boldsymbol{x}^{II}$ if it lies on the tie line that is connecting both phases, or formally if

$$\boldsymbol{x}_i^F = \left(\boldsymbol{x}^{II} - \boldsymbol{x}^I\right)\beta_i + \boldsymbol{x}^I \tag{16}$$

with $\beta_i \in (0,1)$ holds. This fact can be used to generate $n_{\text{aug,TP}}$ samples in the two-phase region for each computed equilibrium point by creating $n_{\text{aug,TP}}$ values for $\beta_i \in (0,1)$ and evaluating equation (16).

For many systems it is known that a liquid-liquid equilibrium exists only for a certain range of temperatures and pressures. In this case, more samples can be created outside of the miscibility gap by using equation (16) with $\beta_i \in [\beta_{\min}, 0) \cup (1, \beta_{\max}]$. Here $\beta_{\min}$ and $\beta_{\max}$ denote the minimum and maximum physically relevant values where all molar fractions are non-negative and below unity. These can be computed as

$$\begin{aligned}\beta_{\max} &= \underset{j}{\operatorname{argmin}}\left(\beta_{\text{edge},j} \mid \beta_{\text{edge},j} \geq 0\right)\\ \beta_{\min} &= \underset{j}{\operatorname{argmax}}\left(\beta_{\text{edge},j} \mid \beta_{\text{edge},j} \leq 0\right)\end{aligned} \tag{17}$$

with

$$x_{i,j}^F(\beta_{\text{edge},j}) = 0 \Rightarrow \beta_{\text{edge},j} = \frac{x_j^I}{x_j^I - x_j^{II}}. \tag{18}$$

Note, that the cutoff at 0 in equation (17) is arbitrary. Any value in $[0,1]$ can be used as $\beta_{\text{edge},j}$ cannot attain values in $(0,1)$. Similar to the approach in the two-phase region, an additional data set in the one-phase region can be generated by sampling $\beta_i \in [\beta_{\min}, 0) \cup (1, \beta_{\max}]$ and applying equation (16).

The sampling of β_i for data augmentation can be done in different ways. Here, we investigate two approaches, random and uniform data augmentation. In random data augmentation, β_i is drawn from a uniform distribution over the corresponding interval. For uniform data augmentation the values of β_i depend on the number of additional samples that are generated, $n_{\text{aug,TP}}$ and $n_{\text{aug,OP}}$. The way the samples are distributed follows the heuristic that samples near the edge of the two-phase region are particularly valuable for the classification surrogate model because they improve the precision of the location of the decision boundary.

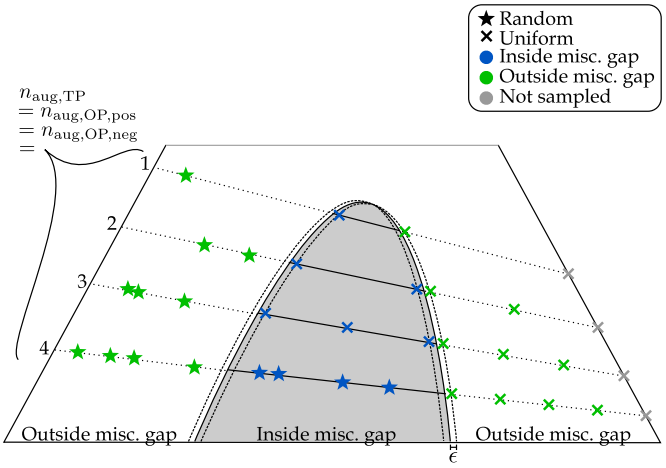


Fig. 7. Uniform and random data augmentation for pT-flash data.

Inside of the miscibility gap new samples are uniformly distributed with β_i between ϵ and $1-\epsilon$. ϵ is used as a threshold to not have samples collapse at the phase boundary. If there is only one sample ($n_{aug,TP} = 1$) it is set to either ϵ or $1-\epsilon$ with equal probability to remove any bias of creating more samples on one side of the miscibility gap. This is formalized in equation (19).

$$\beta_i = \left\{ \begin{array}{ll} \frac{i-1}{n_{aug,TP}}(1-2\epsilon)+\epsilon & \text{if } n_{aug,TP} > 1 \\ X(1-2\epsilon)+\epsilon, & \text{if } n_{aug,TP} = 1 \end{array} \right\}, i = 1 \dots n_{aug,TP} \tag{19}$$

with X being drawn from a Bernoulli distribution: $X \sim \mathcal{B}(0.5)$.

Outside of the miscibility gap, a new sample can lie either in the interval $[\beta_{min}, 0)$ or in $(1, \beta_{max}]$. As discussed, these intervals are approximated as $[\beta_{min}, -\epsilon]$ and $[1+\epsilon, \beta_{max}]$. The number of samples in each interval will be denoted as $n_{aug,OP,neg}$ and $n_{aug,OP,pos}$, with $n_{aug,OP,neg} + n_{aug,OP,pos} = n_{aug,OP}$. Ideally the number of samples in each interval is the same, if $n_{aug,OP}$ is odd the interval that contains one more sample is chosen randomly.

Inside of these intervals the samples are again distributed uniformly though the sample at the outside edge is disregarded because it is not expected to generate much information. The equations to compute β_i outside of the miscibility gap in the one-phase region are shown below.

$$\beta_i = 1 + \epsilon + \frac{i-1}{n_{aug,OP,pos}}(\beta_{max} - (1+\epsilon)) \quad i = 1 \dots n_{aug,OP,pos} \tag{20}$$

$$\beta_i = -\epsilon + \frac{i-1}{n_{aug,OP,neg}}(\beta_{min} + \epsilon) \quad i = 1 \dots n_{aug,OP,neg} \tag{21}$$

These data augmentation strategies are visualized in Fig. 7.

The two described methodologies for data augmentation were applied to the multi-component system from the HyFo process. To this end, data sets of 100, 200, 400 and 800 samples were created. Each of these data sets was augmented by adding 1, 3, 7 or 15 samples per tie line. Including the case without augmentation, 20 different data sets were considered. 80% of the points were augmented inside and 20% were augmented outside of the miscibility gap. ANN models with the hyperparameters from section 3.2 were trained five times on these data sets, the results are shown in Fig. 8.

In Fig. 8 the test set errors are shown for all combinations of the data augmentation factor and the number of samples in the base set. It can be seen that the error decreases in both dimensions for both uniform and random data augmentation. I.e., increasing the size of the data set through sampling or through data augmentation increases the accuracy of the surrogate model. This trend is almost monotonic, the only exception are data sets with a data augmentation factor of 15. This augmentation factor slightly increases the errors compared to a data augmentation factor of 7. The fact that the RMSE increases with very high data augmentation is not surprising because highly augmented data sets show a poor distribution of the samples in the input space and can lead to overfitting the regions around the tie lines that were considered. From this one can infer that data augmentation is useful for phase split surrogate models with an augmentation factor of up to at least 7. Comparing the results using uniform and random data augmentation, little difference can be observed. The augmented data sets are not only used for the phase split surrogate model but also for the phase stability surrogate. The results for the latter are shown in Fig. 9.

It can be seen that data augmentation also leads to better phase stability surrogate models as the error is reduced with an increasing data augmentation factor. When comparing the two kinds of data augmentation it becomes clear that for the phase stability surrogate model the uniform data augmentation methodology leads to significantly better results if the data augmentation factor exceeds 1, i.e. points inside and outside the miscibility gap are added. This can be explained by the fact that with uniform data augmentation samples are created where they are the most useful for the phase stability surrogate model: on both sides close to the miscibility gap.

Uniform data augmentation was also applied for the multi-component system of the hydroaminomethylation process. The improvement coming from the data augmentation with a factor of 7 is shown in Table 5 for all tested combinations. The complete results can be found in the supplementary material.

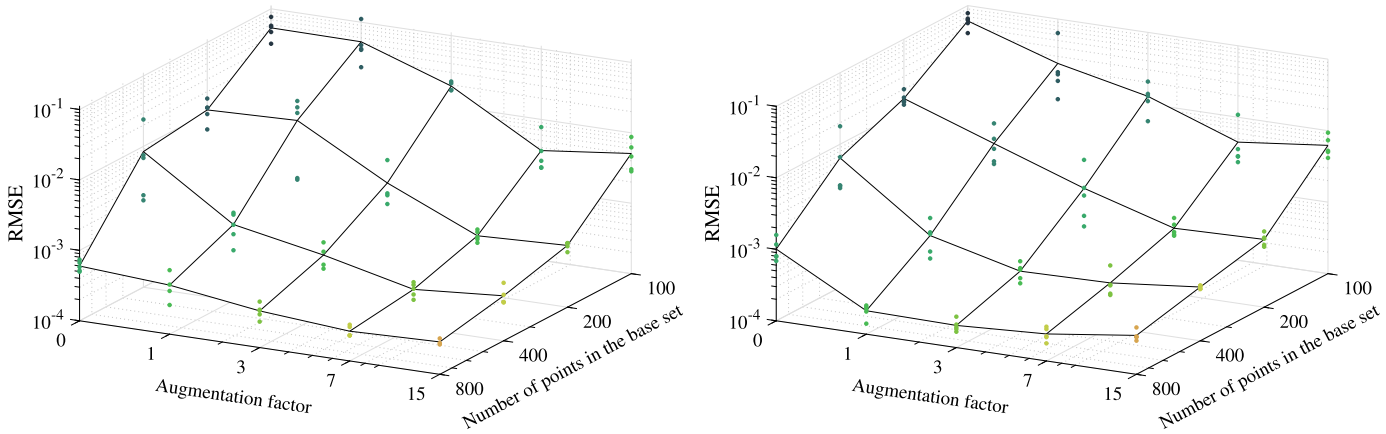
It can be seen that data augmentation reduces the error in all cases. The error reduction is larger for the HyFo system compared to the system from the HAM process. This is not surprising because the base set is much smaller for the HAM system compared to the input space. Thus, the test set error for the HAM system is dominated by points in regions where no tie lines have been evaluated. Based on the results, uniform data augmentation is chosen as the preferred methodology due to the improvement for the phase stability surrogate models and an augmentation factor of 7 is used further in this work.

6. Adaptive sampling

Another technique for making the training of surrogate models more efficient is adaptive sequential sampling. It was originally proposed by Kleijnen and Van Beers (2004), and further developed by Eason and Cremaschi (2014) under the name mixed adaptive sampling (MAS). Nentwich and Engell (2019) extended the methodology for the application of pT-flash surrogate models. In adaptive sequential sampling, a quality criterion $\eta(\underline{x})$ is defined that describes how promising it is to add a sample to the data set at some input location. Nentwich and Engell defined a quality criterion for phase split surrogate models that is shown in the set of equations (22).

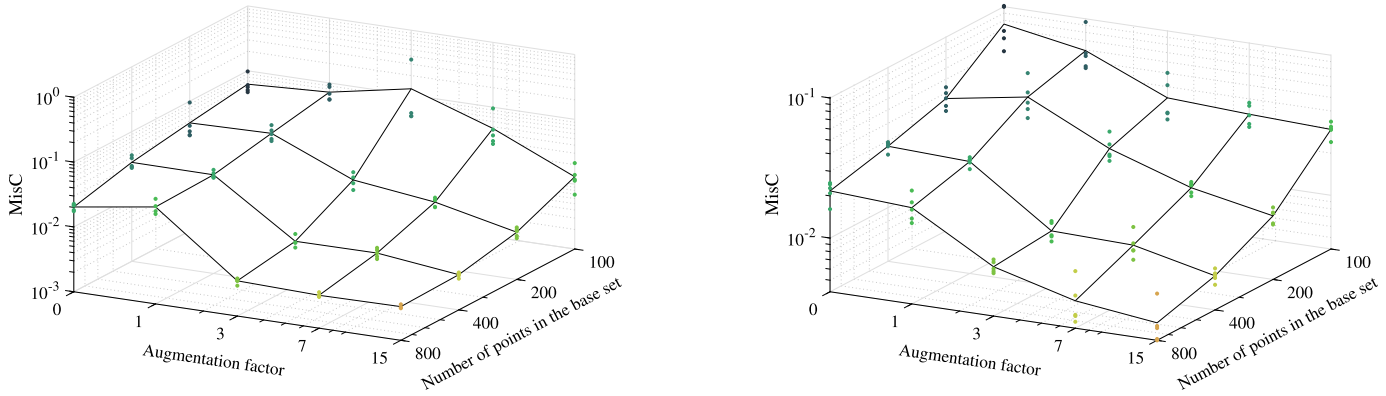
$$\begin{aligned} \eta(\underline{x}_j) &= \frac{d(\underline{x}_j)}{\max_k d(\underline{x}_k)} + \bar{\sigma}^2(\underline{x}_j) \\ \bar{\sigma}^2(\underline{x}_j) &= \frac{1}{2}\frac{\sigma^2_{stability}(\underline{x}_j)}{\max_k \sigma^2_{stability}(\underline{x}_k)} + \frac{1}{2}\bar{\sigma}^2_{split}(\underline{x}_j) \\ \bar{\sigma}^2_{split}(\underline{x}_j) &= \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{\sigma^2_{split,i}(\underline{x}_j)}{\max_k \sigma^2_{split,i}(\underline{x}_k)} \end{aligned} \tag{22}$$

Three indicators are considered in the quality criterion. The nearest neighbor distance of the j^{th} candidate input $\underline{x}_j$ with respect to the sample set and the prediction variances of the phase stability $\sigma^2_{stability}$ and phase split $\sigma^2_{split,i}$ surrogate models. One issue of this approach is the scaling with respect to the number of components. With an increasing number of components the expected maximum value of $\bar{\sigma}^2_{split}$ decreases as an average over more quantities in the interval $[0,1]$ is taken. $\eta(\underline{x})$ then becomes dominated by the nearest neighbor term and samples are also proposed far from the miscibility gap. Therefore, in this work the following modified criterion is applied:



(a) Uniform data augmentation (b) Random data augmentation

Fig. 8. Results of applying data augmentation to the multi-component system of the hydroformylation process.



(a) Uniform data augmentation (b) Random data augmentation

Fig. 9. Errors of the phase stability surrogate model for different augmented data sets as dots, the labels indicate the total number of samples in the data set after augmentation.

Table 5
Relative test set RMSE of surrogate models with 7 fold data augmentation for different numbers of samples in the base set. All RMSE values are averaged for five neural network trainings. The numbers are the ratio of the error with augmentation divided by the error without augmentations.

System Data augmentation Surrogate model	HyFo uniform split	HyFo uniform stability	HyFo random split	HyFo random stability	HAM uniform split	HAM uniform stability
100 samples	0.0673	0.7553	0.0736	0.4136	0.3192	0.5753
200 samples	0.0607	0.2187	0.0559	0.4198	0.5581	0.2726
400 samples	0.0411	0.1455	0.0652	0.3579	0.5470	0.4025
800 samples	0.4413	0.1584	0.2371	0.2989	0.4373	0.3755

$$\eta(\underline{x}) = \sqrt{\sigma^2_{\text{stability}}(\underline{x})} + (1 - \hat{f}_{\text{stability}})(\underline{x}) \sum_{i=1}^{n_c} \sqrt{\sigma^2_{\text{split},i}(\underline{x})}. \tag{23}$$

σ^2_k denotes the estimated prediction error of the surrogate model k. This criterion is continuous even when crossing the border of the miscibility gap. This criterion eliminates the influence of the variance of the phase split surrogate model outside the region where the composition is unstable. The nearest neighbor distance is not considered in order to not favor points in the single phase region far from the miscibility gap, where $\sigma^2_{\text{stability}}$ is small and $\hat{f}_{\text{stability}} \approx 1$. This is because a sample in the biphasic region is significantly more valuable than one outside of this region as it can be used in both surrogate models.

With this criterion, the adaptive sampling loop starts with an initial set of e.g. 100 samples. Surrogate models are trained on the full set and on subsets of the samples for computing the jackknife variance. Then candidate points are randomly distributed in the input space. The number of candidate points is chosen equal to number of samples in the previous iteration. The quality criterion in equation (23) is evaluated for the candidate points and a fraction of these points (e.g. 20%) with the highest values of the quality criterion are evaluated by the original function, here PC-SAFT. With the extended sample set, the next iteration is started. This sampling loop is the same as the one applied by Nentwich and Engell (2019), the only difference is the quality criterion.

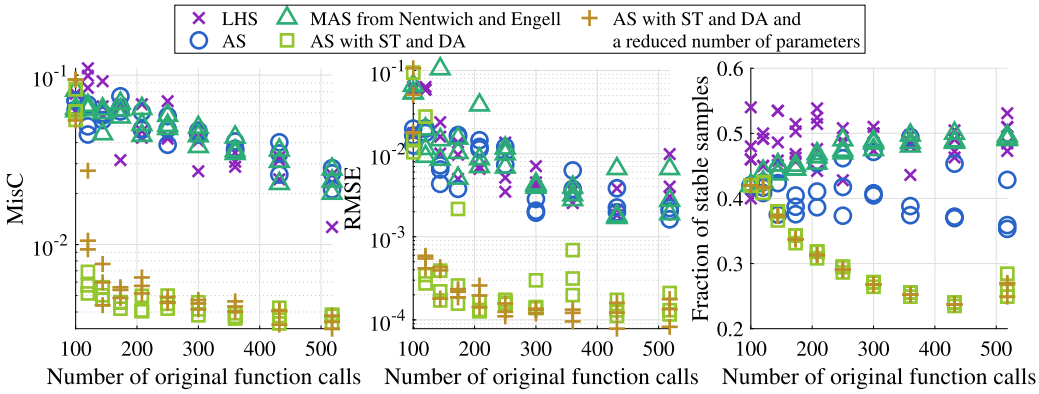


Fig. 10. Results for the multi-component system from the hydroformylation process for Latin hypercube sampling (LHS), mixed adaptive sampling (MAS) and for adaptive sampling (AS) with and without data augmentation (DA) and Sobolev training (ST) with the full and with a reduced number of parameters.

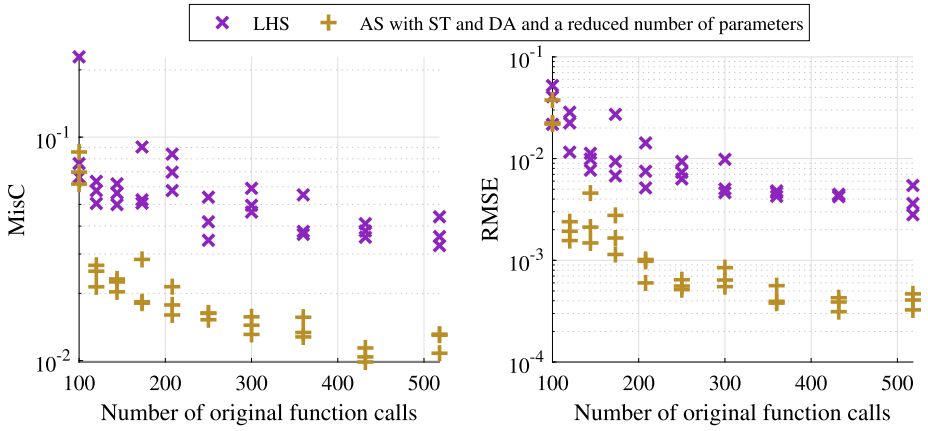


Fig. 11. Results for the multi-component system from the hydroaminomethylation process for Latin hypercube sampling (LHS) and adaptive sampling (AS) with data augmentation (DA) and Sobolev training (ST) with a reduced number of parameters.

In the adaptive sampling loop, after each iteration the sampled data points can also be augmented and the sensitivities can be computed to apply Sobolev training. For the number of parameters per data point that was determined in section 4, 0.393, the combination of Sobolev training, data augmentations and adaptive sampling was tested for the HyFo case.

Due to data augmentation, the number of training samples increases quickly. In addition, with Sobolev training the amount of data is increased by including the values of the derivatives. If the number of parameters is a fixed multiple of the number of data points, this can lead to large parameter spaces and cumbersome training. To remedy this issue, we explored the effect of reducing the number of parameters per data point to 0.1. This value seems promising considering the results in Fig. 5. The adaptive sampling results are shown in Fig. 10.

In this figure the prediction accuracy of the phase stability and phase split surrogate model are shown together with the fraction of samples that are in the stable region outside of the miscibility gap. Without data augmentation, the number of samples equals the number of original function calls. Comparing the prediction accuracy, little differences can be observed between the mixed adaptive sampling approach, the presented adaptive sampling approach using equation (23) as a quality criterion and Latin hypercube sampling. This is probably due to the relatively small number of original function calls that are considered.

Regarding the fraction of samples in the stable region, clear disparities can be observed. The MAS approach using equation (22) leads to roughly the same number of samples in both regions. The quality criterion from this work leads to only 35 to 42% samples in the stable region, which causes that data augmentation in combination with the Sobolev training approach works more efficiently because only unstable samples can be augmented.

The adaptive sampling methodology with Sobolev training and data augmentation was applied twice, using the full number of parameters and using a reduced number of parameters. It can be seen that combining the three approaches from this work, the errors of the phase stability and phase split surrogate model can be reduced by more than one order of magnitude. The training is more stable with a reduced number of parameters which gives consistently very good results. The RMSE using all three techniques in an integrated fashion is on average 1.45×10^{-4} with 500 function evaluations without reducing the number of parameters. For the pure Sobolev sampling the RMSE is on average 2.92×10^{-4} when using 300 unstable samples, which is comparable to 500 overall samples, for $\alpha = 3 \times 10^{-2}$, see Fig. 6. As shown in Fig. 8 the mean RMSE is 3.39×10^{-4} using only data augmentation on a set of 400 samples with an augmentation factor of 7. It can be seen that neither approach alone is responsible for the low error values, it is in fact the synergy between the three approaches that leads to the high accuracy of the resulting surrogate models.

As mentioned already, the reduction of the number of parameters (i.e. the reduction of the complexity of the ANN) did not lead to worse results but in some cases to a more reliable training. Therefore the full approach consisting of Sobolev training, data augmentation, adaptive sampling with a reduced number of parameters per data point was applied on the hydroaminomethylation case study. The results are shown in Fig. 11.

It can be seen that also for the larger system the prediction errors are significantly reduced using the presented approach. This validates

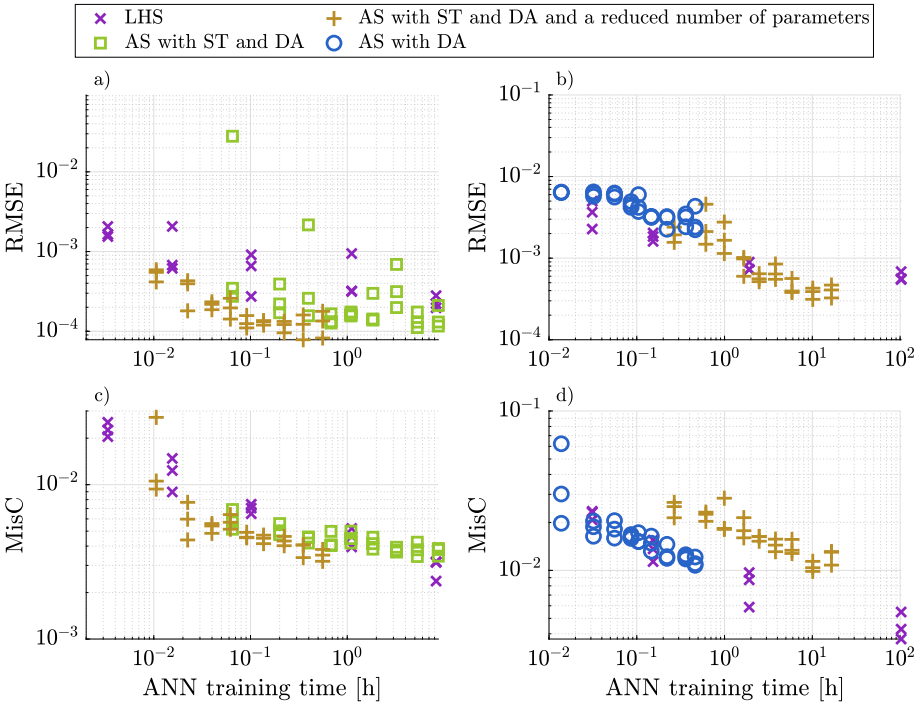


Fig. 12. Results for the multi-component system from the hydroformylation process, a) and c), and from the hydroaminomethylation process, b) and d), for Latin hypercube sampling (LHS), adaptive sampling (AS) with data augmentation (DA) and Sobolev training (ST) with the full and a reduced number of parameters and adaptive sampling with data augmentation. Errors are shown for phase split surrogate models, a) and b), and phase stability surrogate models, c) and d). Sobolev training is not applied for the phase stability surrogate in c) and d).

the assumption that the proposed hyperparameters can be generalized to different multi-component systems.

Up to this point, the aspect of computation time was not taken into account. The computational effort needed for ANN training is shown for the ANN models from the Figs. 10 and 11 in Fig. 12. Additionally, larger data sets that resulted from LHS were evaluated and the results of the training of ANN models are shown. For the system from the hydroaminomethylation process, sets of 1000, 2000, 5000 and 10000 samples were generated, for the hydroformylation system the data sets have the same size with an additional LHS set of 15000 samples. The training time is considered for training both the phase split and phase stability ANN models. The training times are averaged over 3 runs to reduce the variance from ANN trainings that stopped early because the validation error was not decreasing. The computation times do not include the time needed for the evaluation of the thermodynamic model because it depends strongly on the implementation of the PC-SAFT and pT-flash algorithm but it includes the iterative nature of the adaptive sampling, i.e. the time corresponding to the error values of the adaptive sampling in the last iteration contains the cumulative time of the ANN training of all previous iterations. The training times shown in the graphs are for the training of the full model that consists of the phase split model and the classification model. The training of the classification model alone is considerably faster than that of the phase distribution model.

The computations were done using an Intel® Core™i9-13900K processor with 24 cores and a NVIDIA GeForce RTX 4090 GPU.

For the hydroformylation case, the training of the phase split models with the extensions presented in this work is significantly more efficient than using LHS sampling, even when the number of PC-SAFT calls is not taken into account. Additionally, only 518 calls to the pT-flash routine are needed, compared to 15000 for the most accurate models obtained from LHS sampling.

The accuracy of the classification models for the HyFo case is slightly better than using LHS sampling, when excluding the number of PC-SAFT calls for the same computation time (which is indicated for the training of both models together).

For the HAM case, the proposed approach leads to the most accurate phase split models with 19 times less calls of the pT-flash routine but relatively high computation times for the training of the models. LHS sampling with large data sets produces better classification models than our approach for large data sets. This is due to the fact that our procedure only uses 518 samples in a 7-dimensional space, resulting in large regions with no available samples. In contrast, LHS sampling of 5000 or 10000 data sets avoids this issue but also comes with increased effort of data generation. Comparing the results of adaptive sampling with Sobolev training and data augmentation and adaptive sampling with data augmentation only, it can be seen that Sobolev training leads to about one order of magnitude higher computation times but also significantly lower errors which are not obtained when using even the largest LHS data sets.

7. Conclusion and outlook

In this work, different approaches are presented to improve the efficiency of approximating pT-flash computations using surrogate models. Artificial neural network (ANN) models were chosen due to their capability to deal with large numbers of data points. These models are used to predict the phase stability by a classifier and the composition of both phases for a phase split by a regression model on two complex multi-component systems with 4 and 7 species. The tuning of the training algorithms and the efficiency of the extensions of the approach in Nentwich and Engell (2019) were investigated for the quaternary system arising from a hydroformylation process. The best parameterization was then used for validation to the larger system of seven components that is relevant in a hydroaminomethylation process.

For ANN models, many hyperparameters have to be chosen before training. To find the best set of hyperparameters, a broad screening was conducted, where the number of layers in the networks, the number of parameters per data point and the regularization were varied for data sets of different sizes. The performance on a large test set was used for benchmarking. It was shown that with and without regularization sim-

ilar results can be achieved. For phase split surrogate models, deeper networks resulted in smaller errors, while for the phase stability prediction shallow networks performed well. The determined hyperparameters were also close to optimal for the larger multi-component system.

To reduce the number of samples that have to be computed by the pT-flash algorithm, three approaches were presented. Firstly, Sobolev training was introduced, where the ANN is not only trained using the output data of the pT-flash, but also the derivative of the outputs with respect to the inputs is used. This derivative information can be computed by applying a sensitivity analysis to the set of equations that determine the pT-flash solution. For this training approach, also a broad hyperparameter screening was performed. Interestingly it was found that for Sobolev training much deeper networks with up to 9 layers perform best compared to 5 layers for the standard training approach. In the Sobolev training cost function there is a weighting factor of the two error terms. Varying this weighting parameter a value could be found that leads to good performance for both case studies.

Secondly, data augmentation was applied to generate additional samples without evaluating the full thermodynamic model along each tie line. It was found that a seven-fold increase of the data set is optimal and leads to an improved accuracy for all tested surrogate models.

Finally, adaptive sampling was revisited. Here, the idea is to concentrate samples in regions of interest, e.g. in regions, where the surrogate model prediction accuracy is low. In this part we built on previous work and proposed a new quality criterion that concentrates the samples inside the miscibility gap. This is advantageous because both Sobolev training and data augmentation only use samples with an unstable feed concentration. By the new quality criterion, the adaptive sampling produces more points that enable data augmentation and Sobolev training. Another synergistic effect between the approaches occurs because using Sobolev training with data augmentation, the sensitivity analysis has to be performed only for one feed composition and can be generalized to all feed compositions without additional evaluations of the thermodynamic model.

We showed that the combination of the three approaches leads to significantly better performance than previously proposed methods when comparing the number of original function evaluations. When comparing the errors of the phase stability models with respect to the training time of the ANN, the phase split models for the hydroformylation case give significantly smaller RMSE values for the same computation time than brute-force LHS sampling with up to 29 times more samples. The classification errors are about the same as for the LHS benchmark results. The hyperparameters that had been determined for the HyFo system led to good results also for the more complex HAM case with 7 instead of 4 components. The proposed procedure gave the best results for the phase split model although at high training times. This is remarkable because here in a 7-dimensional space only 518 calls of the pT-flash routine were used. Due to the small number of samples however the classification error could not be reduced to the values that could be reached for very large data sets. For systems with many components, it may be more efficient to use only adaptive sampling and data augmentation. This enables the sampling of more distributed data before the training time for the ANN becomes large.

In further work, the generalizability of the approach and of the determined hyperparameters to different kinds of multi-component systems should be studied. In this work, the multi-component systems considered, although different in their dimension, consist of relatively similar components: polar and non-polar solvents and long-chain organic molecules. It needs to be studied how the results can be generalized for very different systems, e.g. electrolyte, polymer or VLLE systems.

Additionally, the presented approaches are expected to work well together with continuation methods that can be used to generate highly accurate initial values for the pT-flash calculations for faster sampling of the component space.

CRediT authorship contribution statement

Joschka Winz: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sebastian Engell:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This research has been supported by the project “KI-Inkubator-Labore in der Prozessindustrie - KEEN”, funded by the Bundesministerium für Wirtschaft und Klimaschutz (BMWK) under grant number 01MK20014T. This support is gratefully acknowledged.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ces.2024.120461.

References

Beale, M.H., Hagan, M.T., Demuth, H.B., 2018. Deep Learning Toolbox, User’s Guide. The MathWorks, Inc., Natick, MA.

Bhosekar, A., Ierapetritou, M., 2018. Advances in surrogate based modeling, feasibility analysis, and optimization: a review. Comput. Chem. Eng. 108, 250–267. https://doi.org/10.1016/j.compchemeng.2017.09.017.

Bianga, J., Künnemann, K.U., Goclik, L., Schurm, L., Vogt, D., Seidensticker, T., 2020. Tandem catalytic amine synthesis from alkenes in continuous flow enabled by integrated catalyst recycling. ACS Catal. 10 (11), 6463–6472. https://doi.org/10.1021/acscatal.0c01465.

Brunsch, Y., Behr, A., 2013. Temperature-controlled catalyst recycling in homogeneous transition-metal catalysis: minimization of catalyst leaching. Angew. Chem., Int. Ed. Engl. 52 (5), 1586–1589. https://doi.org/10.1002/anie.201208667.

Chimowitz, E.H., Anderson, T.F., Macchietto, S.M., Stutzman, L.F., 1983. Local models for representing phase equilibriums in multicomponent, nonideal vapor-liquid and liquid-liquid systems. 1. Thermodynamic approximation functions. Ind. Eng. Chem. Process Des. Dev. 22 (2), 217–225. https://doi.org/10.1021/i200021a009.

Chimowitz, E.H., Macchietto, S., Anderson, T.F., Stutzman, L.F., 1984. Local models for representing phase equilibria in multicomponent, non-ideal vapor-liquid and liquid-liquid systems. 2. Application to process design. Ind. Eng. Chem. Process Des. Dev. 23 (3), 609–618. https://doi.org/10.1021/i200026a034.

Czarnecki, W.M., Osindero, S., Jaderberg, M., Świrszcz, G., Pascanu, R., 2017. Sobolev training for neural networks. arXiv:1706.04859.

Dan Foresee, F., Hagan, M., 1997. Gauss-Newton approximation to Bayesian learning. In: Proceedings of International Conference on Neural Networks (ICNN’97), vol. 3. IEEE, Houston, TX, USA, pp. 1930–1935.

Eason, J., Cremaschi, S., 2014. Adaptive sequential sampling for surrogate model generation with artificial neural networks. Comput. Chem. Eng. 68, 220–232. https://doi.org/10.1016/j.compchemeng.2014.05.021.

Farzi, A., Tarjomannejad, A., 2015. Prediction of phase equilibria in binary systems containing acetone using artificial neural network. Int. J. Sci. Eng. Res. 6, 358.

Gaganis, V., Varotsis, N., 2012. Non-iterative phase stability calculations for process simulation using discriminating functions. Fluid Phase Equilib. 314, 69–77. https://doi.org/10.1016/j.fluid.2011.10.021.

Gross, J., Sadowski, G., 2001. Perturbed-chain SAFT: an equation of state based on a perturbation theory for chain molecules. Ind. Eng. Chem. Res. 40 (4), 1244–1260. https://doi.org/10.1021/ie0003887.

Guimarãaes, P.R.B., McGreavy, C., 1995. Flow of information through an artificial neural network. Comput. Chem. Eng. 19, 741–746. https://doi.org/10.1016/0098-1354(95)87123-3.

Hentschel, B., Kiedorf, G., Gerlach, M., Hamel, C., Seidel-Morgenstern, A., Freund, H., Sundmacher, K., 2015. Model-based identification and experimental validation of the optimal reaction route for the hydroformylation of 1-dodecene. Ind. Eng. Chem. Res. 54 (6), 1755–1765. https://doi.org/10.1021/ie504388t.

Hernandez, R., Dreimann, J., Vorholt, A., Behr, A., Engell, S., 2018. Iterative real-time optimization scheme for optimal operation of chemical processes under uncertainty: proof of concept in a miniplant. Ind. Eng. Chem. Res. 57 (26), 8750–8770. https://doi.org/10.1021/acs.iecr.8b00615.

Huxoll, F., Schlüter, S., Budde, R., Skiborowski, M., Petzold, M., Böhm, L., Kraume, M., Sadowski, G., 2021. Phase equilibria for the hydroaminomethylation of 1-decene. J. Chem. Eng. Data 66 (12), 4484–4495. https://doi.org/10.1021/acs.jced.1c00561.

Iftakher, A., Aras, C.M., Monjur, M.S., Hasan, M.M.F., 2022. Data-driven approximation of thermodynamic phase equilibria. AIChE J. 68 (6), e17624. https://doi.org/10.1002/aic.17624.

Ihunde, T.A., Olorode, O., 2022. Application of physics informed neural networks to compositional modeling. J. Pet. Sci. Eng. 211, 110175. https://doi.org/10.1016/j.petrol.2022.110175.

Kaiser, S., Engell, S., 2023. An integrated approach to fast model-based process design: integrating superstructure optimization under uncertainties and optimal design of experiments. Chem. Eng. Sci. 269, 118453. https://doi.org/10.1016/j.ces.2023.118453.

Kashinath, A., Szulczewski, M.L., Dogru, A.H., 2018. A fast algorithm for calculating isothermal phase behavior using machine learning. Fluid Phase Equilib. 465, 73–82. https://doi.org/10.1016/j.fluid.2018.02.004.

Kiedorf, G., Hoang, D.M., Müller, A., Jörke, A., Markert, J., Arellano-Garcia, H., Seidel-Morgenstern, A., Hamel, C., 2014. Kinetics of 1-dodecene hydroformylation in a thermomorphic solvent system using a rhodium-biphephos catalyst. Chem. Eng. Sci. 115, 31–48. https://doi.org/10.1016/j.ces.2013.06.027.

Kleijnen, J.P., Van Beers, W.C., 2004. Application-driven sequential designs for simulation experiments: Kriging metamodelling. J. Oper. Res. Soc. 55 (8), 876–883. https://doi.org/10.1057/palgrave.jors.2601747.

Kraume, M., Enders, S., Drews, A., Schomäcker, R., Engell, S., Sundmacher, K. (Eds.), 2022. Integrated Chemical Processes in Liquid Multiphase Systems: From Chemical Reaction to Process Design and Operation. De Gruyter, Berlin.

Kunde, C., Keßler, T., Linke, S., McBride, K., Sundmacher, K., Kienle, A., 2019. Surrogate modeling for liquid–liquid equilibria using a parameterization of the binodal curve. Processes 7 (10), 753. https://doi.org/10.3390/pr7100753.

Leesley, M., Heyen, G., 1977. The dynamic approximation method of handling vapor-liquid equilibrium data in computer calculations for chemical processes. Comput. Chem. Eng. 1 (2), 103–108. https://doi.org/10.1016/0098-1354(77)80015-X.

Lopez-Zamora, S., Kong, J., Escobedo, S., de Lasa, H., 2021. Thermodynamics and machine learning based approaches for vapor–liquid–liquid phase equilibria in n-octane/water, as a naphtha–water surrogate in water blends. Processes 9 (3), 413. https://doi.org/10.3390/pr9030413.

Lüken, L., Brandner, D., Lucia, S., 2023. Sobolev Training for Data-efficient Approximate Nonlinear MPC. IFAC-PapersOnLine.

Ma, K., Sahinidis, N.V., Bindlish, R., Bury, S.J., Haghpanah, R., Rajagopalan, S., 2022. Data-driven strategies for extractive distillation unit optimization. Comput. Chem. Eng. 167, 107970. https://doi.org/10.1016/j.compchemeng.2022.107970.

MacKay, D.J.C., 1992. Bayesian interpolation. Neural Comput. 4 (3), 415–447. https://doi.org/10.1162/neco.1992.4.3.415.

McBride, K., Sundmacher, K., 2019. Overview of surrogate modeling in chemical process engineering. Chem. Ing. Tech. 91 (3), 228–239. https://doi.org/10.1002/cite.201800091.

McBride, K., Kaiser, N.M., Sundmacher, K., 2017. Integrated reaction–extraction process for the hydroformylation of long-chain alkenes with a homogeneous catalyst. Comput. Chem. Eng. 105, 212–223. https://doi.org/10.1016/j.compchemeng.2016.11.019.

Merchan, V.A., Wozny, G., 2016. Comparative evaluation of rigorous thermodynamic models for the description of the hydroformylation of 1-dodecene in a thermomorphic solvent system. Ind. Eng. Chem. Res. 55 (1), 293–310. https://doi.org/10.1021/acs.iecr.5b03328.

Michelsen, M.L., Mollerup, J.M., 2007. Thermodynamic Models: Fundamentals & Computational Aspects, 2nd edition. Tie-Line Publications, Holte.

Mohanty, S., 2005. Estimation of vapour liquid equilibria of binary systems, carbon dioxide–ethyl caproate, ethyl caprylate and ethyl caprate using artificial neural networks. Fluid Phase Equilib. 235 (1), 92–98. https://doi.org/10.1016/j.fluid.2005.07.003.

Nentwich, C., Engell, S., 2019. Surrogate modeling of phase equilibrium calculations using adaptive sampling. Comput. Chem. Eng. 126, 204–217. https://doi.org/10.1016/j.compchemeng.2019.04.006.

Nentwich, C., Winz, J., Engell, S., 2019. Surrogate modeling of fugacity coefficients using adaptive sampling. Ind. Eng. Chem. Res. 58 (40), 18703–18716. https://doi.org/10.1021/acs.iecr.9b02758.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: an imperative style, high-performance deep learning library. arXiv: 1912.01703. https://doi.org/10.48550/arXiv.1912.01703, Dec. 2019.

Poort, J.P., Ramdin, M., van Kranendonk, J., Vlugt, T.J.H., 2019. Solving vapor-liquid flash problems using artificial neural networks. Fluid Phase Equilib. 490, 39–47. https://doi.org/10.1016/j.fluid.2019.02.023.

Schäfer, E., Brunsch, Y., Sadowski, G., Behr, A., 2012. Hydroformylation of 1-dodecene in the thermomorphic solvent system dimethylformamide/decane. Phase behavior-reaction performance-catalyst recycling. Ind. Eng. Chem. Res. 51 (31), 10296–10306. https://doi.org/10.1021/ie300484q.

Schlüter, S., Künnemann, K.U., Freis, M., Roth, T., Vogt, D., Dreimann, J.M., Skiborowski, M., 2021. Continuous co-product separation by organic solvent nanofiltration for the hydroaminomethylation in a thermomorphic multiphase system. Chem. Eng. J. 409, 128219. https://doi.org/10.1016/j.cej.2020.128219.

Schmitz, J.E., Zemp, R.J., Mendes, M.J., 2006. Artificial neural networks for the solution of the phase stability problem. Fluid Phase Equilib. 245 (1), 83–87. https://doi.org/10.1016/j.fluid.2006.02.013.

Vaferi, B., Rahnama, Y., Darvishi, P., Toorani, A., Lashkarbolooki, M., 2013. Phase equilibria modeling of binary systems containing ethanol using optimal feedforward neural network. J. Supercrit. Fluids 84, 80–88. https://doi.org/10.1016/j.supflu.2013.09.013.

Wang, P., Stenby, E.H., 1994. Non-iterative flash calculation algorithm in compositional reservoir simulation. Fluid Phase Equilib. 95, 93–108. https://doi.org/10.1016/0378-3812(94)80063-4.

Winz, J., Nentwich, C., Engell, S., 2021. Surrogate modeling of thermodynamic equilibria: applications, sampling and optimization. Chem. Ing. Tech. 93 (12), 1898–1906. https://doi.org/10.1002/cite.202100092.