HybridGamma: A thermodynamically consistent framework for hybrid modelling of activity coefficients

Ulderico Di Caprio, Jan Degrève, Peter Hellinckx, Steffen Waldherr, M. Enis Leblebici

PII: S1385-8947(23)04835-0
DOI: https://doi.org/10.1016/j.cej.2023.146104
Reference: CEJ 146104

To appear in: *Chemical Engineering Journal*

Received Date: 28 April 2023
Revised Date: 6 September 2023
Accepted Date: 15 September 2023

Please cite this article as: U. Di Caprio, J. Degrève, P. Hellinckx, S. Waldherr, M. Enis Leblebici, HybridGamma: A thermodynamically consistent framework for hybrid modelling of activity coefficients, *Chemical Engineering Journal* (2023), doi: https://doi.org/10.1016/j.cej.2023.146104

hybridGamma: a thermodynamically consistent framework for hybrid modelling of activity coefficients

Ulderico Di Caprio[1], Jan Degrève[2], Peter Hellinckx[3], Steffen Waldherr[4,5], M. Enis Leblebici[1,*]

[1] Center for Industrial Process Technology, Department of Chemical Engineering, KU Leuven, Agoralaan Building B, 3590 Diepenbeek, Belgium

[2] Process Engineering for Sustainable Systems, Department of Chemical Engineering, KU Leuven, Celestijnenlaan 200F, B-3001 Heverlee, Belgium

[3] Faculty of Applied Engineering, University of Antwerp, Groenenborgerlaan 171, 2000 Antwerp, Belgium

[4] Chemical Reactor Engineering and Safety (CREaS), Department of Chemical Engineering, KU Leuven, Celestijnenlaan 200F, B-3001 Heverlee, Belgium

[5] University of Vienna, Molecular Systems Biology (MOSYS), Department of Functional and Evolutionary Ecology, Faculty of Life Sciences, 1030 Vienna, Austria

*Corresponding author (muminenis.leblebici@kuleuven.be)

Abstract

Predicting molecular interactions is a crucial step for chemical process modeling. It requires the full knowledge of the analyzed system, however, this is often impossible in complex real-world cases. Machine learning (ML) techniques overcome this bottleneck and enhance systems predictability using data. Hybrid modeling (HM) is an established technique combining first-principle information and ML techniques. This work introduces a mathematical framework to predict activity coefficients employing HM approach. The obtained models are physically consistent and can handle systems with unknown components or external sources of deviation. The framework is validated on experimental and in-silico cases employing different training approaches. In all the tested cases, the HM showed remarkable prediction capabilities with coefficients of determination R^2 above 0.98 for the predicted variables. This work proposes and develops a novel way to approach the HM of molecular interactions by embedding physical laws within the model structure. We encountered three main benefits in applying thermodynamically consistent HMs for activity coefficients: the reduction of tenable parameters, the increased prediction capabilities, and the physical-consistent behavior of the model.

Keywords: activity coefficients, hybrid model, physical consistency, vapor-liquid equilibria, Gibbs-Duhem equation

1. Introduction

The use of computational tools for process optimization and intensification represents state-of-the-practice in the processing system engineering field, but they require using predictive models. Nowadays, the available models are generally able to simulate the behavior of complex processes representing a cheap and fast alternative to real-life experiments. Typically, a process simulation model contains multiple layers of sub-models predicting the physical phenomena driving the system [1]. Among these, the sub-model predicting the molecular interacting energies plays a crucial role [1,2], especially in all the scenarios where the system shows a significant deviation from the ideal behavior. In literature, it is well known how the abovementioned deviations affect several aspects of the physical-chemical system, such as phase stability, vapor-liquid equilibrium, and solubility [3–6]. Running an optimization with such an uncertain model leads to uncertain system behavior predictions [7], severely limiting optimization reliability. For all these reasons, modeling non-ideal behaviors represents one of the leading research fields in process system engineering.

The deviations from the ideal behavior can be characterized by the excess of Gibbs energy or, equivalently, activity coefficients (ACs). Several models have been developed in the last century to estimate the ACs in non-ideal mixtures utilizing various assumptions. These models are usually constructed using first principles, assuming the binary molecular interaction to be the most probable in the system. Widely used and advanced models to predict ACs and excess Gibbs energy are UNIFAC [8,9], NRTL [10] and COSMO-RS [11,12]. The wide application and usage of these models proved their robustness and flexibility. In addition, they have been applied to many systems over the past decades. In addition, recently, machine learning (ML) frameworks predicting ACs using molecular representations of the involved molecules have been released [13–15]. However, all the methodologies available in the literature require fully characterizing the mixture to predict its behavior.

In several cases, the composition of the mixture is partially unknown; therefore, the molecular interactions are hard to estimate. For example, in a telescopic reaction process, the main reaction path is known. Still, side reactions can often take place, and the molecular structure of the side products and their concentrations are often unknown, bringing uncertainties related to the system composition. A second example where the first-principle ACs estimation shows high deviation from the real world is the presence of a complex mechanism of molecular release. It could be the case of a fragrance deposited over a fabric or within a capsule for the release control; here, the fabric fiber nature and the release control mechanism can significantly impact the molecular interaction and, therefore, their evaporation. In literature, few attempts have been made to tackle this problem and estimate ACs of poorly specified mixtures. Mathias proposed using a perturbation scheme on the ACs based on the Margules equation [16]. The proposed framework worked with any model predicting ACs, and the perturbation was drawn as a set of pseudo binary interactions, modeled using Margules equations. A similar perturbation scheme was expanded by Burger et al. to be applied to any equation of state [17]. Another significant contribution from Baumeister and Burger highlighted the possibility of utilizing a perturbation scheme to describe the deviation of ACs due to unknown components [18]. The three methods listed above do not rely on analytical techniques to estimate the activity coefficients in poorly specified mixtures. However, analyzing the mixture using spectroscopic techniques and involving group contribution methods to estimate the mixture excess properties is still possible. In this regard, many works have been published [19–24]. Furthermore, some frameworks have been developed to estimate the propagation of the uncertainty from the physical parameters (e.g., activity coefficients and excess properties) to the flowsheet simulation and optimization [25,26].

Recently, ML techniques have shown significant potential in chemical engineering [27] especially in scenarios where process parameters are hard to estimate using first-principle modeling [28–31]. Such methods utilize experimental data and a statistical function structure (e.g., polynomial, rational functions, and artificial neural networks) to approximate the underlying function by characterizing the described quantity of interest while simultaneously creating an internal map of the relation between the input and the described quantity. In recent years, ML approaches are gaining popularity in describing chemical systems. Despite the advanced modeling capabilities and the interest gained in chemical engineering, the ML techniques still demand data respecting strict conditions [32] and they are known for the low extrapolation capabilities behind experimental conditions explored in the training phase [33]. Hybrid modeling is gaining popularity for chemical engineering applications to solve the problems mentioned above, as it combines first-principle and ML models. In such a technique, physical laws and information derived from data are combined, increasing the data contextualization within the system. Consequently, the known part of the system is described using a first-principle model, while the ML technique describes only the unknown part of the system [34]. By doing so, the required data variety and volume are drastically reduced. This modeling approach has been used in several fields of chemical engineering, including reaction kinetics estimation [35,36], separations [29,37] and overall optimization [38]. Despite the wide application of hybrid modeling in process systems engineering for chemical applications, to our best knowledge, the literature lacks papers and methodologies where hybrid modeling is applied to estimate the interactions between molecules for systems with uncertainty over the interactions between the molecules and their nature. A hybrid model of the excess properties is a non-trivial task since the models describing molecular interactions must respect a-priori laws such as the Gibbs-Duhem equation for excess properties [39]. If the model violates these rules, the obtained predictions violate a fundamental constrain of thermodynamic. This drastically decreases the reliability, generalization, and extrapolation capabilities of the model since its prediction have higher degrees of freedom than is in the reality. Therefore, when predicting excess properties with ML techniques, it is crucial to constrain the model with such an equation. Recently, Carranza-Abaid et al. applied neural network programming to create ML models predicting activity coefficients while respecting the Gibbs-Duhem equations [40]. This effort represents a significant step toward integrating first-principle knowledge within ML techniques; however, it still requires complete knowledge of the system in analysis. This work investigates the application of machine learning and hybrid modeling techniques to predict molecular interactions for poorly specified conditions.

Despite the lack of effort in applying hybrid models for poorly specified mixtures, hybrid modeling has been widely applied in the thermodynamic modeling of fluid mixtures. In their review, Jirasek and Hasse identify four main ways this has been done: dataset combination, model interconnections, embedding ML techniques in physical models, and embedding physical models in ML techniques [41]. One of the main problems in applying ML for thermodynamic tasks is the scarcity of data. However, data combinations can overcome this issue. Simulation data can pre-condition the model in a preliminary training step. Following, the parameters are refined using experimental data. An example of this approach is reported in recent works applying matrix completion techniques to predict activity coefficients at infinite dilution and Henry's coefficients [42,43]. Thermodynamic properties predictions can rely on a wide availability of first-principle models able to calculate the excess properties of many mixtures; however, the prediction could deviate from the measured values. Nevertheless, correcting the obtained values using ML techniques creates and accurate hybrid model for predicting the excess properties. Sanches Medina et al. proposed a framework able to correct the predictions of the mechanistic models for the activity coefficients using the molecular structure of the components within the mixture and graph network to predict the deviation level [44]. A further way to create a hybrid model for thermodynamic properties is by embedding first-principle information in ML techniques or vice versa. In this regard, physics-informed neural networks consider the numerical satisfaction of physical relationships (e.g., the Gibbs-Duhem equations) in the training step and ensure that the obtained predictions satisfy given constraints. An example of their application in predicting excess properties is the work by Felton et al. [45]. They embed in the neural network the Gibbs-Duhem relationship while predicting activity coefficients.

This paper introduces a novel framework to construct physics-aware hybrid models to predict the macroscopic effects of molecular interactions. The framework can also work with poorly specified systems predicting their ACs. It is flexible to all the first-principle models available in the literature, and it can also be used stand-alone to have an entirely statistical description of the interactions occurring within the systems. In addition, the framework respects the main a-priori rules governing the excess properties (i.e., Gibbs-Duhem equation and activity for pure components), still utilizing functional structures typical of the ML techniques. The physical characterization is done to 1) reduce the number of parameters that the optimizer should identify in the training phase and 2) avoid solutions not having any physical sense.

2. Framework characterization

This section proves and develops the main mathematical groundwork supporting the physical characterization of the hybrid model. The framework considers two application cases: 1) the deviation of the ACs from the first-principle model predictions is caused by one or more unknown chemicals contained in the mixture component (*internal disturbance*), and 2) the deviation of the ACs from the first-principle model predictions is caused by an external agent affecting the mixture (*external disturbance*) (Figure 1). In this paper, the disturbance is defined as a phenomenon causing the ACs of a mixture deviate from their first-principle predictions. The disturbance is internal if the mixture is partially unknown (i.e., it is impossible to determine part of the components contained in it) and the molar fraction of the known part does not sum up to 1. On the other hand, the disturbance is said to be external when the molar fraction of the known part sums up to 1, but the ACs still deviate from the first-principle model prediction. Let's consider a mixture containing two solvents (e.g., toluene and water) and an active pharmaceutical ingredient (API). If the interaction between the solvents and the API is unknown, this case can be considered to have internal disturbance. The known part of the mixture comprises the two solvents, while the API represents the unknown part. On the other hand, an example of a case with external disturbance is the case of a spectroscopic analysis (e.g., GC-MS) calibrated only for a set of molecules. In this case, the molar fraction of the various components is calculated as the ratio between the heights of the peaks; therefore, the overall concentration of the known components will sum up to 1 by definition. However, the measurement techniques cannot detect all the components within the mixture. As a last example, one of the demonstration cases reported in this work involves a mixture of organic solvents and LiCl. The same mixture is used as a system with internal or external disturbance. LiCl is one of the chemical components in the mixture; if its concentration is known, but its nature is unknown, the system can be considered with internal disturbance. On the other hand, a component could be present within the mixture, but the analytical instrument does not detect it; therefore, its measured concentration is equal to 0. If the AC values deviate from the first-principle prediction, the mixture can be considered with external disturbance. One can apply the following rule as a differentiation between cases with internal and external disturbance. A case can be considered to have internal disturbance if the concentration of one or more components is known, but their nature is unknown. On the other hand, a case can be considered to have external disturbance if the unknown components have unknown concentrations. A diagram guiding the case selection with internal or external disturbance is reported in the Supplementary Information section of this paper. In Figure 1a, the components *K* are the organic solvents while the component *U* is the LiCl. Therefore, the molar fraction of LiCl can be obtained from the concentration of the other components. It is essential to highlight here that, for the sake of demonstration, the chemical nature of LiCl is known and only one component; however, the same definition can be applied if the unknown components are more than one and their chemical nature is unknown. If the mixture deviates its activity coefficients from the first-principle prediction and the concentration of LiCl is unknown and impossible to estimate from the concentration of the known components, the system has external disturbance. In Figure 1b, the components K are the organic solvents; in this case, the sum of their molar fraction equals one. Therefore, it is impossible to estimate the concentration of unknown components (i.e., LiCl) through the concentration of the known components.

Baumeister and Burger analyzed the case of internal disturbance for systems with one known component [18]. The hybridGamma framework also predicts when the internal disturbance affects the behavior of a mixture and not only a single component. Therefore, the case reported by Baumeister and Burger is a specific case that can be covered by the framework proposed in this work. Besides the internal disturbances, the hybridGamma framework can also operate with unquantifiable external disturbances. This is possible by creating a model considering only measurable variables as input (e.g., temperature and the concentration of the known and measurable chemical components). This way, a model of the disturbance is created without using the level of the external disturbance.

The mathematical restrictions are applied and proven over a general statistical function in Section 2.1. Further, the mathematical restrictions are applied over the two disturbance cases in Section 2.2. Then, the characterization framework is applied to a polynomial statistical function. In Section 2.3, a 3rd-order polynomial statistical functional form is described and characterized for systems with internal and external disturbances. At the end of this paragraph, the steps involved within the library automatically characterizing the statistical model are introduced (Section 2.4).

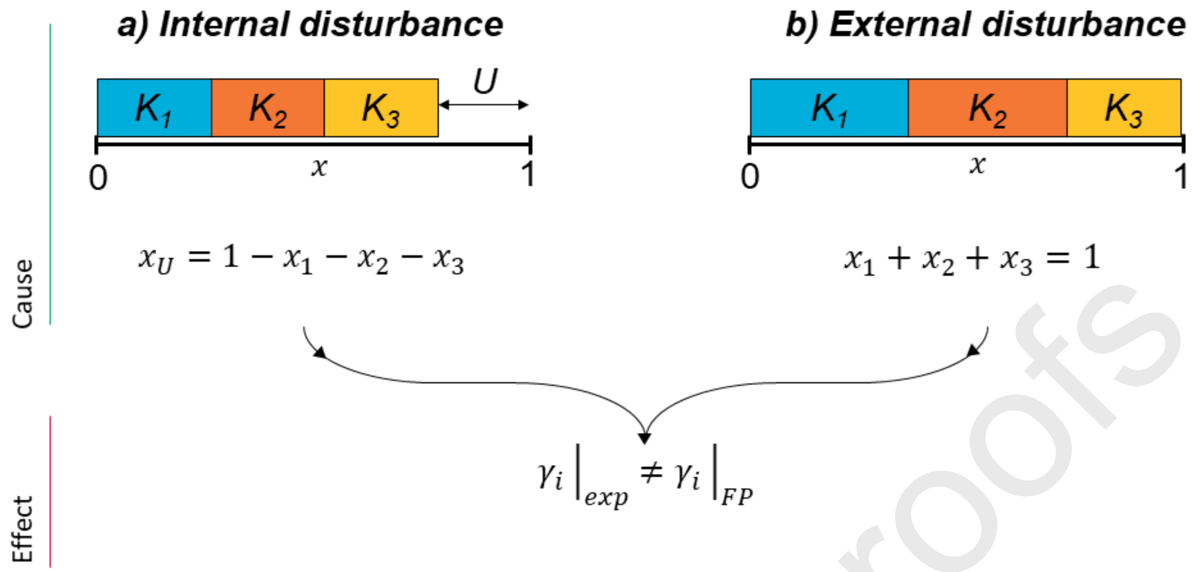

Figure 1. Graphical representation of the cases reported in this paper. In this figure, x is the molar fraction of the chemical components in the liquid [mol/mol], γ is the activity coefficient [-], K is the known component, U is the unknown component, $|_{exp}$ is the experimentally measured property, $|_{FP}$ is the property predicted by the first-principle model. In the case of internal disturbance, the molar fractions of the known components in the mixture do not sum to 1. The presence of unknown components creates a deviation from the first-principle predictions. In the case of external disturbance, the molar fractions of the known components in the mixture sum to 1. However, the experimental ACs still deviate from the first-principle predictions.

2.1. Structure and proof of the general framework

The proposed hybrid model considers the sum of the logarithms of the AC. Therefore, the hybrid AC models are represented as

$$\ln(\gamma_i)|_{HM} = \ln(\gamma_i)|_{FP} + \ln(\gamma_i)|_{SM}, \quad (1)$$

where γ are the activity coefficients. The subscripts HM, FP and SM refer to the quantities calculated using a hybrid-model, a first-principle model, and a statistical model respectively.

Because the hybrid model represents a physical system, it must satisfy the Gibbs-Duhem equation for excess thermodynamic property [46], that characterized for Gibbs excess energy at constant temperature and pressure becomes

$$\sum_{n=1}^{N_{components}} x_n d(ln(\gamma_n)|_{HM}) = 0\Bigg|_{T,P}, \quad (2)$$

where γ are the activity coefficients, x_n is the molar fraction of the n-th component within the mixture and the subscript HM refers to the quantity calculated with the hybrid model.

In addition, it must let the pure component AC be equal to one in case of internal disturbances. The application of the limit condition is not possible for systems with external disturbance because the deviation still applies to a pure component. It could also interact among the molecules of such a component, causing a deviation from the first-principle predictions. These conditions are mathematically formulated as

$$\lim_{x_i \to 1} \gamma_i|_{HM} = 1 \Rightarrow \lim_{x_i \to 1} ln\,(\gamma_i)|_{HM} = 0, \quad (3)$$

where γ are the activity coefficients and the subscript HM refers to the quantity calculated with the hybrid model.

For the case of internal disturbance, equation (3) is appliable only for the fully known chemical molecules (i.e., both the nature of the molecule and the interactions with all the other known molecules within the system are known).

From the characterization of equation (3) on a binary system, it is possible to write

$$x_i \frac{d(ln\,(\gamma_i))}{dx_i} + x_k \frac{d(ln\,(\gamma_k))}{dx_i} = 0, \quad (4)$$

where γ are the activity coefficients, x is the molar fraction of the i-th and k-th component within the liquid. The independency of (4) from the derivation variable x_i is proven in the Supplementary Information of this paper in section 1s. As previously stated, the hybrid model combines a first-principle and a statistical model. It is assumed, in this work, that the first-principle model already satisfies the restrictions imposed by equations (3) and (4). For example, this is already the case if one of the local composition models (e.g., UNIFAC, NRTL, PC SAFT) is utilized as the first-principle part of the hybrid model. Section 2.1.1 proves the separability of the Gibbs-Duhem equation over the first-principle and statistical part of the hybrid model. The separability of the limit condition (3) over the first-principle and statistical part of the hybrid model is trivial and not reported in this paper.

2.1.1. Gibbs-Duhem equation separability over the hybrid model structure

Applying the Gibbs-Duhem equations over the hybrid model structure (1) results in

$$x_i * \frac{d\left(ln\,(\gamma_i)|_{FP} + ln\,(\gamma_i)|_{SM}\right)}{dx_i} + x_k * \frac{d\left(ln\,(\gamma_k)|_{FP} + ln\,(\gamma_k)|_{SM}\right)}{dx_i} = 0, \quad (5)$$

using the sum property of the differentiation, it can be written as

$$x_i * \frac{d(ln\,(\gamma_i)|_{FP})}{dx_i} + x_k * \frac{d(ln\,(\gamma_k)|_{FP})}{dx_i} + x_i * \frac{d(ln\,(\gamma_i)|_{SM})}{dx_i} + x_k * \frac{d(ln\,(\gamma_k)|_{SM})}{dx_i} = 0. \quad (6)$$

The expression in (6) implies that the Gibbs-Duhem equation is separable over the selected hybrid model structure. In addition, considering the hypothesis of the validity of the Gibbs-Duhem restriction on the first-principle model, the first two summations of the left-hand side in (6) are null. Therefore, it is possible to simplify (6) as

$$x_i * \frac{d(ln\,(\gamma_i)|_{SM})}{dx_i} + x_k * \frac{d(ln\,(\gamma_k)|_{SM})}{dx_i} = 0. \quad (7)$$

2.1.2. *A corollary of the hybrid model characterization*

Characterizing the hybrid model to respect the Gibbs-Duhem equation (4) and the limit condition (3) implies that the statistical function must respect the conditions reported in Section 2.1. Therefore, the same restrictions can also be applied to a purely statistical model of the ACs. In other words, the framework proposed in this paper and executed for a hybrid model applies to an entirely statistical modeling approach of the ACs without any modification.

2.2. Characterization of the framework for internal and external disturbance

2.2.1. *The case of the internal disturbance*

Let us consider a solution containing two classes of components. The first group of components is fully characterized in concentration and chemical nature. The interactions among the molecules within this group are fully described utilizing only first-principle models. The second group of components is not characterizable in concentration, chemical nature, or both. Therefore, it is impossible to describe this part of the mixture, represent the interaction among the molecules within this group using a first-principle model, or both. Let us say *M* is the total number of components within the mixture, *K* is the number of components fully characterized, and *U* is the number of components impossible to characterize. Let us say x_1,x_2,..., x_k the measurable molar fractions of the *K* known components. For the remaining part of the mixture, it is impossible to measure the molar fractions of the unknown components individually; however, it is possible to estimate the sum of the molar fractions of the unknown components using (8). Here, even though the molar fractions of the unknown *U* components are still unmeasurable, it is still possible to estimate their sum

$$x_U = \sum_{j=1}^{U} \overline{x_j} = 1 - \sum_{j=1}^{K} x_j, \quad (8)$$

where x_U is the sum of the liquid molar fraction of the unknown components within the mixture $\overline{x_j}$, and x_j is the liquid molar fraction of the known components within the mixture. A system respecting the above-illustrated rules is defined as a *system with internal disturbance*. In such a system, it is possible to utilize first-principle models to describe the interactions between the molecules only for the *K* known components. On the contrary, describing the interactions involving any unknown component is impossible by employing the first-principle equation. In addition, the interactions between the known components are not encoded in the statistical model because they are already included in the first-principle model. Therefore, the statistical part of a hybrid model describing a system with internal disturbance only considers the concentration of the unknown part, as expressed in (8), and its interaction with the known concentrations and temperature.

A system with an internal disturbance requires a modification over the molar fraction definition in the first-principle model. The modification is needed to let the first-principle part of the hybrid model (1) respect the Gibbs-Duhem equation and the limit condition for the pure components. For this reason, the first-principle part of the hybrid model (1) should be calculated utilizing a normalized molar fraction over the known components as described by

$$x_j^* = \frac{x_j}{\sum_{i=1}^{K} x_i} \text{ for } j \text{ in } K. \quad (9)$$

Therefore, considering the abovementioned model simplification strategies and equation (9), the hybrid model (1) for the system is

$$\ln(\gamma_i)|_{HM} = \ln(\gamma_i)|_{FP} + \ln(\gamma_i)|_{SM} = \ln\left(\gamma_i\left(x_1^*, x_2^*, \ldots x_k^*, T\right)\right)|_{FP} + \ln\left(\gamma_i(x_U, x_1 x_U, x_2 x_U, \ldots, x_k x_U, x_U T)\right)|_{SM}. \quad (10)$$

In the case of internal disturbance, the unknown components contained in *U* are part of the chemical mixture; therefore, it is possible to draw the limit condition. This statement will not be possible for the case with external disturbance; the reader is invited to read Section 2.2.2 of this manuscript for further information. In the case of internal disturbance, the unknown part of the mixture generally does not behave as a pure component at the limit condition unless it is composed of only one chemical. Consequently, it is possible to derive the following equation for the known components:

$$\lim_{x_i \to 1} \ln(\gamma_i)|_{HM} = 0 \text{ for } i \text{ in } K. \quad (11)$$

One could argue that the case with internal disturbance can be solved by utilizing the pseudo-component hypothesis, applying a first-principal model, and identifying unknown parameters over experimental data. However, this approach has two limitations: 1) it is implicitly assuming that the first-principle model utilized for the known mixture also works for the unknown part, 2) it is implicitly assuming that the unknown component is pure since the limit condition is also respected for the pseudo-component because of the structure of the first-principle model. Therefore, using the pseudo-component hypothesis could drastically reduce the extrapolation capabilities of the obtained model and force the prediction to be valid only within the range of the experimental data.

2.2.2. The case of the external disturbance

In a solution having an external disturbance, the components within the mixture are fully characterized, and the sum of their concentrations is 1. The solution contains *K* components, and it is possible to know the concentration and chemical nature of all the *K* components. Let us say $x_1, x_2, \ldots, x_K$ the molar fraction of the *K* fully known components. Despite the system being fully characterized from a chemical point of view, the solution still shows a deviation in the ACs since an external agent is acting on it. For example, this could be the case of a perfume interacting with a fabric fiber. In this case, the deviation is not caused by a component within the mixture, but by an external agent acting on the mixture. Therefore, the case with an external disturbance considers a deviation of the ACs caused not by species contained in the solution but by external factors (e.g., interaction of the liquid with a surface or a particle dispersion within the liquid). Because of external disturbance, all the chemical interactions deviate from the ones predicted by first-principles. Thus, the statistical model describing the deviation of the ACs considers only the concentration of known components. Therefore, the hybrid model is constructed as

$$\ln(\gamma_i)|_{HM} = \ln(\gamma_i)|_{FP} + \ln(\gamma_i)|_{SM} = \ln(\gamma_i(x_1, x_2, \ldots x_K, T))|_{FP} + \ln\left(\gamma_i(x_1, x_2, \ldots x_K, T)\right)|_{SM}. \quad (12)$$

Since the external disturbance also affects the pure component activity applying the limit conditions to the hybrid model equations is impossible. In addition, in case the level of the external disturbance is quantifiable, it is possible to include this information within the coefficients of the statistical model. In this case, the utilized function represents the model coefficient as zero in case of a null disturbance level. Although the reported case is a further development in this framework, this work illustrates only the case with an unknown level of external disturbance. In addition, the level of external disturbance is supposed to be constant.

2.3. Proof of the 3rd-order polynomial statistical function

This section illustrates how characterizing the statistical part of the hybrid model utilizing physical restriction has two main benefits: 1) adding physical information into the system and increasing the model generalization capabilities and 2) reducing the number of the free parameter that the optimizer must identify during the training. So far, the focus of the paper was about increasing the physical awareness of the model; however, the second point is also crucial since it allows the reduction of the required training data points.

To show how the statistical models are constrained via the proposed framework, a third-order polynomial of a ternary mixture is considered

$$\ln(\boldsymbol{\gamma})|_{SM} = \boldsymbol{A} \cdot \boldsymbol{x} + \boldsymbol{B} \cdot \boldsymbol{x}^2 + \boldsymbol{C} \cdot \boldsymbol{x}^3, \quad (13)$$

$\ln(\boldsymbol{\gamma})|_{SM}$ is a 3x1 vector containing the statistically modelled values of the ACs. $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$ are matrixes containing the parameters tunable by the search algorithm; their shape is 3x4, 3x10 and 3x20 respectively. Therefore, the overall system has 102 free parameters to identify prior to the physical characterization. The vector $\boldsymbol{x} = (x_1, x_2, x_3, T)^T$ contains all model inputs, here x_i is the concentration of the i-th chemical component contained in the mixture, and T is the temperature of the mixture. The vector $\boldsymbol{x}^2 = (x_1^2,\ x_1x_2,\ x_1x_3, x_1T, x_2^2, x_2x_3, x_2T, x_3^2, x_3T, T^2)^T$ contains all the 2nd-order interactions of the variables contained in the vector $\boldsymbol{x}$. The vector $\boldsymbol{x}^3 = (x_1^3, x_1^2x_2, x_1^2x_3, x_1^2T, x_1x_2^2, x_1x_2x_3, x_1x_2T, x_1x_3^2, x_1x_3T, x_1T^2, x_2^3, x_2^2x_3, x_2^2T, x_2x_3^2, x_2x_3T, x_2T^2, x_3^3, x_3^2T, x_3T^2, T^3)^T$ contains all the 3rd-order interactions of the variables contained in the vector $\boldsymbol{x}$. In the final hybrid model, equation (13) is combined with the first-principle model describing the interaction between the known part of the system as reported in equation (1).

2.3.1. Application of the 3rd-order polynomial over a system with internal disturbance

As described in section 2.2.1, the system can be characterized by equation (10). Both physical constraints can be applied, namely the Gibbs-Duhem equation (7) and the limit condition (3). In this example, a three components system is analyzed. Component 3 is the unknown part of the system. Therefore, it is possible to describe the interaction between components 1 and 2 using first-principle models, while the statistical model describes the interactions between 1-3 and 2-3. Consequently, the model becomes

$$\begin{cases} \ln(\gamma_1)|_{SM} = A(1,3)*x_3 + B(1,3)*x_1x_3 + B(1,6)*x_2x_3 + B(1,8)*x_3^2 + \\ \quad +B(1,9)*x_3T + C(1,3)*x_1^2x_3 + C(1,6)*x_1x_2x_3 + C(1,8)*x_1x_3^2 + \\ \quad +C(1,9)*x_1x_3T + C(1,12)*x_2^2x_3 + C(1,14)*x_2x_3^2 + C(1,15)*x_2x_3T + \\ \quad C(1,17)*x_3^3 + C(1,18)*x_3^2T + C(1,19)*x_3T^2 \\ \ln(\gamma_2)|_{SM} = A(2,3)*x_3 + B(2,3)*x_1x_3 + B(2,6)*x_2x_3 + B(2,8)*x_3^2 + \\ \quad B(2,9)*x_3T + C(2,3)*x_1^2x_3 + C(2,6)*x_1x_2x_3 + C(2,8)*x_1x_3^2 + \\ \quad C(2,9)*x_1x_3T + C(2,12)*x_2^2x_3 + C(2,14)*x_2x_3^2 + C(2,15)*x_2x_3T + \\ \quad C(2,17)*x_3^3 + C(2,18)*x_3^2T + C(2,19)*x_3T^2 \\ \ln(\gamma_3)|_{SM} = A(3,3)*x_3 + B(3,3)*x_1x_3 + B(3,6)*x_2x_3 + B(3,8)*x_3^2 + \\ \quad B(3,9)*x_3T + C(3,3)*x_1^2x_3 + C(3,6)*x_1x_2x_3 + C(3,8)*x_1x_3^2 + \\ \quad C(3,9)*x_1x_3T + C(3,12)*x_2^2x_3 + C(3,14)*x_2x_3^2 + C(3,15)*x_2x_3T + \\ \quad C(3,17)*x_3^3 + C(3,18)*x_3^2T + C(3,19)*x_3T^2 \end{cases} \quad (14)$$

Applying the Gibbs-Duhem equation and the limit condition constraints to the reported case, the physical characterization of the statistical model, reduces the number of free parameters identifiable by the optimizer from 102 to 28.

2.3.2. *Application of the 3rd-order polynomial over a system with external disturbance*

As described in section 2.2.2, the system can be characterized by only equation (12) and the Gibbs-Duhem restriction (7). The nature of the system blocks the possibility of applying the limit condition restriction. In addition, the statistical model requires as input the concentration of the components described already in the first-principle model and the temperature (i.e., x_1, x_2 and T). In the reported case, the physical characterization of the statistical model has reduced the number of free parameters identifiable by the optimizer from 102 to 86.

2.4. Description of the characterization algorithm

2.4.1. *Description of the algorithm*

The framework developed for characterizing the Gibbs-Duhem equation and limiting conditions over polynomial statistical functions is deployed in a python library available on request mailing the authors. This section describes the most relevant steps performed by the algorithm. The algorithm initiation requires a few definition parameters. The most relevant definition parameters are 1) the number of components within the system (i.e., *K+1* in the case of internal disturbance or *K* in case of external disturbance), 2) the order of the polynomial function to implement as the statistical model, and 3) the presence of an unknown component. The resolution of the equations applying the physical restriction within the algorithm is entirely symbolic. However, at the end of the resolution, the resulting characterized equations describing the ACs are deployed within a python function more responsive than a symbolic relation. The algorithm developed for this work is fully developed in python, employing SymPy 1.10.1 [47] for the symbolic manipulation and resolution of the equations. The procedure described in the following steps is deployed in a python class. The various parts of the algorithm are deployed in the class methods.

When executed, the procedure creates a symbolic vector having many elements as the number of components within the mixture. If the temperature is added to the input space, the variables vector also contains a symbolic element for the temperature. The description performed in this section hypothesizes that the temperature is selected as the input element. Therefore, the dimension of the vector is *M+1*. For example, looking at (13), this procedure generates the symbolic representation of $\boldsymbol{x}$. After creating such an array, the higher-order monomial terms are symbolically computed. They are generated by multiplying each term of the variable vector both by itself and all the other terms. The resulting terms are stored in another array containing all the monomial terms of the set order. The procedure is repeated until the monomial order set during the class definition is reached. In equation (13), this procedure generates the symbolic representation of $\boldsymbol{x^2}$ and $\boldsymbol{x^3}$. Once all the monomial terms are generated, the procedure computes the matrices containing the model parameters. Each matrix is associated with a monomial order and contains symbolic elements. It has as many rows as the number of components within the mixture and as many columns as the number of monomial terms associated with the monomial order. In equation (13), this procedure generates the symbolic representation of $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$. The statistical model is generated by summing the dot products between the vector and the matrix associated to the same monomial order. After creating the statistical functions, they are characterized to the case in the analysis. More specifically, the parameters of the variables to exclude from the computation are set to 0. For example, in the case of the presence of an unknown component, the statistical model considers only the interactive and non-interactive terms containing the unknown species. Therefore, all the parameters associated with the known species and their interactions are set to 0. This procedure creates the symbolic equations of the statistical model. The symbolic representation is then utilized to compute the equations to physical-restrict the statistical models. The first step is calculating the limit conditions if needed. It is done by iteratively setting to 1 the molar fraction of one specie and 0 the molar fractions of all the other chemical species in the symbolic equations of the statistical models. For each component, this procedure returns an equation containing the model parameters. Some of the model parameters are multiplied by the temperature. These equations are stored in a python list and will be solved later in the procedure. After characterizing the limit conditions, the restrictions for the binary interaction of the Gibbs-Duhem equation are computed. It is performed by iteratively considering binary mixtures within the symbolic equations of the statistical model and implementing equation (7). It creates a polynomial function in the concentration of one of the two species. This relation must be equal to 0 for any temperature and molar fraction value, as expressed in (7); the only way to achieve this is to set all the coefficients of the equation to 0. Thus, the procedure sets all the polynomial coefficients to 0. This operation creates an equation set containing the model parameters and the system temperature. The equations are stored in the equations list to be solved. Later, the equations obtained with the Gibbs-Duhem correlation and the limit condition are united. This procedure generates an equations system to solve in order to constrain the statistical function. This equation set is linear in the model parameters. The only non-linear term is the product between the temperature and the model parameters; however, this does not restrict the resolution since the validity of the Gibbs-Duhem equation is at a constant temperature. The same can be claimed for the limit condition. Because of the linearity of the problem, the system admits only one solution that can be obtained analytically. The system is then solved in a symbolic manner returning correlations between variables. The obtained restrictions are implemented on the variable matrices, and the restricted equations are generated in a symbolic form.

2.4.2. The computational complexity of the algorithm

An analysis of the computational time required for the characterization of the system was performed. The investigation considered a system with external disturbance containing interactive terms between the variables. In addition, the limit condition was added to the system to cover the worst-case scenario, even if it does not make physical sense. The polynomial order and the number of components were considered for the analysis. The polynomial order varied from 1 to 5, while the number of components varied from 2 to 5. A full factorial analysis was performed involving 20 data points. The analysis returned an exponential complexity on the polynomial order and the number of components. Therefore, using the O-notation, the system has a model complexity of $O(e^{NC * PO})$ where NC is the number of components within the mixture and PO is the polynomial order utilized in the model. Characterizing the function is the slowest step of the algorithm, and the abovementioned model complexity is related only to this part. Once the function has been characterized, it is converted to a python lambda function, and the computational time is negligible. Further information about the algorithm complexity and computational time are reported in the Supplementary Information of this paper in section s2. The computational complexity of the algorithm reported in this work is very unscalable since it grows exponentially with the number of components and the polynomial order. In this work, we focus on the methodology to obtain the hybrid model rather than the scalability of the algorithm. Further research can be executed to increase the scalability of the characterization algorithm by reducing the computational complexity.

3. Application of the framework and results

This section reports and analyses the application of the hybridGamma framework over experimental cases obtained from the literature and simulation. The investigated cases involved vapor-liquid equilibria (VLE) systems with 1) two organic components and one electrolyte, obtained from literature data, and 2) a multi-components organic mixture obtained from simulations. The literature data were obtained by Iliuta et al. [48]. They reported the VLE data for a system involving acetone and methanol as molecular liquids and LiCl as the dissolved salt. The paper refers to this dataset as the *"Iliuta dataset"*. The simulation data were obtained by calculating the VLE profiles of mixtures, including tetrahydrofuran, cyclohexane, acetonitrile, and benzene, obtained using Aspen Plus V10 using the NRTL model [10]. This paper refers to this dataset as the *"Aspen dataset"*. Most of the mixture in the Aspen dataset was composed of tetrahydrofuran and cyclohexane, with a total molar composition above 80%. On the other hand, the molar fraction of acetonitrile and benzene is lower than 20%. Therefore, the sub-mixture acetonitrile-benzene is considered the disturbance when operating with the Aspen dataset. Further information about the simulations executed to obtain the Aspen dataset is reported in the Supplementary Information section of this paper, in paragraph s3.

Both datasets were utilized to test the framework for the cases with external disturbance (in Section 3.1) and internal disturbance (in Section 3.2). Only the known components were considered for the external disturbance, and their mass fraction was normalized to 1. Although both experimental cases were employed to validate the framework, the approaches for the parameter identification of the statistical model differ between the two cases. For the Iliuta dataset, the hybrid models were trained using information about the VLE data. Therefore, the models were not trained directly on the ACs values; however, the cost functions aim to minimize the error between the boiling temperature of the system and the vapor concentration. For this reason, the framework validation includes Raoult's law for non-ideal mixtures

$$P_{tot} y_i = P_i^0(T) \gamma_i|_{HM}(\boldsymbol{x}, T) x_i, \quad (15)$$

where the ACs $\gamma_i|_{HM}$ are calculated using the hybrid model (1), P_{tot} is the total system pressure, y_i is the vapor molar-fraction of the i-th component within the mixture, $P_i^0(T)$ is the vapor pressure of the i-th component and x_i is the liquid molar fraction of the i-th component within the mixture. A schematic representation of the training loop employed for the Iliuta dataset is given in Figure 2a. For the Aspen dataset, the hybrid models were trained using the ACs experimental values of the known components directly in the cost function (Figure 2b). The two training paradigms were designed to evaluate the performances of the framework both when direct information about the ACs is available and when this is missed but are available other measurable variables affected by the ACs (e.g., the boiling temperature for the evaluation case reported in this work). The first-principle model in the hybrid model is the non-random two liquids (NRTL) [10] for all the cases.

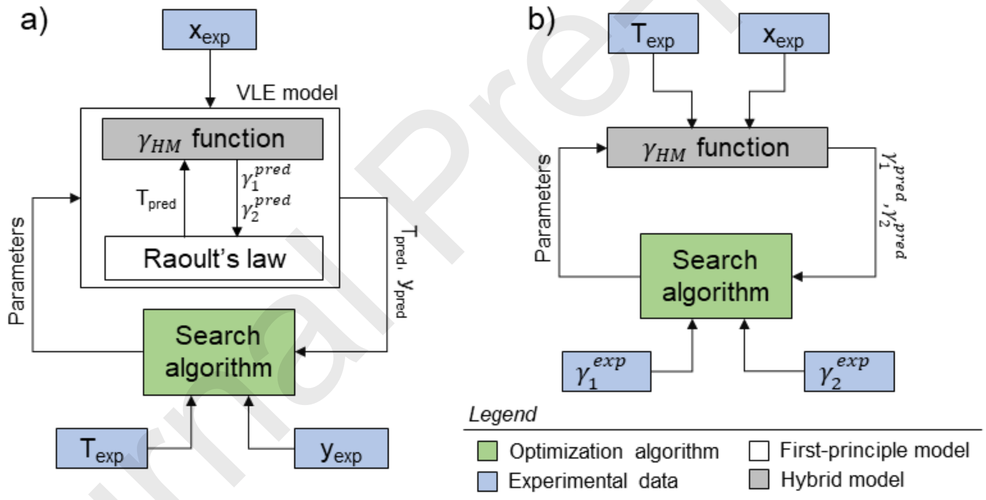

Figure 2: Schematic representation of the training approach for the two datasets. In this figure, x is the molar fraction of the chemical components in the liquid, γ is the activity coefficient, T is the temperature of the system [K], y is the molar fraction of the chemical components in the vapor [mol/mol], the superscript and the subscript exp is the experimentally measured property, the superscript and the subscript pred is the property predicted by the first-principle model, $|_{HM}$ is the property predicted by the hybrid model a) Training loop used for the Iliuta dataset. Here, the ACs are combined with Raoult's law for non-ideal mixtures, and the parameter identification is executed on the temperature and vapor concentration profile predicted by it. b) Training loop used for the Aspen dataset. Here, during the training, the predicted values of the ACs are compared with the experimental ones to compute the loss function.

To train the model, as many data points as the number of model parameters were selected. All the other experimental data points serve as the validation set for an a-posteriori evaluation of the model and to assess its generalization capabilities. The generalization capabilities are evaluated utilizing the root mean squared error (RMSE), the mean absolute percentage error (MAPE) and the coefficient of determination (R^2).

Further validation cases can be found in the Supplementary Information section of this paper, paragraph 6s.

3.1. Validation of the framework on a system with external disturbance

The statistical function utilized for both the systems with external disturbance employed a second-order polynomial, including the system temperature in the input space. To further reduce the number of parameters included in the system, the statistical function of the AC does not consider the interaction terms between the input variable. Because of the nature of the system, the limit condition was not applied (for further information, see section 2.2.2). The resulting system after the physical restriction with the Gibbs-Duhem equation is

$$\begin{cases} ln(\gamma_1)|_{SM} = A(1,1)*x_1 + A(1,2)*x_2 + A(1,3)*T + B(1,1)*x_1^2 + B(1,2)*x_2^2 + B(1,3)*T^2 \\ ln(\gamma_2)|_{SM} = A(2,1)*x_1 + A(2,2)*x_2 + A(2,3)*T + B(2,1)*x_1^2 + B(2,2)*x_2^2 + B(2,3)*T^2 \end{cases} \quad (16)$$

where $A(i,j)$ and $B(i,j)$ are the parameters that are identified during the training of the statistical model, x_1 and x_2 are the molar fraction of the known components, T is the temperature of the system. After the characterization using the hybridGamma framework, the model (16) has 9 parameters identifiable via the search algorithm. In this work, SciPy [49] 1.7.3 was used to fit the model, namely the algorithm Broyden–Fletcher–Goldfarb–Shanno implemented in the *optimize.minimize* function of SciPy was used. The minimization was executed on the mean absolute error of the predictions.

3.1.1. Framework validation with external disturbance on the Iliuta dataset

The VLE data reported by Iliuta et al. [48] with a molar fraction of LiCl of 0.1 were utilized for the validation executed in this section. In equation (16), component 1 is the acetone, and component 2 is the methanol. LiCl is the external disturbance of the system; the hybrid model is unaware of the LiCl concentration. As many data points as the number of model parameters were selected for the training set. Figure 3a reports the performances of the trained hybrid model on the VLE system acetone-methanol with the external disturbance. Moreover, these plots include the prediction obtained with the first-principle model considering only the known part of the mixture and the data points utilized for the training and test of the model. Figure 3a reports the excellent generalization and extrapolation capabilities of the hybrid model. The result is achieved using only 9 data points in the train set. The model matches almost perfectly the test set points. In addition, the model shows excellent extrapolation capabilities toward parts of the investigation space not explored during the training set (i.e., the points located at the minimum and the maximum of the input space). This behavior is related to the physical awareness of the statistical model achieved via the characterization of the functions with the Gibbs-Duhem restrictions. The prediction accuracies over the test set are also confirmed by the evaluation metrics reported in Table 1. In this table, the model returns high R^2 scores on the test set and very low RMSE and MAPE, highlighting the outstanding capabilities of the obtained model.

Table 1. Evaluation metrics for the Iliuta case with the external disturbance over the ACs on the test set.

	RMSE	MAPE	R^2
Temperature	0.296	0.403%	0.992
Vapor concentration	0.0415	5.77%	0.972

$\gamma_{Acetone}$	0.128	6.10%	0.855
$\gamma_{Methanol}$	0.0856	12.3%	0.279

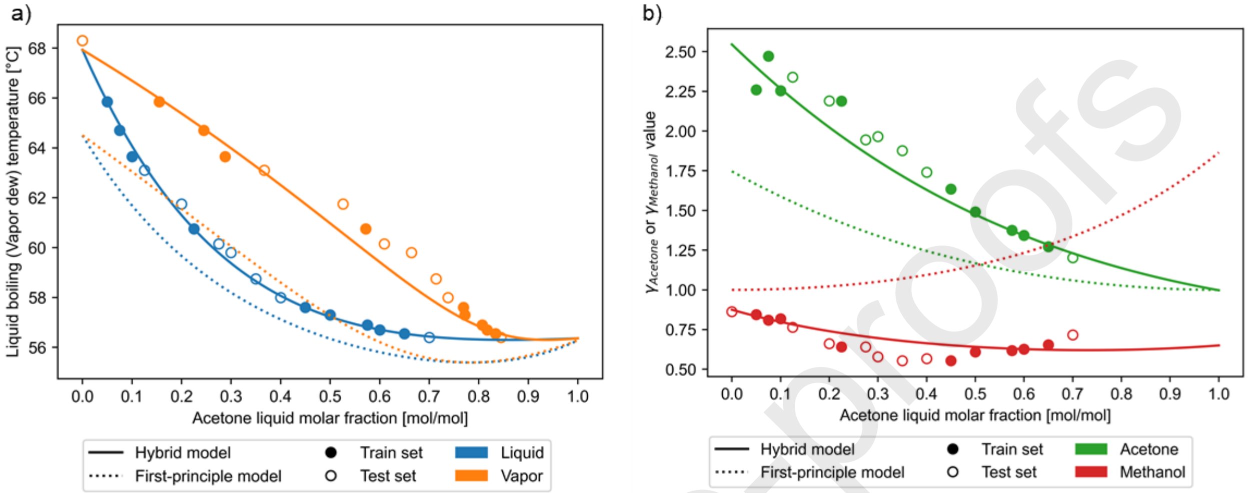

Figure 3: Performance of the hybrid model framework to predict Acetone-Methanol-LiCl system VLE having as the external disturbance the LiCl concentration. a) Profile of the boiling point varying the acetone concentration in the liquid phase. b) profile of the acetone content in the vapor varying the acetone concentration in the liquid phase. The obtained ACs model shows significant generalization and extrapolation capabilities.

Figure 3b reports the experimental and predicted profiles of the ACs. The ACs reports in this plot were calculated using the information reported in the Iliuta dataset and Raoult's law for non-ideal mixtures (15) using the training loop in Figure 2a. In Figure 3b, it is possible to assess the model prediction quality within the entire range, both on training and test sets. The hybrid model profile matches the entire dataset for the acetone both on the train and on the test data points. Moreover, they return excellent scores when the models are evaluated over the test set (Table 1). The prediction over the methanol ACs profile is less accurate than the case of acetone. The profile reported in Figure 3b related to the methanol ACs presents a significant deviation in the range of acetone concentration from 0.2 to 0.5. In this area, the model systematically overestimates the ACs values. However, the hybrid model predictions of the methanol ACs are better in the rest of the experimental space, both on the train and test sets. For this reason, the evaluation metrics on the test set about the methanol ACs are lower than one of the acetone. Most of the points included in the test set for the methanol ACs were from 0.2 to 0.5, with few points outside this range. This worsens the score values. Despite the bias, the evaluation performed over the test set for the methanol ACs returns acceptable RMSE and MAPE values, with values lower than 0.5 and 10%, respectively; however, the R^2 results are lower than the acetone ACs predictions (Table 1). In addition, from Figure 3, it is possible to assess the correction level made by the hybrid model over the first-principle model predictions. Figure 3b shows the ACs curves predicted by the first-principle model to intersect at $x_{Ac} = 0.5$; however, this is not the case for the experimental data. The hybrid model performed this correction only using the system output and without any information about the actual value of the ACs. From Figure 3 and Table 1, it is possible to conclude that the framework shows excellent parameter identification and generalization capabilities for a system with external disturbance when only the output variables are used for the training. In this case, the MAPE on the estimated activity coefficient is around 10%, and the RMSE is lower than 0.2. All of this makes the developed hybrid model prediction overlap the experimental points in Figure 3b.

3.1.2. Framework validation with external disturbance on the Aspen dataset

Table 2. Evaluation metrics for the Aspen dataset with the external cause of disturbance over the ACs on the test set.

	RMSE	MAPE	R^2
$\gamma_{Tetrahydrofuran}$	1.46e-3	0.075%	0.999
$\gamma_{Cyclohexane}$	6.65e-3	0.351%	0.999

For this evaluation case, were selected the data points having the molar fraction of the unknown system of 0.2. This dataset was selected because it has the highest value of disturbance and, therefore, it gives the highest effect of the internal deviation. The training points were randomly selected within the experimental dataset, excluding the two points at the higher and lower molar fraction of tetrahydrofuran employed in the test set. As many data points as the number of model parameters were selected for the training set. Referring to equation (16), component 1 is the tetrahydrofuran, and component 2 is the cyclohexane. The training was executed using a loss function employing the predicted and experimental ACs values rather than the system output, as in section 3.1.1. Figure 4 reports the simulations obtained with the hybrid ACs functions identified during the training phase. More specifically, it reports the experimental profile value, hybrid model predictions and first-principle predictions of the tetrahydrofuran (reported in green) and cyclohexane (reported in red). From Figure 4, it is possible to evaluate the model performances within the training boundaries and its extrapolation capabilities. The hybrid model correctly predicted the entire experimental curve on both the training points and the test points. In addition, the training points covered the area between $x_{THF} = 0.2$ and $x_{THF} = 0.9$. The model did not experience any point outside these boundaries during the training phase; however, it can still accurately predict the system behavior for points outside the training set. It highlights the excellent extrapolation capabilities and generalization performance offered by the framework. They are also proved by evaluating the metrics reported in Framework validation with external disturbance on the Aspen dataset

Table 2. In addition, based on Figure 4, it is possible to evaluate how the statistical model prediction corrects the first-principle model predictions. The first-principle model was executed, considering only the mixture containing tetrahydrofuran and cyclohexane. The first-principle model predicts an intersection between the profiles around $x_{THF} = 0.5$; however, in the experimental data, the intersection between the lines happens around 0.12 because of the external disturbance. The hybrid model detects the deviation given by the external disturbance and corrects it based on the concentration of the known components.

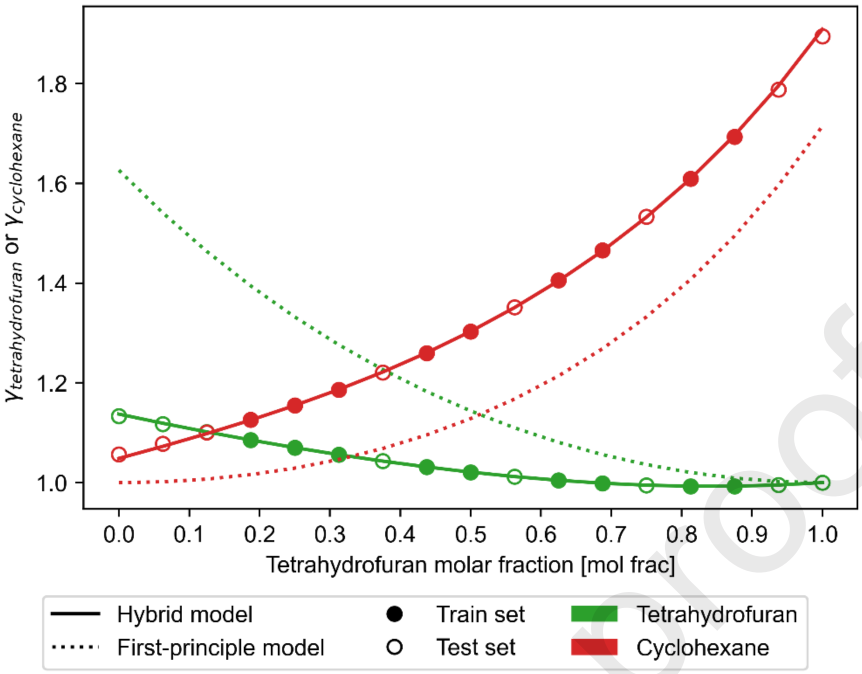

Figure 4. Performances of the hybrid model framework predict the tetrahydrofuran-cyclohexane system ACs having the acetonitrile-benzene presence as an external disturbance. The plot reports the prediction of the ACs obtained with the hybrid model and the first-principle predictions. The hybrid model describing the ACs shows significant generalization capabilities with MAPE lower than 0.5% on the test points.

3.2. Validation of the framework with internal disturbance

The statistical function used in this case employed a third-order polynomial, including the system temperature within the input space. Because of the nature of the system, the limit condition is applied (for further information, see section 2.2.1 of this paper). For brevity, the obtained equation after the physical characterization is given in the Supplementary Information of this paper in Section s4. Overall, the characterized function has 28 parameters to be identified by the optimizer.

3.2.1. Framework validation with internal disturbance on the Iliuta dataset

The entire dataset of the VLE data reported by Iliuta et al. [48], referred to Methanol-Acetone-LiCl, was utilized for this validation. The model was trained on 28 points randomly selected from the dataset; the number of the training points were chosen to be equal to the number of model tunable parameters. The points for the training set have been randomly chosen within the entire dataset except for the data having the concentration of $x_U = x_{LiCl} = 0.10$; these points were excluded from the training to evaluate the interpolation capabilities of the model over a new concentration of unknown component. Acetone and methanol were assumed to be the known chemicals, and LiCl was the internal disturbance of the system (i.e., the unknown component). Therefore, the molar fraction of LiCl was utilized as the molar fraction of the unknown component. The model training was executed as reported in section 3 of this paper and Figure 2a, considering the output variables (i.e., temperature and vapor concentration) in the loss function. The prediction capabilities of the model on the test set with $x_U = 0.10$ are reported in Figure 5. From this plot, it is possible to assess how the prediction capabilities of the first-principle model are enhanced by using the hybrid modeling framework. Although the experimental points reported in Figure 5 were not included in the training set, the model still achieves better prediction accuracies than the first-principle model employed alone. Figure 5a reports the prediction of the VLE obtained with the hybrid model trained. In this figure, the predictions of the hybrid model match the experimental points for low acetone concentration in the area $x_{Ac} < 0.3$, after this threshold, the hybrid model starts deviating, and the deviation increases with the acetone concentration within the system. However, the maximum absolute deviation detected is 1.5°C. Figure 5b reports the value of the ACs obtained experimentally from the data, as reported in section 3.1.1, and the values obtained by the first-principle and hybrid models. The ACs for the acetone and methanol are well predicted without any bias. The experimental ACs points are spread around the hybrid model prediction line. The model was not trained for the data points reported in Figure 5; therefore, it shows excellent generalization capabilities. The excellent model generalization capabilities are also confirmed by the metrics reported in Table 3. Here, two kinds of evaluations are carried out: the evaluation over the entire test set and over the data points characterized by $x_U = 0.1$. In both cases, the model results show significant accuracy and generalization with MAPE generally lower than 10%; however, the metrics evaluated over the overall test set results are better than those evaluated over the data with $x_U = 0.1$. In other words, the model has better generalization capabilities over the known chemicals rather than the unknown. This behavior is related to the structure of the hybrid model since, for the known components, the model can already rely on the information contained in the first-principle model.

Table 3. Performance metrics of the hybrid model over the test set of the system containing acetone-methanol-LiCl. The system was hypothesized to have an internal disturbance represented by LiCl

		RMSE	MAPE	R^2
Overall test set	Temperature	0.33	0.44%	0.990
	Vapor concentration	0.015	2.43%	0.991
	$\gamma_{Acetone}$	0.124	2.60%	0.948
	$\gamma_{Methanol}$	0.068	4.59%	0.913
Test set with $x_U = 0.1$	Temperature	0.352	0.52%	0.990
	Vapor concentration	0.024	3.66%	0.991
	$\gamma_{Acetone}$	0.0961	3.75%	0.954
	$\gamma_{Methanol}$	0.0665	7.60%	0.651

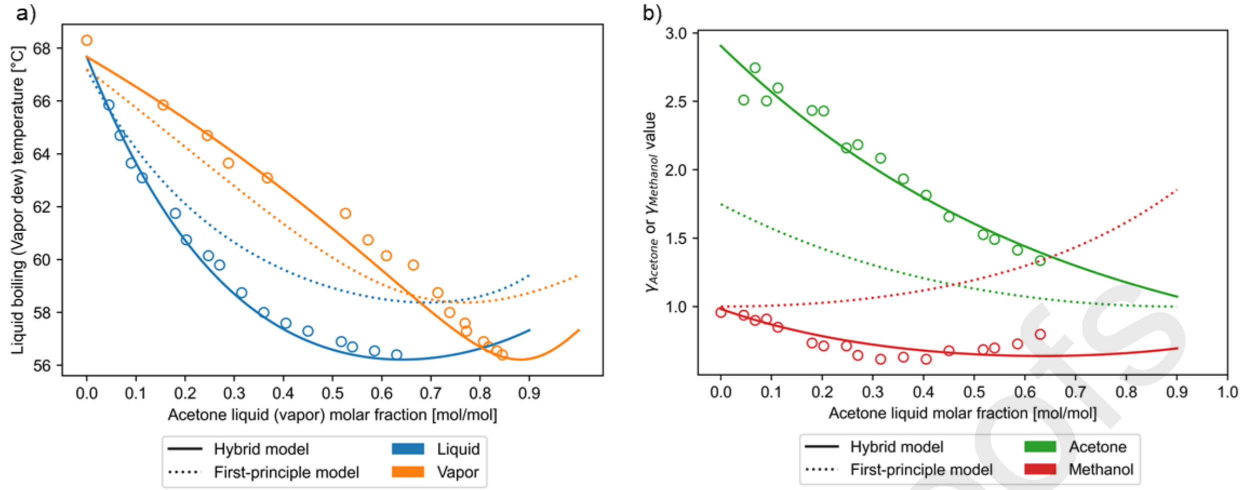

Figure 5: Performance of the hybrid model over the test set over the system containing acetone-methanol-LiCl. The system was hypothesized to have an internal disturbance represented by LiCl. a) VLE profile of the mixture. b) ACs profiles estimated using the state variables of the system. From these two plots, it is possible to assess how the statistical model enhances the prediction capabilities of the first-principle component of the hybrid model.

3.2.2. Framework validation with internal disturbance on the Aspen dataset

The entire Aspen dataset was employed to train the model with internal disturbance. The training was executed with 28 data points randomly selected within the dataset; the number of the training points were chosen to be equal to the number of model tunable parameters. In addition, the points having $x_U = x_{Imp} = 0.20$ were excluded from the training. The points of this dataset were utilized in the validation phase to assess the extrapolation capabilities of the model. Tetrahydrofuran and cyclohexane were used as known components of the system, while acetonitrile and benzene were considered the internal disturbance of the system (i.e., unknown component). Figure 6 reports the model capabilities on the test set with $x_U = 0.20$. This figure reports the excellent extrapolation capabilities of the model. The training space has a maximum impurity concentration of $x_U = 0.1$; therefore, the case reported in this figure is significantly outside the last training point. The predictions reported in this figure are remarkable; this is possible because of the physical characterization of the characterized hybrid model for the ACs prediction. The remarkable generalization capabilities are also highlighted by the evaluation metrics values in Table 4. Here the MAPE values are lower than 1% for all the variables and the R^2 is always larger than 0.99. This table reports the metrics over the test set, for the entire test set (i.e., the union of the random points within the training space and the value at $x_U = 0.2$) and the results obtained only with the extrapolation test points (i.e., the points at $x_U = 0.2$). The R^2 reported in this table is very close to one, and the RMSE and MAPE are significantly small for both test sets. However, the evaluations executed over the overall test set are better than the ones executed on the test set with $x_U = 0.2$. This happens because the entire test set also contains points within the investigation space (i.e., interpolative points). The interpolations are generally more accurate than the extrapolations; therefore, the error metrics are lower on the entire test set than the extrapolation set alone. Despite this scenario, the evaluation metrics computed only over the test set with $x_U = 0.2$ does not differ much from the values of the evaluation metrics computed on the entire dataset.

Table 4. Performance metrics of the hybrid model over the test set of the system containing tetrahydrofuran, cyclohexane, acetonitrile, and benzene. The system was hypothesized to have an internal disturbance represented by acetonitrile and benzene.

		RMSE	MAPE	R^2
Entire test set	$\gamma_{Tetrahydrofuran}$	1.94e-3	0.136%	0.999
	$\gamma_{Cyclohexane}$	1.04e-2	0.606%	0.998
Test set with $x_U = 0.2$	$\gamma_{Tetrahydrofuran}$	2.35e-3	0.191%	0.997
	$\gamma_{Cyclohexane}$	1.33e-2	0.972%	0.997

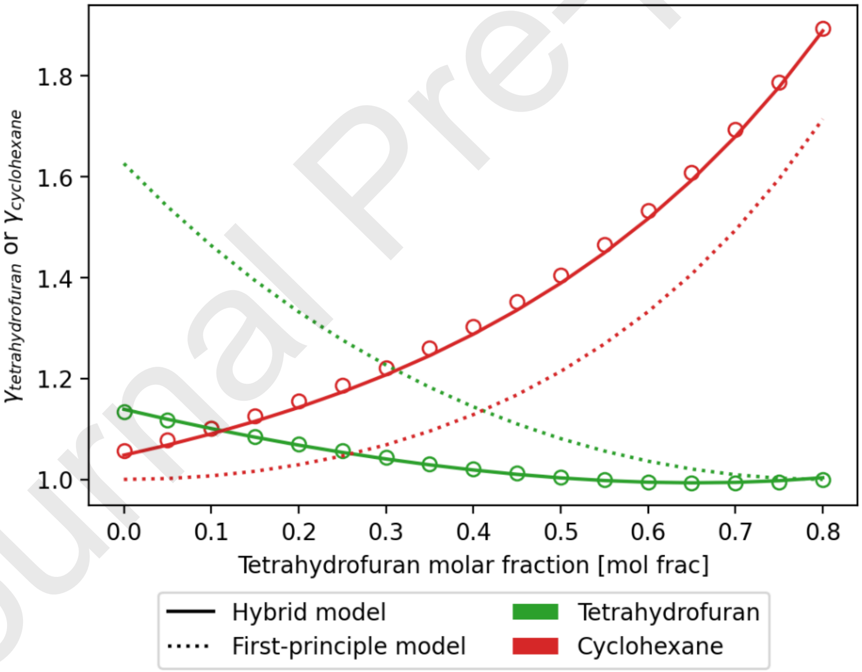

Figure 6. Extrapolation capabilities of the framework on the Aspen dataset considering an internal disturbance. The data contained in this figure refers to $x_U = 0.2$.

4. Conclusions

In this paper, we introduced a novel framework to construct physic-aware hybrid models of ACs. It contains two main parts, a first-principle model utilizing established knowledge about the known part of the system (e.g., NRTL or UNIFAC) and a statistical model able to characterize the hybrid model to the system in analysis via parameters to be identified on experimental cases. The framework was designed to handle the behavior of systems with internal or external disturbances. The methodology constructs physic-aware hybrid models by constraining the statistical part of the model using physical equations (i.e., Gibbs-Duhem equation and limit condition). It increases the generalization capabilities of the model and reduces the number of free parameters to identify during the training phase. The reduced number of parameters reduces the number of data points required for the model training, increasing the data efficiency of the methodology compared to an entirely statistical approach. This paper mathematically proved the fundamental paradigms on which the framework relies. In addition, the algorithm to construct the physics-aware hybrid models was described in detail. The algorithm was implemented in a python library available on request mailing the authors. The implemented algorithm is available on request and easy to use scientists and researchers. The application and validation of the framework were performed on two cases involving experimental data obtained from literature and a dataset containing simulation data obtained from a first-principle model calculation. The datasets refer to vapor-liquid equilibria (VLE) profiles; however, different training strategies were applied to investigate the framework flexibility over various data sources. Both cases have been treated as containing an internal and an external disturbance. The models were trained using the minimum amount of data required for this task to demonstrate the data-efficiency of the methodology. The training of the hybrid models describing the ACs was performed in combination with Raoult's law for non-ideal mixtures to prove the robustness of the methodology. All the training returned outstanding prediction accuracy, generalization, and extrapolation capabilities on the test set.

This paper focused on a single type of statistical function (i.e., polynomial) and simple validation cases. A natural prosecution of this work is the application of the methodology over other statistical functions (e.g., multivariate rational functions and artificial neural networks). Furthermore, the framework can be applied to more extensive hybrid models involving fermentative or pharmaceutical processes.

Table of symbols

Latin letters

$\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \ldots$	Matrices containing the parameters of the statistical
K	Number of known chemicals present within the system
M	Number of total chemicals present within the system
T	Temperature of the system
U	Number of unknown chemicals present within the system
x	Molar fraction of a component in the liquid phase
$\boldsymbol{x}$	Vector containing the molar fractions of the components within the liquid
y	Molar fraction of a component in the vapor phase

Greek letters

γ	Activity coefficient

Superscripts

$\bar{\square}$	Molar fraction of the unknown components
$*$	Normalized molar fraction overall the known components

Subscripts

FP	First-principle model
HM	Hybrid model
SM	Statistical model
U	Unknown component

Acknowledgements

The authors acknowledge funding from VLAIO-Catalisti projects "Real-time data-assisted process development and production in chemical applications" (HBC.2020.2455 – DAP^{2}CHEM). The authors declare no competing interests.

References

[1] J. Haydary, Introduction to Computer-Aided Process Design and Simulation, in: Chemical Process Design and Simulation, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2019: pp. 1–14. https://doi.org/10.1002/9781119311478.ch1.

[2] J. Gmehling, M. Kleiber, B. Kolbe, J. Rarey, Applications of Thermodynamics in Separation Technology, in: Chemical Thermodynamics for Process Simulation, 2nd, Completely Revised and Enlarged Edition, Wiley, Weinheim, Germany, 2019: pp. 469–504.

[3] G. Vicente, A. Paiva, T. Fornari, V. Najdanovic-Visak, Liquid-liquid equilibria for separation of tocopherol from olive oil using ethyl lactate, Chemical Engineering Journal. 172 (2011) 879–884. https://doi.org/10.1016/j.cej.2011.06.077.

[4] D. Soukup-Carne, X. Fan, J. Esteban, An overview and analysis of the thermodynamic and kinetic models used in the production of 5-hydroxymethylfurfural and furfural, Chemical Engineering Journal. 442 (2022). https://doi.org/10.1016/j.cej.2022.136313.

[5] K. Gonzalez, L. Boyer, D. Almoucachar, B. Poulain, E. Cloarec, C. Magnon, F. de Meyer, CO2 and H2S absorption in aqueous MDEA with ethylene glycol: Electrolyte NRTL, rate-based process model and pilot plant experimental validation, Chemical Engineering Journal. 451 (2023). https://doi.org/10.1016/j.cej.2022.138948.

[6] U. Domańska, M. Zawadzki, M. Królikowski, A. Lewandrowska, Phase equilibria study of binary and ternary mixtures of {N-octylisoquinolinium bis{(trifluoromethyl)sulfonyl}imide+hydrocarbon, or an alcohol, or water}, Chemical Engineering Journal. 181–182 (2012) 63–71. https://doi.org/10.1016/j.cej.2011.10.001.

[7] E. Forte, F. Jirasek, M. Bortz, J. Burger, J. Vrabec, H. Hasse, Digitalization in Thermodynamics, Chem Ing Tech. 91 (2019) 201–214. https://doi.org/10.1002/cite.201800056.

[8] A. Fredenslund, R.L. Jones, J.M. Prausnitz, Group-contribution estimation of activity coefficients in nonideal liquid mixtures, AIChE Journal. 21 (1975) 1086–1099. https://doi.org/10.1002/aic.690210607.

[9] U. Weidlich, J. Gmehling, A modified UNIFAC model. 1. Prediction of VLE, hE, and .gamma..infin., Ind Eng Chem Res. 26 (1987) 1372–1381. https://doi.org/10.1021/ie00067a018.

[10] H. Renon, J.M. Prausnitz, Local compositions in thermodynamic excess functions for liquid mixtures, AIChE Journal. 14 (1968) 135–144. https://doi.org/10.1002/aic.690140124.

[11] A. Klamt, Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena, J Phys Chem. 99 (1995) 2224–2235. https://doi.org/10.1021/j100007a062.

[12] A. Klamt, V. Jonas, T. Bürger, J.C.W. Lohrenz, Refinement and parametrization of COSMO-RS, Journal of Physical Chemistry A. 102 (1998) 5074–5085. https://doi.org/10.1021/jp980017s.

[13] K.C. Felton, H. Ben-Safar, A.A. Alexei, DeepGamma: A deep learning model for activity coefficient prediction, in: 1st Annual AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE), 2022.

[14] B. Winter, C. Winter, J. Schilling, A. Bardow, A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing, Digital Discovery. 1 (2022) 859–869. https://doi.org/10.1039/D2DD00058J.

[15] J.G. Rittig, K. Ben Hicham, A.M. Schweidtmann, M. Dahmen, A. Mitsos, Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids, Comput Chem Eng. (2023) 108153. https://doi.org/10.1016/j.compchemeng.2023.108153.

[16] P.M. Mathias, Sensitivity of process design to phase equilibrium-A new perturbation method based upon the Margules equation, J Chem Eng Data. 59 (2014) 1006–1015. https://doi.org/10.1021/je400748p.

[17] J. Burger, N. Asprion, S. Blagov, M. Bortz, Simple perturbation scheme to consider uncertainty in equations of state for the use in process simulation, J Chem Eng Data. 62 (2017) 268–274. https://doi.org/10.1021/acs.jced.6b00633.

[18] E. Baumeister, J. Burger, General Perturbation Scheme to Model Activities in Poorly Specified Liquid Mixtures, Ind Eng Chem Res. 59 (2020) 413–422. https://doi.org/10.1021/acs.iecr.9b05042.

[19] F. Jirasek, J. Burger, H. Hasse, Method for Estimating Activity Coefficients of Target Components in Poorly Specified Mixtures, Ind Eng Chem Res. 57 (2018) 7310–7313. https://doi.org/10.1021/acs.iecr.8b00917.

[20] F. Jirasek, J. Burger, H. Hasse, NEAT - NMR Spectroscopy for the Estimation of Activity Coefficients of Target Components in Poorly Specified Mixtures, Ind Eng Chem Res. 58 (2019) 9155–9165. https://doi.org/10.1021/acs.iecr.9b01269.

[21] T. Specht, H. Hasse, F. Jirasek, Predictive Thermodynamic Modeling of Poorly Specified Mixtures and Applications in Conceptual Fluid Separation Process Design, Ind Eng Chem Res. (2023). https://doi.org/10.1021/acs.iecr.3c01096.

[22] F. Jirasek, J. Burger, H. Hasse, NEAT - NMR Spectroscopy for the Estimation of Activity Coefficients of Target Components in Poorly Specified Mixtures, Ind Eng Chem Res. 58 (2019) 9155–9165. https://doi.org/10.1021/acs.iecr.9b01269.

[23] A.J.B. Ten Kate, J. Gerretzen, H.J. Van Manen, G.M. Kontogeorgis, G. Bargeman, Methodology to Predict Thermodynamic Data from Spectroscopic Analysis, Ind Eng Chem Res. 59 (2020) 21548–21566. https://doi.org/10.1021/acs.iecr.0c04043.

[24] F. Jirasek, J. Burger, H. Hasse, Application of NEAT for determining the composition dependence of activity coefficients in poorly specified mixtures, Chem Eng Sci. 208 (2019). https://doi.org/10.1016/j.ces.2019.115161.

[25] J. Frutiger, M. Jones, N.G. Ince, G. Sin, From property uncertainties to process simulation uncertainties – Monte Carlo methods in SimSci PRO/II process simulator, in: Computer Aided Chemical Engineering, Elsevier B.V., 2018: pp. 1489–1494. https://doi.org/10.1016/B978-0-444-64241-7.50243-3.

[26] J. Frutiger, C. Marcarie, J. Abildskov, G. Sin, A Comprehensive Methodology for Development, Parameter Estimation, and Uncertainty Analysis of Group Contribution Based Property Models-An Application to the Heat of Combustion, J Chem Eng Data. 61 (2016) 602–613. https://doi.org/10.1021/acs.jced.5b00750.

[27] A.M. Schweidtmann, E. Esche, A. Fischer, M. Kloft, J.U. Repke, S. Sager, A. Mitsos, Machine Learning in Chemical Engineering: A Perspective, Chem Ing Tech. 93 (2021) 2029–2039. https://doi.org/10.1002/cite.202100083.

[28] U. Di Caprio, M. Wu, F. Elmaz, Y. Wouters, N. Vandervorst, A. Anwar, S. Mercelis, S. Waldherr, P. Hellinckx, M.E. Leblebici, Hybrid modelling of a batch separation process, Comput Chem Eng. (2023) 108319. https://doi.org/10.1016/j.compchemeng.2023.108319.

[29] U. Di Caprio, M. Wu, F. Vermeire, T. Van Gerven, P. Hellinckx, S. Waldherr, E. Kayahan, M.E. Leblebici, Predicting overall mass transfer coefficients of CO2 capture into monoethanolamine in spray columns with hybrid machine learning, Journal of CO2 Utilization. 70 (2023) 102452. https://doi.org/10.1016/j.jcou.2023.102452.

[30] V.Y. Valera, M.C. Codolo, T.D. Martins, Artificial neural network for prediction of SO2 removal and volumetric mass transfer coefficient in spray tower, Chemical Engineering Research and Design. 170 (2021) 1–12. https://doi.org/10.1016/J.CHERD.2021.03.008.

[31] M.M. Ghiasi, S. Abedi-Farizhendi, A.H. Mohammadi, Modeling Equilibrium Systems of Amine-Based CO2 Capture by Implementing Machine Learning Approaches, Environ Prog Sustain Energy. 38 (2019) 13160. https://doi.org/10.1002/ep.13160.

[32] L. Chiang, B. Lu, I. Castillo, Big Data Analytics in Chemical Engineering, Annu Rev Chem Biomol Eng. 8 (2017) 63–85. https://doi.org/10.1146/annurev-chembioeng-060816-101555.

[33] S. Feyo De Azevedo, B. Dahm, F.R. Oliveira, Hybrid modelling of biochemical processes: A comparison with the conventional approach, Comput Chem Eng. 21 (1997) S751–S756. https://doi.org/10.1016/s0098-1354(97)87593-x.

[34] M. von Stosch, R. Oliveira, J. Peres, S. Feyo de Azevedo, Hybrid semi-parametric modeling in process systems engineering: Past, present and future, Comput Chem Eng. 60 (2014) 86–101. https://doi.org/10.1016/j.compchemeng.2013.08.008.

[35] M.J. Willis, M. von Stosch, Simultaneous parameter identification and discrimination of the nonparametric structure of hybrid semi-parametric models, Comput Chem Eng. 104 (2017) 366–376. https://doi.org/10.1016/j.compchemeng.2017.05.005.

[36] P. Azadi, J. Winz, E. Leo, R. Klock, S. Engell, A hybrid dynamic model for the prediction of molten iron and slag quality indices of a large-scale blast furnace, Comput Chem Eng. 156 (2022) 107573. https://doi.org/10.1016/J.COMPCHEMENG.2021.107573.

[37] K. McBride, E.I. Sanchez Medina, K. Sundmacher, Hybrid Semi-parametric Modeling in Separation Processes: A Review, Chemie Ingenieur Technik. 92 (2020) 842–855. https://doi.org/10.1002/cite.202000025.

[38] A. Hamid, A. Heryanto, S. Nurfaqihah, Z. Harun, Z.A. Putra, Hybrid modelling for remote process monitoring and optimisation, Digital Chemical Engineering. 4 (2022) 100044. https://doi.org/10.1016/j.dche.2022.100044.

[39] S.I. Sandler, Chemical, biochemical, and engineering thermodynamics, John Wiley & Sons, 2017.

[40] A. Carranza-Abaid, H.F. Svendsen, J.P. Jakobsen, Thermodynamically consistent vapor-liquid equilibrium modelling with artificial neural networks, Fluid Phase Equilib. 564 (2023). https://doi.org/10.1016/j.fluid.2022.113597.

[41] F. Jirasek, H. Hasse, Combining Machine Learning with Physical Knowledge in Thermodynamic Modeling of Fluid Mixtures, Annu Rev Chem Biomol Eng. 14 (2023) 31–51. https://doi.org/10.1146/annurev-chembioeng.

[42] F. Jirasek, R. Bamler, S. Mandt, Hybridizing physical and data-driven prediction methods for physicochemical properties, Chemical Communications. 56 (2020) 12407–12410. https://doi.org/10.1039/d0cc05258b.

[43] N. Hayer, F. Jirasek, H. Hasse, Prediction of Henry’s law constants by matrix completion, AIChE Journal. 68 (2022). https://doi.org/10.1002/aic.17753.

[44] E.I. Sanchez Medina, S. Linke, M. Stoll, K. Sundmacher, Graph neural networks for the prediction of infinite dilution activity coefficients, Digital Discovery. 1 (2022) 216–225. https://doi.org/10.1039/d1dd00037c.

[45] J.G. Rittig, K.C. Felton, A.A. Lapkin, A. Mitsos, Gibbs-Duhem-Informed Neural Networks for Binary Activity Coefficient Prediction, (2023). http://arxiv.org/abs/2306.07937.

[46] S.I. Sandler, Estimation of the Gibbs Energy and Fugacity of a component in a mixture, in: Chemical, Biochemical and Engineering Thermodynamics, 4th ed., John Wiley & Sons, Inc., 2006: pp. 399–488.

[47] A. Meurer, C.P. Smith, M. Paprocki, O. Čertík, S.B. Kirpichev, M. Rocklin, Am. Kumar, S. Ivanov, J.K. Moore, S. Singh, T. Rathnayake, S. Vig, B.E. Granger, R.P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M.J. Curry, A.R. Terrel, Š. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, A. Scopatz, SymPy: symbolic computing in Python, PeerJ Comput Sci. 3 (2017) e103. https://doi.org/10.7717/peerj-cs.103.

[48] M.C. Iliuta, I. Iliuta, O.M. Landauer, F.C. Thyrion, Salt effect of LiCl on vapor–liquid equilibrium of the acetone–methanol system, Fluid Phase Equilib. 149 (1998) 163–176. https://doi.org/10.1016/S0378-3812(98)00365-3.

[49] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, \.Ilhan Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, {SciPy} 1.0: Fundamental Algorithms for Scientific Computing in Python, Nat Methods. 17 (2020) 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Declaration of interests

☐ 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.

☒ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Mumin Enis Leblebici reports financial support was provided by Flanders Innovation & Entrepreneurship.

- A methodology constructing hybrid models for activity coefficients was proposed
- The proposed hybrid models can predict mixtures under uncertainty
- The hybrid model development is automatic and thermodynamically consistent
- The thermodynamic consistency improves both data efficiency and models reliability
- The hybrid models are trainable from laboratory data as well as plant data