Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Accelerating Solvent Design Optimisation with Group-Contribution Machine Learning Surrogate Classifiers

Lifeng Zhang[a], Benoît Chachuat[a], and Claire S. Adjiman[a]*

[a] Department of Chemical Engineering, The Sargent Centre for Process Systems Engineering, Imperial College London, London SW7 2AZ, United Kingdom
* Corresponding Author: c.adjiman@imperial.ac.uk.

ABSTRACT

Asserting the phase stability of multi-component mixtures is an important task in computer-aided mixture/blend design (CAM^{b}D), but it is often hindered by the lack of reliable and tractable models. In this paper, we propose a group-contribution machine-learning (GC-ML) method to predict phase coexistence for a large set of ternary mixtures consisting of two solvents and one (fixed) solute. Each solvent is represented by a vector of functional group numbers, encoded by integer values. The solvent vectors are combined with mixture composition and temperature to form the input features to a GC-ML surrogate classifier, which distinguishes between four types of stable phase configurations as possible outputs: liquid (L), solid-liquid (SL), liquid-liquid (LL) or solid-liquid-liquid (SLL). To explore the performance of the trained GC-ML multi-classifier, it is embedded as a surrogate phase-stability constraint in the optimisation of an ibuprofen crystallisation process. A two-step solution strategy is proposed, iterating between a surrogate-based subproblem and a rigorous UNIFAC-based subproblem, to design binary solvent mixtures that improve the yield of ibuprofen. A high classification accuracy score of over 0.96 is achieved in identifying the correct phase configurations with the surrogate model, making it possible to accurately predict whole phase diagrams for numerous mixtures. Moreover, reliable solutions to the crystallisation design problem are generated with the two-step strategy, yielding objective function values close to those with the UNIFAC-based model at a lower computational cost. These results reveal the value of the proposed surrogate model in guiding the search for better solvent mixtures in CAM^{b}D applications.

Keywords: Solvent design, Machine learning, Group contribution, Optimisation, Phase stability

INTRODUCTION

Identifying optimal multicomponent mixtures is critical in process design, e.g., extraction, absorption and crystallisation [1]. In the past few decades, the computer-aided mixture/blend design (CAM^{b}D) formalism has been widely used to guide the design of solvent mixtures as part of process optimisation efforts. The CAM^{b}D problem is often formulated as a mixed-integer nonlinear programming (MINLP) model, which includes process constraints and property prediction constraints [2]. In addition, ensuring the phase stability of the designed blends is important to avoid undesirable demixing behaviour of the mixture during operation. Given a mixture at a specific temperature, pressure, and composition, phase stability is achieved at the global minimum of the Gibbs free energy of the mixture. It can be enforced by ensuring the sign of the minimum tangent plane distance (TPD) is nonnegative [3]. However, the need to use nonlinear thermodynamic models to model complex mixtures makes TPD minimisation a nonconvex optimisation problem. Embedding this nonconvex problem within the CAM^{b}D results in a bilevel problem that is highly challenging and expensive to solve within reasonable runtime in practice. Although tailored optimisation strategies and algorithms have been developed to address the phase stability problem [4,5] or to embed heuristic phase stability checks within CAM^{b}D [6], this remains a hurdle to designing stable mixtures.

Recently, the idea of replacing the lower-level phase stability constraint in the CAM^{b}D problem with a data-

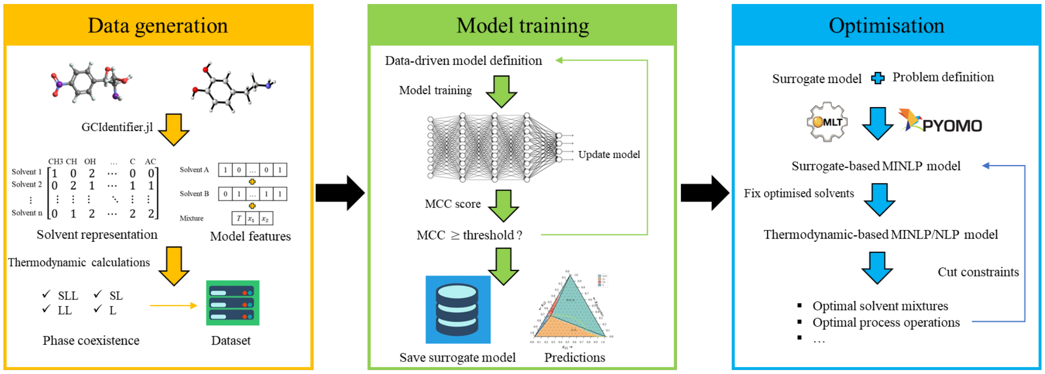


Figure 1: Overview of the main steps, namely data generation, model training and optimisation in proposed framework to develop a surrogate ANN multi classifier and implement it in the solvent design optimisation problem.

driven machine learning model has been gaining attention. The use of artificial neural networks (ANNs) as surrogate models has been investigated due to their ability to achieve high accuracy and the potential to embed them as linear constraints, which are preferable in optimisation [7]. In principle, this approach provides a way to simplify nonconvex problems and to reduce computational cost. ANNs [8,9] and support vector machines (SVMs) [10] have previously been trained to predict the phase stability of a specific mixture and have been found to reduce computational cost. However, these mixture-specific models have limited potential for use in CAM^{b}D when many mixture candidates are considered in the design space. To address this, Karia et al. [11] trained an ANN classifier to predict the phase stability of all binary mixtures for a set of predefined solvent candidates, using parameters that describe the solvent identities along with composition and temperature as inputs. The classifier model was used to replace the phase stability criterion and was found to lead to improvements in optimisation efficiency and solution quality. This work was extended by considering SVM classifiers and introducing an additional surrogate model to replace the phase equilibrium (solubility) constraints [12], demonstrating further the potential of data-driven models in CAM^{b}D.

Although these studies revealed that training surrogate models for phase stability prediction can lower computational cost, the exploration of solvent mixtures beyond the dataset is restricted as the solvent identities are linked to a subset of the UNIFAC parameters [13], namely surface and volume. Furthermore, the need to train different surrogate models for solubility and phase stability prediction increases model complexity. To overcome these limitations, a group-contribution machine learning (GC-ML) ANN is developed in this work to predict the phase behaviour of ternary mixtures. In GC-ML, the number and identity of functional groups in a molecule are used as inputs to an ML model to predict a target property. It has been used for pure component properties [14,15], but, to the best of our knowledge, not for mixture properties. Here, each solvent is represented by a vector of functional groups, resulting in a more general surrogate model than that in [11,12]. In addition, a multi-class classifier is developed to determine phase coexistence, instead of relying on a binary classifier for phase stability; this gives the ability to generate entire phase diagrams. The framework for constructing the GC-ML model is described in detail in the next section. The design of a binary solvent mixture for ibuprofen crystallisation is next presented as a case study. Conclusions are developed in the final section.

METHODOLOGY

The model construction is described in the context of a binary solvent mixture and a solute at conditions where the solute may crystallise and/or the ternary mixture may exhibit one or more liquid phases. It is assumed that the pressure p is fixed but the temperature T is variable. The approach can be extended to other mixtures or phase equilibria. Figure 1 illustrates the steps required to construct the surrogate model, including data generation and model training, and to use it in optimisation.

Data generation

A set of N_S common solvent candidates is collected from [16]. Each solvent is decomposed into N_G groups by GCIdentifier.jl [17] in Julia. The groups can either be UNIFAC functional groups [13] or SAFT-γ Mie [18] functional groups. The input features for the GC-ML model are two N_G-dimensional vectors of group numbers describing each solvent, a 2-dimensional mole fraction vector and the temperature. To obtain a diverse dataset that includes all the solvent candidates, each component is paired with N_p other solvents to create $N_S \times N_p$ solvent mixtures. Notice that not all possible mixtures within the design space are selected. A dataset with N points is generated by adding the solute (e.g., ibuprofen) to each solvent pair and sampling the space of compositions via a Sobol' sequence [19] and the space of temperatures at regular intervals. A pT flash calculation is executed for each sample with the rigorous model (phasepy [20] with UNIFAC), to identify the stable phase configuration. There are four possible outcomes or categories: liquid (L), solid-liquid (SL), liquid-liquid (LL) or solid-liquid-liquid (SLL). The resulting dataset is split into a training dataset ($80\%N$) and a test dataset ($20\%N$). The sampled mixtures are distributed across both datasets to ensure that the model effectively learns from each functional group.

Model training

The ANN classifier is developed in Pytorch [21]. The classifier structure, including the number of layers, the number of neurons per layer and the activation function, is defined before training. For the i^{th} input, the ANN yields four real output values, $z_{i,k}$, where $k \in K = \{L, SL, LL, SLL\}$ represents the category. These outputs are normalised into the likelihood $\hat{y}_{i,k}$ via a softmax function:

$$\hat{y}_{i,k} = \frac{\exp(-z_{i,k})}{\sum_{kk\in K}\exp\,(-z_{i,kk})} \quad \forall k \in K, i = 1, \dots, N \qquad (1)$$

The binary label $y_{i,k}$is then set to 1 for $k = \underset{k\in K}{\operatorname{argmax}}\, \hat{y}_{i,k}$ and to 0 for other values of k and the cross-entropy loss function L is given by:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k\in K} y_{i,k}\log(\hat{y}_{i,k}) \qquad (2)$$

Once the loss function has been minimized, the chosen architecture is assessed. Given that the dataset is extremely imbalanced since data points in categories LL and SLL are rare, the Matthews Correlation Coefficient (MCC) [22] is applied as a metric to evaluate model performance instead of the more usual accuracy metrics. The MCC score is defined as follows:

$$MCC = \frac{\sum_k\sum_l\sum_m C_{kk}C_{lm} - C_{kl}C_{mk}}{\sqrt{(\sum_k\sum_l C_{kl})(\sum_k\sum_l C_{lk})(\sum_k\sum_m C_{km})(\sum_l\sum_m C_{lm})}} \qquad (3)$$

Here C_{lm} refers to the number of samples in the confusion matrix which are predicted as belonging to category m while the true category is l. The MCC score takes into account all the prediction results and varies from -1 (a totally wrong prediction) to 1 (a perfect prediction). The MCC score for the test dataset is compared to a threshold MCC_{min} and the model architecture is modified by either adding hidden layers or increasing the number of neurons until the threshold is reached.

Optimisation strategy

To use the surrogate within an optimisation problem, it is converted into a set of algebraic constraints via the OMLT [23] and embedded into the solvent design optimisation problem in Pyomo [24]. To solve the overall design problem, a two-step approach is implemented. First, the surrogate-based model is formulated by replacing all equations related to phase equilibrium or stability calculations, including the thermodynamic model, with the trained ANN classifier. The surrogate-based model involves binary and continuous variables and may be linear or nonlinear depending on the activation function. Additionally, process constraints may also be nonlinear. With the surrogate model, the phase stability and phase equilibria are replaced by constraints based on the predicted probability values $\hat{y}_k, k \in K$,. To ensure a stable liquid phase, the following constraints are used to ensure $k = L$ has the highest probability:

$$\hat{y}_L \geq \hat{y}_k \quad \forall k \in \{SLL, SL, LL\} \qquad (4)$$

To approximate the phase equilibrium constraint, such as SL equilibrium (solubility), the difference between the likelihoods of the two relevant categories, $\hat{y}_{SL}$ and $\hat{y}_L$, should be within a predefined error ε:

$$|\hat{y}_L - \hat{y}_{SL}| \leq \varepsilon \qquad (5)$$

The resulting surrogate-based MILP or MINLP is then solved to obtain optimal solvent identities, compositions, and process conditions. Only the optimal solvent identities are retained for the second step. The second optimisation subproblem is an NLP (fixed solvent identities) which includes the rigorous phase equilibrium constraints and a surrogate phase stability check, formulated as the following constraint [11]:

$$\hat{y}_L + \hat{y}_{SL} \geq \hat{y}_{LL} + \hat{y}_{SLL} \qquad (6)$$

This constraint ensures that the probability of having one stable liquid phase is larger than that of having two stable liquid phases. If desired, the process of generating candidate solvent mixtures can be repeated by adding integer cuts to eliminate previous solutions.

CASE STUDY

Problem description

The proposed framework is demonstrated on an ibuprofen crystallisation process [25]. At the initial state, the ibuprofen is dissolved in a binary solvent mixture consisting of solvents s_1 and s_2 at temperature T_0. There are no crystals in this state. During the crystallisation, more solvent s_2 can be added to the mixture as an antisolvent and the mixture can be cooled to temperature T. In the present case study, the ranges of T_0 and T are taken to be $[293.15\,K,\ 313.15\,K]$ and $N_S = 48$ solvent candidates are considered. The goal of the optimisation is to select the binary solvent mixture from the N_s solvent candidates (2,256 possible mixtures), as well as the mixture composition and temperature, that maximise the yield of ibuprofen. The solvents combine any of $N_G = 33$ UNIFAC functional groups and the predicted mixture properties are compared against UNIFAC [13] as the rigorous thermodynamic model. More details on this case study can be found in [11,25].

All models are trained and optimised on Windows11 using Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz with 64GB RAM and NVIDIA RTX A4000 GPU with 16GB RAM.

Performance of multi-class classifier

A parametric study is carried out for the given number of solvent candidates and temperature ranges to determine the appropriate size of the dataset for training. As a result, N_p=30 is applied, leading to $N_s \times N_p$=1,440 randomly generated solvent mixtures. The temperature range is sampled every 2 K, leading to 11 temperature values. For each solvent mixture at each temperature, 400 mole fraction vectors are sampled, leading to a large set of inputs that consists of 6,336,000 data points. The phasepy calculations are conducted in parallel and require nearly 3 hours of wall time. An overview of the dataset is presented in Table 1.

Table 1: Overview of the dataset.

Category	# input vectors	% category
LLS	371,350	5.86%
LL	183,962	2.90%
LS	3,428,021	54.10%
L	2,333,993	36.84%
E	18,674	0.30%
# inputs	6,336,000	100.00%
# valid inputs	6,317,326	99.70%

Category E refers to input vectors for which phasepy does not converge. This small subset is discarded. As expected, the LS and L categories are most abundant in the dataset, making up 90% of the data.

Since 33 functional groups are used to define one solvent, the input vector has 69 elements consisting of the two solvents, the temperature as well the mole fraction of each solvent. As shown in Table 2, the final ANN includes a first hidden layer of 80 neurons and a second layer with 30 neurons. The number of epochs is set to 30, and the batch size is set to 2,056 with a learning rate of 0.005. The tanh activation function is used based on previous experience to avoid introducing redundant binary variables that can slow down convergence [12]. The training time is mostly determined by the large dataset.

Table 2: Training statistics for the final ANN model.

Parameters	Values
Activation function	tanh
Model complexity	[69,80,30,4]
Epochs	30
Batch size	2,056
Learning rate	0.005
Training time /s	14,032
MCC for training dataset	0.9771
MCC for test dataset	0.9698

Encouragingly, for a value of $MCC_{min} = 0.96$, a surrogate model that achieves a high MCC score for both the training dataset (0.9771) and the test dataset (0.9698) is obtained. This indicates that the surrogate model captures the nature of the phase coexistence at different temperatures and for different mixtures. To investigate the performance of the trained ANN multi-class classifier, the model is used to generate phase diagrams for four combinations of solvent 1/solvent 2/temperature that do not appear in the dataset: chloroform-tetrachloromethane (CCl4) at T=300.15 K, water-chloroform at T=300.15 K, water-toluene at T=308.15 K and ethylene glycol-anisole at T=308.15 K. These are shown in Fig. 2, where the predictions from the classifier are marked with different colours. The true phase boundaries are also obtained with the rigorous UNIFAC model and are denoted by black solid curves. The yellow dashed curves represent metastable solid-liquid equilibrium points in a region of solid-liquid-liquid equilibrium. The predicted phase boundaries are in good agreement between the two models. For example, the chloroform-CCl_4 mixture (Fig. 2a) only displays SL and L, indicating that no liquid-liquid demixing occurs. The solubility of ibuprofen, denoted by the black solid line, is described well by the classifier. For the remaining mixtures, which all exhibit two liquid phases, SLL and LL regions are indeed identified. For water-chloroform and water-toluene (Figs. 2b-c), misclassification occurs when the composition is close to pure s_1, predicting SL instead of the true LL and SLL regions. This probably arises from the presence of other mixtures in the dataset that exhibit this behaviour. For ethylene glycol-anisole (Fig. 2d), the surrogate model does not capture the extent of the SLL region very well. There are 254,808 data points that involve anisole, only 4.69% of which are labelled as SLL. The model tends to predict ethylene glycol-anisole mixtures forming a single liquid (SL or L), leading to an underestimation of the SLL region but accurate identification of a tiny L region of pure $s1$ between L and LS, which is most important in solvent design for crystallisation. Overall, the data-driven model captures the phase coexistence of multi-component mixtures well. Finally, parts of the solid-liquid equilibrium boundary that correspond to a metastable equilibrium are shown with yellow curves in Fig. 2. If the phase stability constraint were not included in the solvent design, solutions that fall on these undesirable parts of the curve would likely be identified.

Optimisation performance

The problem is solved with the proposed two-step approach. Details on the process can be found in [11]. Here a summary of the surrogate-based model is given:

$$\max \ \text{yield of ibuprofen} \tag{7}$$

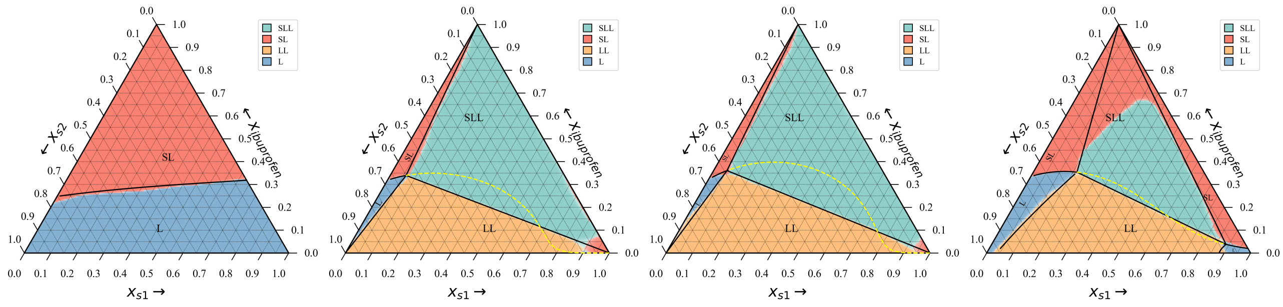
$$\text{s.t. process constraints} \tag{8}$$

$$\text{mixture formulation constraints} \tag{9}$$

$$\hat{y}_{k,st} = f_{ANN}(x_{in,st}) \ \forall k \in K, st \in \{\text{initial}, \text{final}\} \tag{10}$$

$$|\hat{y}_{L,st} - \hat{y}_{LS,st}| \le \varepsilon \ \forall st \in \{\text{initial}, \text{final}\} \tag{11}$$

$$\hat{y}_{L,st} \ge \hat{y}_{k,st} \ \forall k \in K \backslash \{L\}, st \in \{\text{initial}\} \tag{12}$$



a). Chloroform : CCl_4 at 300.15 K

b). Water : Chloroform at 300.15 K

c). Water : Toluene at 308.15 K

d). Ethylene glycol : Anisole at 308.15 K

Figure 2: Phase diagrams generated by the surrogate model and the rigorous model for four solvent mixtures at 300.15 K or 308.15 K. The mixture is shown as $s_1 : s_2$ in each figure caption. The colours denote the prediction from the surrogate model. The solid curves are the phase boundaries generated by pT flash calculations with UNIFAC. The yellow curves denote parts of the solid-liquid equilibrium boundaries that are metastable, as computed with UNIFAC. The regions are labelled L, LL, SL and SLL based on the results of the rigorous model.

$$\hat{y}_{SL,st} \geq \hat{y}_{k,st} \; \forall k \in K \backslash \{SL\}, st \in \{\text{final}\} \qquad (13)$$

Equation (10) represents the ANN with $x_{in,st}$ referring to the input vector at state $st \in \{\text{initial, final}\}$. A value of $\varepsilon = 0.1$ is used in Equation (11). Equation (12) ensures the mixture at the initial state is in region L. It is not necessary to impose that the mixture be saturated at the initial temperature, although it is desirable as it leads to lower solvent consumption. Equation (13) ensures the final mixtures is in the SL region after crystallisation, while Equation (11) ensures it is close to equilibrium solubility.

Neither of the models can be solved within 3 hours with global solvers. Instead, sBB [26] is used to solve the surrogate-based MINLP and CONOPT [27] is used for the UNIFAC-based NLP. Five solutions are listed in Table 3, including the runtime needed to generate each solution, the yield of ibuprofen from each step, and the outcome of the phase stability check in phasepy. All the solutions have the same optimal temperature range, from 313.15 K to 293.15 K, except for the first solution whose final temperature is 302.06 K. Every subproblem can be solved within 3 minutes. This small runtime makes a more extensive search of the design space possible. Ibuprofen yields of around 0.77 are obtained from all instances of step 1. In all cases apart from iteration 3, good agreement between the yields from the two subproblems is found. In iteration 3, a deviation of 0.051 is observed, arising from misclassifications for the benzene-p-cymene mixture. All the solvent mixtures designed only present one liquid phase at both the initial and final states, which satisfies the stability check. This is in contrast to approaches in which this constraint is only enforced *a posteriori* [25]. Applying data-driven surrogates of phase stability to assist solvent design in CAM^{b}D results not only in lower computational cost but, more importantly, in more reliable solutions.

CONCLUSIONS AND FUTURE WORK

We proposed a multi-class classifier for predicting phase coexistence and stability. The group contribution concept was adopted to represent solvent candidates with a vector of functional groups, leading to a GC-ML ANN that can predict a whole phase diagram. This framework was tested on a crystallisation optimisation problem with ibuprofen as the target compound. The results showed that the trained classifier can generate phase diagrams for unseen mixtures with sufficient accuracy for use in CAM^{b}D. A two-stage strategy was proposed to obtain solutions to the crystallisation process optimisation problem. High-quality designs were obtained, exhibiting a homogeneous liquid phase as desired. The proposed classifier surrogate offers a promising alternative for the formulation of CAM^{b}D problems. However, misclassifications arising from the data-driven model can occasionally misdirect the search to a suboptimal or infeasible solution. Future work will thus focus on developing strategies such as active learning to improve model accuracy.

Table 3: Optimal binary solvent mixtures for the ibuprofen crystallisation process via the optimisation.

Iter.	**Runtime /s**	**Solvent A: Solvent B**	**Yield** (Step 1)	**Yield** (Step 2)	**Phase stability (initial: final)**
1	137.20	DMSO: Formic acid	0.784	0.799	L: L
2	74.14	Formic acid: Anisole	0.779	0.770	L: L
3	112.72	Benzene: p-Cymene	0.779	0.728	L: L
4	145.23	Anisole: Formic acid	0.766	0.770	L: L
5	124.84	Chloroform: Hexane	0.764	0.752	L: L

ACKNOWLEDGEMENTS

The authors gratefully acknowledge funding from the UK Engineering and Physical Sciences Research Council (EPSRC) under grant EP/W003317/1.

REFERENCES

1. Jonuzaj, S. *et al*. The formulation of optimal mixtures with generalized disjunctive programming: A solvent design case study. *AIChE J.* **62**, 1616–1633 (2016).
2. Gani, R. Computer-Aided Methods and Tools for Chemical Product Design. *Chem. Eng. Res. Des.* **82**, 1494–1504 (2004).
3. Baker, L. E., Pierce, A. C. & Luks, K. D. Gibbs Energy Analysis of Phase Equilibria. *Soc. Pet. Eng. J.* **22**, 731–742 (1982).
4. Mitsos, A. & Barton, P. I. A dual extremum principle in thermodynamics. *AIChE J.* **53**, 2131–2147 (2007).
5. McDonald, C. M. & Floudas, C. A. Global optimization for the phase stability problem. *AIChE J.* **41**, 1798–1814 (1995).
6. Karunanithi, A. T., Achenie, L. E. K. & Gani, R. A computer-aided molecular design framework for crystallization solvent design. *Chem. Eng. Sci.* **61**, 1247–1260 (2006).
7. Schweidtmann, A. M. & Mitsos, A. Deterministic Global Optimization with Artificial Neural Networks Embedded. *J. Optim. Theory Appl.* **180**, 925–948 (2019).
8. Schmitz, J. E., Zemp, R. J. & Mendes, M. J. Artificial neural networks for the solution of the phase stability problem. *Fluid Phase Equilibria* **245**, 83–87 (2006).
9. Lopez-Ramirez, E. *et al*. Artificial Neural Networks (ANNs) for Vapour-Liquid-Liquid Equilibrium (VLLE) Predictions in N-Octane/Water Blends. *Processes* **11**, 2026 (2023).
10. Gaganis, V. & Varotsis, N. Non-iterative phase stability calculations for process simulation using discriminating functions. *Fluid Phase Equilibria* **314**, 69–77 (2012).
11. Karia, T. *et al*. Classifier surrogates to ensure phase stability in optimisation-based design of solvent mixtures. *Digit. Chem. Eng.* 100200 (2024).
12. Zhang, L. *et al.* Comparative study of classifier models to assert phase stability in multicomponent mixtures. in *Computer Aided Chemical Engineering* (eds. Manenti, F. & Reklaitis, G. V.) vol. 53 1465–1470 (Elsevier, 2024).
13. Published Parameters UNIFAC - DDBST GmbH. https://www.ddbst.com/published-parameters-unifac.html.
14. Mann, V., Gani, R. & Venkatasubramanian, V. Group contribution-based property modeling for chemical product design: A perspective in the AI era. *Fluid Phase Equilibria* **568**, 113734 (2023).
15. Paduszyński, K. & Domańska, U. Viscosity of Ionic Liquids: An Extensive Database and a New Group Contribution Model Based on a Feed-Forward Artificial Neural Network. *J. Chem. Inf. Model.* **54**, 1311–1324 (2014).
16. Prat, D. *et al.* CHEM21 selection guide of classical- and less classical-solvents. *Green Chem.* **18**, 288–296 (2016).
17. Walker, P. J., Riedemann, A. & Wang, Z.-G. GCIdentifier.jl: A Julia package for identifyingmolecular fragments from SMILES. *J. Open Source Softw.* **9**, 6453 (2024).
18. Haslam, A. J. *et al.* Expanding the Applications of the SAFT-γ Mie Group-Contribution Equation of State: Prediction of Thermodynamic Properties and Phase Behavior of Mixtures. *J. Chem. Eng. Data* **65**, 5862–5890 (2020).
19. Sobol', I. M. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. Math. Phys.* **7**, 86–112 (1967).
20. Chaparro, G. & Mejía, A. Phasepy: A Python based framework for fluid phase equilibria and interfacial properties computation. *J. Comput. Chem.* **41**, 2504–2526 (2020).
21. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. Preprint at http://arxiv.org/abs/1912.01703 (2019).
22. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
23. Ceccon, F. *et al.* OMLT: Optimization & Machine Learning Toolkit.
24. Bynum, M. L. *et al. Pyomo — Optimization Modeling in Python*. vol. 67 (Springer International Publishing, Cham, 2021).
25. Watson, O. L. *et al.* Computer Aided Design of Solvent Blends for Hybrid Cooling and Antisolvent Crystallization of Active Pharmaceutical Ingredients. *Org. Process Res. Dev.* **25**, 1123–1142 (2021).
26. SBB. https://www.gams.com.html.
27. Drud, A. CONOPT: A GRG code for large sparse dynamic nonlinear optimization problems. *Math. Program.* **31**, 153–191 (1985).