Chemical Engineering Journal 504 (2025) 158667

Contents lists available at ScienceDirect

Chemical Engineering Journal

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

Advancing thermodynamic group-contribution methods by machine learning: UNIFAC 2.0

Nicolas Hayer [a], Thorsten Wendel [a], Stephan Mandt [b], Hans Hasse [a], Fabian Jirasek [a],*

[a] Laboratory of Engineering Thermodynamics, RPTU Kaiserslautern, Erwin-Schrödinger-Str. 44, 67663 Kaiserslautern, Germany

[b] Department of Computer Science, University of California, Irvine, CA 92617, USA

ARTICLE INFO

Keywords:
Group-contribution methods
Machine learning
UNIFAC
Matrix completion
Activity coefficients
Phase equilibria

ABSTRACT

Accurate prediction of thermodynamic properties is pivotal in chemical engineering for optimizing process efficiency and sustainability. Physical group-contribution (GC) methods are widely employed for this purpose but suffer from historically grown, incomplete parameterizations, limiting their applicability and accuracy. In this work, we overcome these limitations by combining GC with matrix completion methods (MCM) from machine learning. We use the novel approach to predict a complete set of pair-interaction parameters for the most successful GC method: UNIFAC, the workhorse for predicting activity coefficients in liquid mixtures. The resulting new method, UNIFAC 2.0, is trained and validated on more than 224,000 experimental data points, showcasing significantly enhanced prediction accuracy (e.g., nearly halving the mean squared error) and increased scope by eliminating gaps in the original model's parameter table. Moreover, the generic nature of the approach facilitates updating the method with new data or tailoring it to specific applications.

1. Introduction

Understanding the thermodynamic properties of mixtures is indispensable in chemical engineering and various related disciplines. However, the vast combinatorial diversity of mixtures makes it impossible to study each relevant mixture experimentally, necessitating reliable prediction methods. Group-contribution (GC) methods address this challenge by deconstructing components into structural groups, significantly reducing the number of parameters since the number of structural groups is much smaller than those of individual components. These methods rely on modeling pair interactions between these structural groups to describe mixture behavior. The effectiveness of GC methods hinges on selecting suitable groups and accurately determining their interaction parameters, both of which depend crucially on the database used for method development and parameterization.

Among GC methods, UNIFAC stands out as the most sophisticated and widely adopted approach for predicting activity coefficients in liquid mixtures. Since its introduction in 1975 [1], UNIFAC has undergone continuous refinement and improvement [2–7], becoming integral to industrial process simulations. Available in both public [7] and commercial [8] formats, UNIFAC supports diverse applications, including variants like UNIFAC LLE [9] for predicting liquid–liquid equilibria. All UNIFAC variants rely on the same equations but differ in the number and type of groups considered and their parameterization. The process of finding suitable UNIFAC parameters was, in the past, sequential and based on a stepwise extension whenever data became available. This tedious process makes it very difficult to modify decisions taken at early steps.

This study addresses the challenges of updating and improving UNIFAC by leveraging modern computational techniques, aiming to enhance prediction accuracy and expand its applicability across a broader range of components and mixtures.

Throughout this work, we reference the latest published version of UNIFAC. It was trained on a broad data basis focusing on vapor–liquid equilibrium data to develop a widely applicable model, not one for some specific purpose [7]. It is astonishing that, despite the importance of UNIFAC, this version is about 20 years old. The leading developers of UNIFAC have updated the method since then, but they have not disclosed these updates – they are only available for members of the UNIFAC consortium. One might ask why no one else has updated this important method since then. The answer to this question is undoubtedly related to the considerable effort required to do this when the conventional strategy is used. Another issue is the accessibility of suitable data. For simplicity, we will label the reference version of UNIFAC [7] as UNIFAC 1.0 here.

UNIFAC describes the molar excess Gibbs energy g^{E}, of a mixture as a function of temperature T, and composition. From g^{E}, the activity coefficients of the components i, γ_i, in the mixture are obtained. UNIFAC

* Corresponding author.
E-mail address: fabian.jirasek@rptu.de (F. Jirasek).

https://doi.org/10.1016/j.cej.2024.158667
Received 16 August 2024; Received in revised form 27 November 2024; Accepted 14 December 2024
Available online 20 December 2024
1385-8947/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

contains group-specific parameters, namely, a size parameter (R_k) and a surface parameter (Q_k), as well as binary pair-interaction parameters (there are two for each group combination $a_{mn} \neq a_{nm}$, which we will often refer to simply as a_{mn} for simplicity). UNIFAC 1.0 considers 54 *main groups*, subdivided into 113 *subgroups* [7].

Applying UNIFAC 1.0 to a given mixture requires the following: (i) all components of the mixture must be decomposable into the 113 subgroups, (ii) the parameters R_k and Q_k must be available for each relevant subgroup k, and (iii) the pair-interaction parameters a_{mn} must be available for each binary combination of the relevant main groups m and n (all subgroups of a given main group share the same interaction parameters). The group parameters R_k and Q_k are available for all 113 groups [10], but interaction parameters a_{mn} are missing for many pairs of groups. Specifically, numbers for the interaction parameters are only available for 44% of all pairs of groups; Fig. S.1 in the Supporting Information illustrates this. The missing pair-interaction parameters, in some cases due to the challenging fitting process and in other cases due to the lack of experimental data for direct fitting, severely hampers the applicability of UNIFAC 1.0 (a single missing relevant parameter prevents the application of the model).

In this work, we introduce a new way of determining the interaction parameters of GC methods based on machine learning. The pair-interaction parameters can be treated as elements of a square matrix with dimensions 54 × 54, where the size corresponds to the number of main groups. Since experimental data are only available for a fraction of the pair-interaction parameters, many entries of this matrix cannot be fitted directly, resulting in a matrix completion problem that can, in general, be solved by matrix completion methods (MCM) [11–13]. As numbers for all entries are found, the problem of missing parameters does not exist anymore. In the MCM, so-called group features are determined for all groups from a fit to experimental data on activity coefficients. The entire data set is considered during the fit, and a well-defined learning algorithm (in our case, a Bayesian one) is applied. This method replaces the sequential, intuitively guided procedure previously used to determine pair-interaction parameters. As the number of features to be determined scales linearly with the number of main groups N_{MG} ($\mathcal{O}(N_{\mathrm{MG}})$), it is much lower than the number of interaction parameters ($\mathcal{O}(N_{\mathrm{MG}}^2)$). Consequently, the parameterization of the MCM is significantly more robust than a direct fit of the interaction parameters to the experimental data.

From the features of any two groups m and n of interest, the entries of the interaction parameter matrix a_{mn} are found by a simple matrix multiplication, resulting in a complete set of interaction parameters, thus facilitating the prediction of the activity coefficients for any binary mixture given its structural group composition at any temperature and concentration.

The result is UNIFAC 2.0, a hybrid model consisting of the framework of the physical UNIFAC model, in which an MCM from machine learning is embedded. While the MCM used for predicting missing interaction parameters from group-specific features is rather simple, UNIFAC 2.0 fully retains the non-linear UNIFAC equations, allowing it to also describe complex interactions between structural groups.

In prior work, we have already employed MCMs for directly predicting thermodynamic properties of binary mixtures [14–18]. We have also shown that MCMs are suitable for predicting pair-interaction parameters between components [19] and structural groups using synthetic training data [20]. The synthetic training data in Ref. [20] were derived from the existing parameter tables of UNIFAC 1.0, providing a practical starting point. However, the limited prediction accuracy of this approach underscores the need for a more comprehensive approach. In this work, we present the first application of MCMs to the development of GC methods for predicting activity coefficients with direct end-to-end training on several hundred thousand experimental data points.

2. Development of UNIFAC 2.0

2.1. General framework

Fig. 1 illustrates UNIFAC 2.0 with end-to-end training of MCM features, which is compared to UNIFAC 1.0 with sequential parameter fitting. Both UNIFAC variants are based on the same structural groups and physical model equations. UNIFAC 2.0 was trained on experimental logarithmic activity coefficients ($\ln \gamma_i$) in binary mixtures derived from vapor–liquid equilibrium data for binary mixtures, cf. Section “Data” for details.

The MCM can only work if the available entries of the matrix are correlated. The MCM learns these correlations and represents them by the features. This enables the prediction of missing matrix entries through learned features. Each pair-interaction parameter a_{mn} is thereby modeled as follows:

$$a_{mn} = \boldsymbol{\theta}_m^{\mathrm{T}} \cdot \boldsymbol{\beta}_n. \tag{1}$$

Here, $\boldsymbol{\theta}_m$ and $\boldsymbol{\beta}_n$ are column vectors of length K, with K representing the latent dimension, a hyperparameter that was determined in preliminary studies and set to $K = 8$. The feature vectors $\boldsymbol{\theta}_m$ and $\boldsymbol{\beta}_n$ are an abstract characterization of the structural groups determining their interactions with other groups.

A Bayesian approach is applied to train the model, treating each logarithmic activity coefficient $\ln \gamma_i$, each feature, and each interaction parameter a_{mn} as a random variable following a probability distribution, detailed further in the Section “Probabilistic Model”. From the model training, we obtain a probability density for each a_{mn}, the mean of which is used to obtain the scalar value for each parameter. These scalar values are then used in all subsequent evaluations. The completed set of interaction parameters a_{mn}, derived from training on all considered binary data, and the subgroup-specific size parameters R_k and Q_k for using UNIFAC 2.0 are provided freely in the Supporting Information. The size parameters are identical to those of the published UNIFAC 1.0 version.

The relevance of UNIFAC 2.0 becomes apparent when analyzing the applicability of UNIFAC 1.0 and 2.0 considering an example: the Dortmund Data Bank (DDB) [10], which is the most extensive database for thermodynamic properties, presently lists 39,587 unique components that can be broken down into the published UNIFAC subgroups, which translates into more than 783 million possible binary mixtures. Of these binary mixtures, UNIFAC 1.0 is limited to predicting only 58% due to missing pair-interaction parameters, whereas UNIFAC 2.0 can be applied to all these mixtures. For multi-component mixtures, the fraction of mixtures that can only be predicted with UNIFAC 2.0 increases dramatically with an increasing number of components, as a mixture drops out if only a single parameter (pair) is missing.

Besides the hybrid model described above, a variant that is based on symmetrical pair-interaction energies $U_{mn} = U_{nm}$ between main groups instead of the asymmetric parameters a_{mn} was developed and tested. The symmetric model has fewer parameters and performs almost as well as the asymmetric model. We report on the asymmetric model here, as it is the standard way to use UNIFAC, and the results can be implemented and used in a very simple manner. Details on the symmetric model are given in the Supporting Information. For a short background discussion of the two variants applied to component-wise pair interactions, see Ref. [19].

2.2. Probabilistic model

Our proposed probabilistic model integrates observations ($\ln \gamma_i$) and the latent variables (LVs) that characterize UNIFAC main groups ($\boldsymbol{\theta}_m$, $\boldsymbol{\beta}_n$) within a Bayesian framework. UNIFAC 2.0 adheres to Bayes' theorem by incorporating three probability distributions: prior, likelihood, and posterior. The prior describes knowledge about the LVs *prior* to

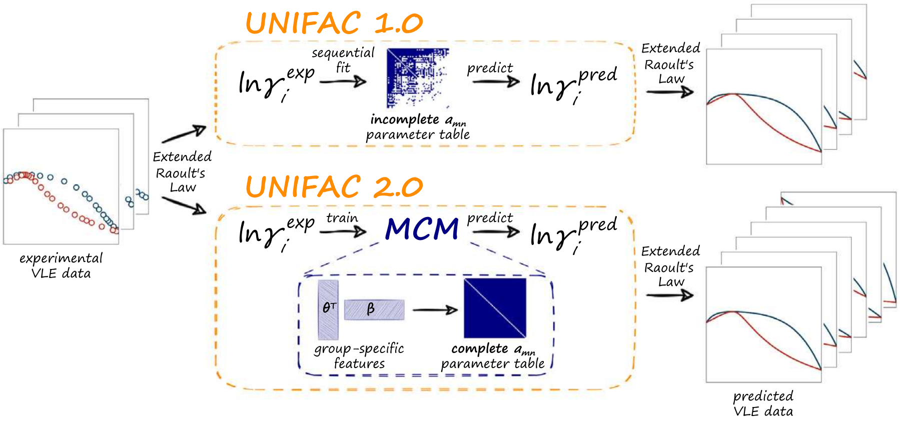

Fig. 1. Comparison of UNIFAC 1.0 and UNIFAC 2.0. UNIFAC 1.0 relies on sequential parameter fitting guided by intuition, whereas UNIFAC 2.0 integrates a matrix completion method (MCM) for predicting pair-interaction parameters into the UNIFAC framework. UNIFAC 2.0 is trained end-to-end on experimental logarithmic activity coefficients ($\ln\gamma_i$) derived from binary vapor–liquid equilibrium (VLE) data. After training, the completed pair-interaction parameter matrix facilitates accurate predictions of phase diagrams for a wide range of binary or multi-component mixtures.

fitting the model to the training data. The likelihood constitutes a probability distribution over the observable quantity ($\ln\gamma_i$ here) conditioned on the LVs, i.e., it specifies how the LVs manifest themselves in the data for $\ln\gamma_i$. The aim of Bayesian inference is finding the posterior, which is the probability distribution over the LVs that encapsulates the updated beliefs about the LVs after considering both prior information and empirical data. A more detailed explanation of the key terms in Bayesian modeling is given in the Supporting Information.

Specifically, all $\ln\gamma_i$ and LVs are modeled as independent random variables. A standard normal distribution, i.e., a normal distribution with the mean $\mu = 0$ and the standard deviation $\sigma = 1$, is used as prior for each LV. The likelihood of observing $\ln\gamma_i$, given the LVs, follows a Cauchy distribution centered around the predicted activity coefficients $\ln\gamma_i^{\text{UNIFAC 2.0}}$ with scale parameter λ:

$$p(\ln\gamma_i \mid \boldsymbol{\theta}_m, \boldsymbol{\beta}_n) = \text{Cauchy}(\ln\gamma_i^{\text{UNIFAC 2.0}},\ \lambda), \tag{2}$$

where $\ln\gamma_i^{\text{UNIFAC 2.0}}$ is determined via the standard UNIFAC equations [7] using the predicted interaction parameters a_{mn}:

$$\ln\gamma_i^{\text{UNIFAC 2.0}} = \text{UNIFAC}(a_{mn}, R_k, Q_k, \boldsymbol{x}, T). \tag{3}$$

Here, R_k and Q_k are the subgroup-specific size parameters, T is the temperature, and $\boldsymbol{x}$ corresponds to the composition (expressed as mole fractions) of the considered binary mixture. The use of a Cauchy distribution for the likelihood is motivated by its robustness to outliers in the experimental data. Unlike the normal distribution, the heavy-tailed nature of the Cauchy distribution reduces the influence of extreme values, ensuring that the training process remains stable even when the data set contains flawed data points.

Written in Pyro, a probabilistic programming language based on Python and PyTorch support [21], our probabilistic model adopts stochastic variational inference (VI) [22] for posterior approximation. This approach leverages the Adam optimizer [23], with a learning rate of 0.15. A normal distribution is employed as the variational distribution, with all LVs being treated independently. During training, this approach facilitates learning variational parameters, specifically the mean and standard deviation, for each LV. Based on preliminary studies that have shown robust behavior in terms of predictive performance, the hyperparameters $K = 8$ and $\lambda = 0.4$ were chosen.

Post-training, the LVs inferred from the posterior enable, via Eqs. (1) and (3), the prediction of $\ln\gamma_i$ for any binary or multi-component mixture, including unstudied ones, whose components can be decomposed in the 113 UNIFAC subgroups.

2.3. Data

Experimental data on vapor–liquid equilibria (VLE) and activity coefficients at infinite dilution in binary mixtures were taken from the largest database for thermodynamic properties, the DDB [10]. In the preprocessing phase, data points identified as poor quality by the DDB were excluded, and the focus was narrowed to binary mixtures whose components can be decomposed into UNIFAC subgroups. Furthermore, only VLE data points from which the activity coefficients γ_i of components i in the mixture could be calculated using the extended Raoult's law

$$\gamma_i = \frac{p\,y_i}{p_i^{\text{vap}}\,x_i} \tag{4}$$

were used. Here, p is the total pressure and p_i^{vap} the vapor pressure of component i, while x_i and y_i correspond to the mole fractions of component i in the liquid and vapor phases, respectively.

3. Results

3.1. Overall performance of UNIFAC 2.0

To evaluate the performance of UNIFAC 2.0 and compare it to that of the original UNIFAC 1.0, we employ the mean absolute error (MAE) and the mean squared error (MSE) in the logarithmic activity coefficients $\ln\gamma_i$, which are calculated mixture-wise (from the scores for each binary mixture) to ensure that each mixture is weighted equally in the final score and frequently measured mixtures do not lead to a false impression of the model quality.

In the following, we focus on the predictions of UNIFAC 2.0 obtained after training the hybrid model on all available data points from our database. We have chosen this way for assessing our model since this is likely also the case for UNIFAC 1.0, as the people maintaining UNIFAC and the DDB are essentially the same (although the exact training set of UNIFAC 1.0 has not been disclosed), so we consider the comparison fair. Nevertheless, as described in the following subsections, two additional extrapolation tests were carried out with UNIFAC 2.0 to dispel doubts about its predictive capacity.

The performance of UNIFAC 2.0 on all available experimental data is shown in Fig. 2 and compared to UNIFAC 1.0. Since UNIFAC 2.0 has a more extensive scope than UNIFAC 1.0, a distinction is made: all data points that can be predicted with both methods are labeled as the “UNIFAC 1.0 horizon”, whereas all data points that can only be predicted with UNIFAC 2.0 are labeled as “UNIFAC 2.0 only”.

Fig. 2(a) clearly shows the superior prediction accuracy of UNIFAC 2.0 over UNIFAC 1.0 in both error scores. The MSE can

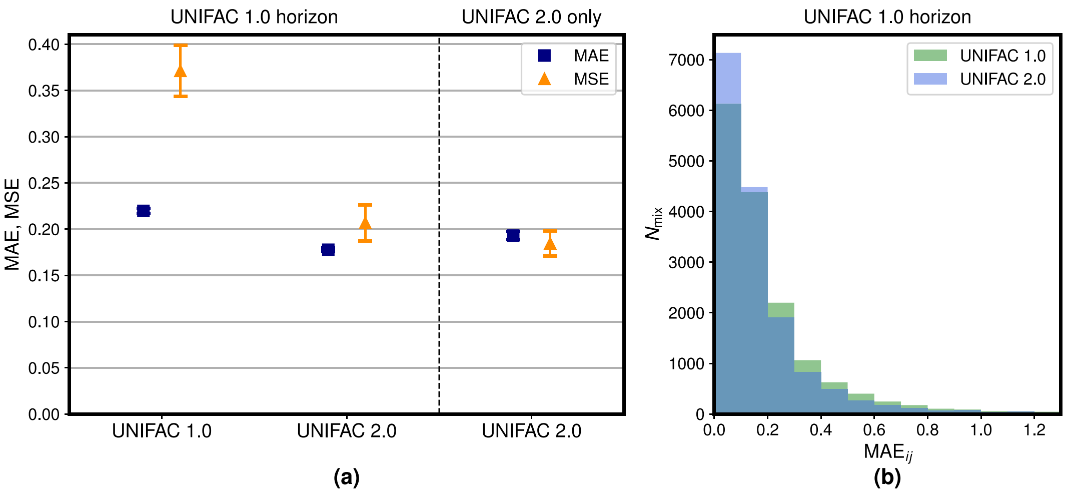

Fig. 2. Comparison of results for $\ln \gamma_i$ with UNIFAC 1.0 and UNIFAC 2.0 for different data sets: the "UNIFAC 1.0 horizon" comprises 210,767 data points for 15,758 binary mixtures, while an additional 13,795 experimental data points for 2957 binary mixtures can only be predicted with UNIFAC 2.0 ("UNIFAC 2.0 only"). (a) Mean absolute error (MAE) and mean squared error (MSE) of the model predictions. Error bars denote standard errors of the means. (b) Histogram of the number of binary mixtures N_{mix} that can be predicted with an MAE in a certain interval. The MAE range shown in panel (b) comprises 98.8% (UNIFAC 1.0) and 99.4% (UNIFAC 2.0) of all mixtures.

almost be halved compared to the original, demonstrating UNIFAC 2.0's effectiveness in reducing the occurrence of outliers. Table S.1 in the Supporting Information highlights the 20 binary mixtures with the largest improvement in prediction accuracy (MSE) achieved by UNIFAC 2.0 compared to UNIFAC 1.0. Notably, mixtures involving methoxy groups paired with silane groups and those with water paired with chlorinated aromatic components show significant improvements, indicating that these specific interactions benefit greatly from the updated parameters in UNIFAC 2.0. Even more importantly, the new method not only improves accuracy for data points within the predictive range of UNIFAC 1.0, but it also maintains this accuracy for data points beyond the scope of UNIFAC 1.0, cf. the results for the "UNIFAC 2.0 only" set.

In Fig. 2(b), a detailed analysis of the MAE for the UNIFAC 1.0 horizon in the form of a histogram of individual binary mixture scores is shown. It underpins that UNIFAC 2.0 achieves an exceptional prediction accuracy: for 7133 mixtures, the MAE is below 0.1, and thereby in the range of the experimental uncertainty. This accuracy is achieved for only 6133 mixtures with UNIFAC 1.0.

The activity coefficients obtained by UNIFAC 2.0 can be used directly to predict phase equilibria of mixtures, which are at the core of many tasks in chemical engineering. For instance, vapor–liquid phase diagrams are crucial for designing and optimizing distillation processes, where the separation efficiency relies on accurate predictions of boiling and dew points. They also play a key role in azeotropic and extractive distillation, where deviations from ideality must be accurately modeled in order to select suitable entrainers. Beyond distillation, they are also directly applicable in absorption and stripping processes, where the vapor–liquid phase equilibrium determines the efficiency of gas capture or solvent recovery. In Fig. 3, we show six examples of isothermal vapor–liquid phase diagrams predicted by UNIFAC 2.0, cf. Section "Data" for computational details. All six mixtures are part of the "UNIFAC 2.0 only" set, i.e., they cannot be modeled with the original UNIFAC 1.0. UNIFAC 2.0 accurately describes the phase behavior of all these mixtures. The examples shown in Fig. 3 represent typical cases and were selected to cover different types of phase behavior, ranging from small deviations of the ideal behavior to low-boiling azeotropes.

Furthermore, although no data on multi-component mixtures were used for training UNIFAC 2.0, the underlying physical framework of UNIFAC also enables predictions for such mixtures. As examples, Fig. 4 shows isothermal vapor–liquid phase diagrams for two ternary mixtures selected from the "UNIFAC 2.0 only" set, i.e., for which UNIFAC 1.0 is not applicable. For each data point, the temperature and the liquid-phase composition (blue symbols in Fig. 4) were specified and used to predict the corresponding vapor-phase composition in equilibrium with UNIFAC 2.0 (shown as filled orange symbols), which was then compared to the experimentally determined vapor-phase composition (open orange symbols). Excellent accuracy is found.

The results demonstrate the very good performance of UNIFAC 2.0, which outperforms UNIFAC 1.0 not only in terms of applicability by closing all gaps in its parameter table but even in terms of prediction accuracy. This highlights UNIFAC 2.0 as a compelling approach to predicting activity coefficients, particularly as it retains the classic UNIFAC framework. This retention facilitates straightforward implementation in process simulators, ensuring broad accessibility and automatic applicability to multi-component mixtures – a significant advantage over other state-of-the-art machine learning approaches. Among these, HANNA, a recently developed hard-constraint neural network [24], is, to our knowledge, currently the most accurate model for predicting activity coefficients in binary mixtures. HANNA's accuracy is achieved through a much more flexible architecture, using more than 70 times the number of parameters compared to UNIFAC 2.0, complicating its direct use in process simulators. Furthermore, HANNA is presently restricted to binary mixtures, whereas UNIFAC 2.0 can intrinsically and consistently extrapolate to multi-component mixtures. These trade-offs highlight the complementary strengths of UNIFAC 2.0 and other machine learning approaches like HANNA, which address different aspects of activity coefficient prediction and meet different user needs.

3.2. Extrapolation to unknown components

In a study to evaluate the capacity of UNIFAC 2.0 to extrapolate to unknown components, 100 randomly selected components were intentionally excluded from the training by withholding all data points for systems containing any of these components from the training set and using the systems removed from the training set as the test set. This test set contains 27,287 data points and covers 2603 different binary mixtures. The results for this test set are shown in Fig. 5, which, again, contains the result from UNIFAC 1.0 for comparison.

Fig. 5 shows that the accuracy of the true predictions with UNIFAC 2.0 obtained by withholding the test data during the training (open symbols) is only marginally lower than that of the UNIFAC 2.0 version that was trained on all data points (closed symbols); this holds for both the "UNIFAC 1.0 horizon" and the "UNIFAC 2.0 only" data sets. Furthermore, also in this true predictive test case, UNIFAC 2.0

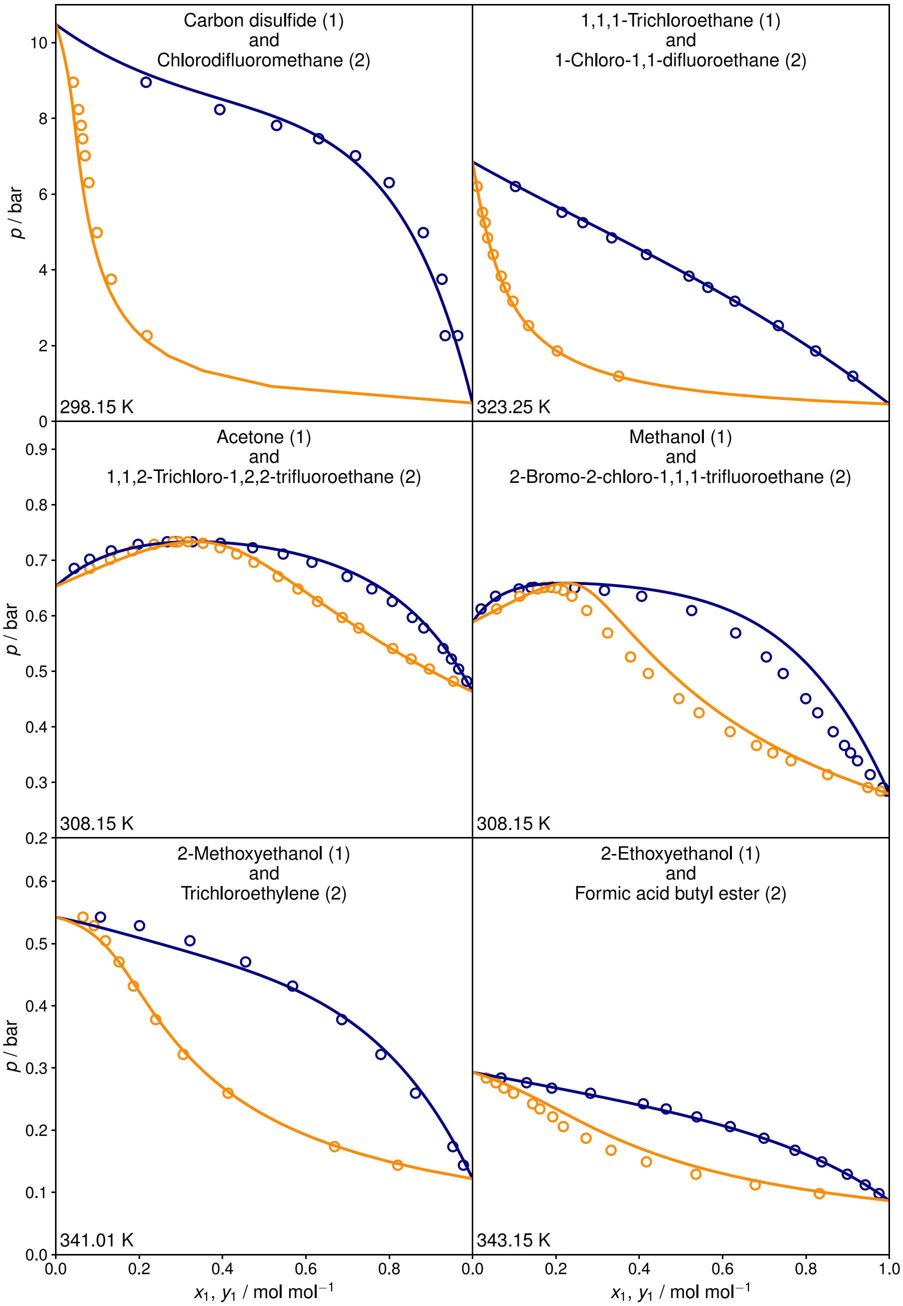

Fig. 3. Prediction of isothermal vapor–liquid phase diagrams for binary mixtures with UNIFAC 2.0 (lines) and comparison to experimental data from the DDB (symbols). Blue: bubble point curves. Orange: dew point curves.

outperforms UNIFAC 1.0, especially considering the MSE, even though it is likely that UNIFAC 1.0 has been trained on most of the test data points, as discussed above. These findings highlight, on the one hand, the robustness of UNIFAC 2.0 and, on the other hand, the predictive qualities of this hybrid model.

3.3. Extrapolation to unknown pair-interaction parameters

Another, even more challenging, test was carried out by randomly choosing 100 combinations of UNIFAC main groups for which experimental data are available and withholding the data on all systems in which any of the chosen combinations of groups occurs from the training of UNIFAC 2.0. In this way, the capacity of the hybrid model to predict pair-interaction parameters a_{mn} that cannot be obtained by direct fitting is investigated. For each of the 100 selected main group combinations, illustrated in Fig. S.4 in the Supporting Information, a test set was created that includes the data for those systems in which the selected group combination occurs. All other data points were used to train the model, and the predictions on the test set were evaluated. This process was repeated for all selected main group combinations. MAE and MSE were calculated for each test set. Fig. 6 shows the average error scores over all 100 test sets. Again, the results are compared to those of UNIFAC 1.0 and the UNIFAC 2.0 version trained on all data points. Note that the 100 test sets vary strongly in the number of data points and different binary mixtures, as shown in Table S.2 in the Supporting Information. This table also includes the MAE for each individual test set.

The comparison of the UNIFAC 2.0 predictions to the UNIFAC 1.0 predictions on the “UNIFAC 1.0 horizon” in Fig. 6 reveals that the *truly predicted* pair-interaction parameters of UNIFAC 2.0 outperform those

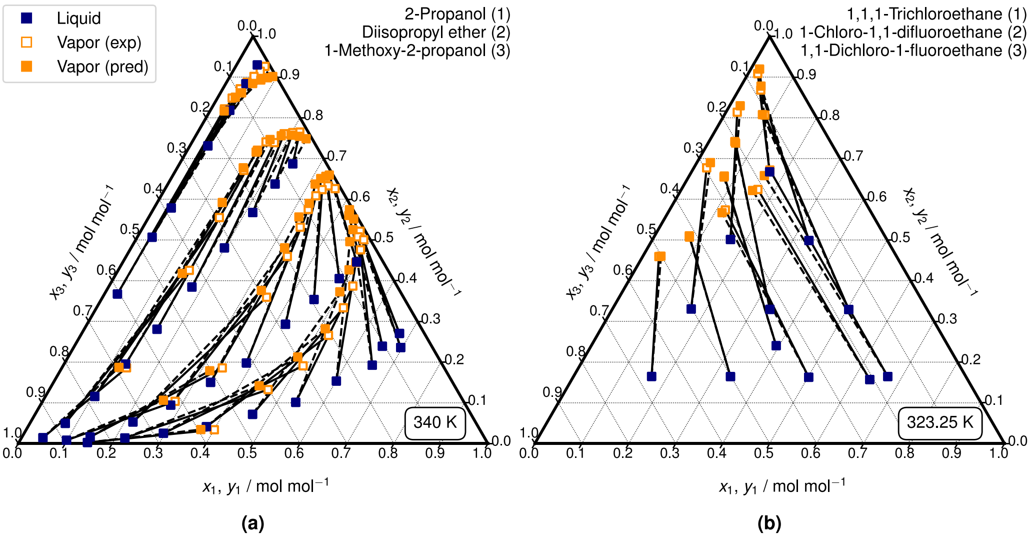

Fig. 4. Prediction of isothermal vapor–liquid phase diagrams for ternary mixtures with UNIFAC 2.0 (pred) and comparison to experimental data (exp) from the DDB. The temperature and the composition of the liquid phase were specified, and the composition of the corresponding vapor phase in equilibrium was predicted. Solid lines are experimental conodes, dashed lines are predicted conodes.

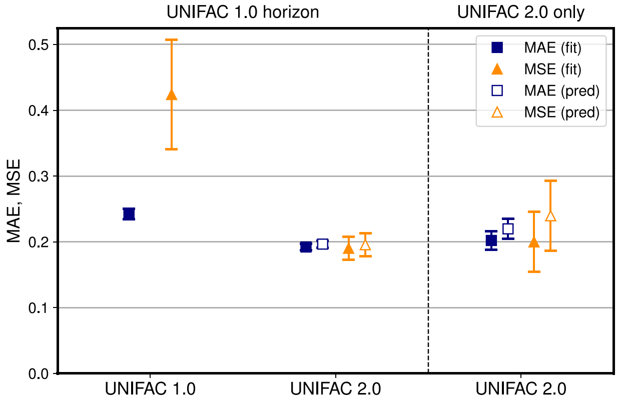

Fig. 5. Mean absolute error (MAE) and mean squared error (MSE) of the predicted $\ln \gamma_i$ of mixtures containing unobserved components with UNIFAC 2.0 (pred). For comparison, the results of UNIFAC 2.0 trained on all experimental data and UNIFAC 1.0 are also shown (fit). The “UNIFAC 1.0 horizon” comprises 25,998 data points for 2202 binary mixtures, while an additional 1289 experimental data points for 401 binary mixtures can only be predicted by UNIFAC 2.0 (“UNIFAC 2.0 only”). Error bars denote standard errors of the means.

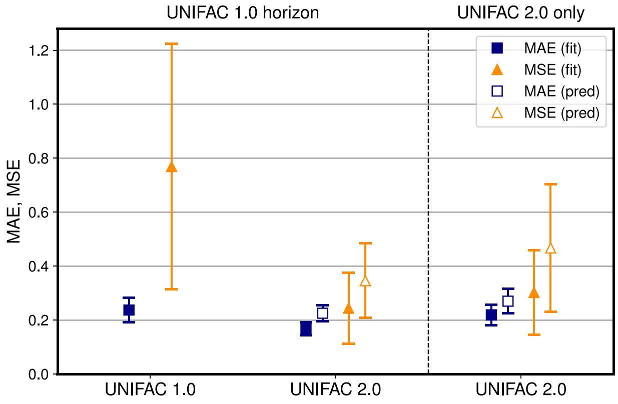

Fig. 6. Mean absolute error (MAE) and mean squared error (MSE) of the predicted $\ln \gamma_i$ averaged over 100 test sets with UNIFAC 2.0 (pred). The test sets were created by selecting all data points for which a specific interaction parameter a_{mn} is relevant, cf. Table S.2 in the Supporting Information. The results for UNIFAC 2.0 trained on all experimental data and UNIFAC 1.0 are shown for comparison (fit). Error bars denote standard errors of the means.

of UNIFAC 1.0, which were presumably *fitted* to the experimental data used for evaluation here; this is particularly evident considering the MSE. When comparing the true predictions with UNIFAC 2.0 (open symbols) to those of UNIFAC 2.0 trained on the whole experimental database (full symbols), a slight reduction in prediction accuracy is observed, as expected. However, the differences are small, which demonstrates the robustness of UNIFAC 2.0. The increased standard error associated with the MSE for UNIFAC 1.0 can be attributed to individual test sets for which the predictions are extremely poor.

The results of these tests demonstrate the capability of UNIFAC 2.0 to accurately predict pair-interaction parameters, which enormously increases the scope of this group-contribution method. UNIFAC 2.0 is not only much more applicable than UNIFAC 1.0, but its predictions are also more accurate, as shown by the comparison on the shared horizon. Hence, UNIFAC 2.0 should not only be used when UNIFAC 1.0 cannot be applied, but it should replace UNIFAC 1.0 as the default method for predicting activity coefficients. The fact that UNIFAC 2.0 performs better than UNIFAC 1.0 as measured by lumped criteria, such as the MAE and MSE, that we have used here for describing the performance on a broad database does not exclude, of course, that for specific systems, UNIFAC 1.0 may give better results. Implementing UNIFAC 2.0 is as simple as possible: one must only substitute the original (incomplete) UNIFAC parameter table, e.g., in an established process simulator, with the completed one, which we provide in the Supporting Information. This ease of implementation clearly distinguishes our hybrid model from other machine learning methods for property prediction.

4. Conclusions

Group-contribution (GC) methods are widely used workhorses for the prediction of thermodynamic properties of materials. Here, we study how they can be combined with methods from machine learning to obtain hybrid models that outperform their physical parent models. This is demonstrated here for the GC model UNIFAC for predicting activity coefficients in liquid mixtures. UNIFAC is one of the most important GC methods, broadly used in engineering, and implemented in basically all process simulation packages. Like most GC methods for predicting properties of mixtures, UNIFAC is based on the concept of group pair interactions. We demonstrate that these pair interactions can be learned and predicted with matrix completion methods (MCM) from machine learning. The resulting new hybrid model, UNIFAC 2.0, is systematically compared to its physical parent model, UNIFAC 1.0.

In contrast to the UNIFAC 1.0 parameter table, which has significant gaps, the parameter table of UNIFAC 2.0 obtained from the MCM has no gaps, leading to a substantial increase in the range of applicability. One could expect to have to pay for this increase in applicability with a deterioration of the accuracy of the predictions — but this is not the case: UNIFAC 2.0 is better than its parent model in both regards.

The hybrid approach described here also has essential advantages regarding the workflow: as the physical framework is kept, the new model can be implemented very easily in existing software packages; only parameter tables have to be updated to use its advantages. The full UNIFAC 2.0 parameter table is provided in the Supporting Information accompanying this paper. Furthermore, the end-to-end training of the hybrid model to experimental data can be carried out in an automated manner so that updates can be supplied easily if new data become available or targets shift; also, tailored versions of the model, adapted to special needs, can be obtained easily.

CRediT authorship contribution statement

Nicolas Hayer: Writing – original draft, Methodology, Conceptualization. **Thorsten Wendel:** Methodology. **Stephan Mandt:** Writing – review & editing, Conceptualization. **Hans Hasse:** Writing – review & editing, Conceptualization. **Fabian Jirasek:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

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

Acknowledgments

We gratefully acknowledge financial support by Carl Zeiss Foundation in the project “Process Engineering 4.0”, as well as by Deutsche Forschungsgemeinschaft in the Priority Program 2363, and in the Emmy Noether Project of FJ.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cej.2024.158667.

Data availability

The training data are available in the Dortmund Data Bank. The completed UNIFAC parameter set predicted in this work and the Python source code of the new model will be made available.

References

[1] A. Fredenslund, R.L. Jones, J.M. Prausnitz, Group-contribution estimation of activity coefficients in nonideal liquid mixtures, AIChE J. 21 (1975) 1086–1099.

[2] S. Skjold-Jorgensen, B. Kolbe, J. Gmehling, P. Rasmussen, Vapor-liquid equilibria by UNIFAC group contribution. Revision and extension, Ind. Eng. Chem. Process Des. Dev. 18 (1979) 714–722.

[3] J. Gmehling, P. Rasmussen, A. Fredenslund, Vapor–liquid equilibriums by UNIFAC group contribution. Revision and extension. 2, Ind. Eng. Chem. Process Des. Dev. 21 (1982) 118–127.

[4] E.A. Macedo, U. Weidlich, J. Gmehling, P. Rasmussen, Vapor–liquid equilibriums by UNIFAC group contribution. Revision and extension. 3, Ind. Eng. Chem. Process Des. Dev. 22 (1983) 676–678.

[5] D. Tiegs, P. Rasmussen, J. Gmehling, A. Fredenslund, Vapor–liquid equilibria by UNIFAC group contribution. 4. Revision and extension, Ind. Eng. Chem. Res. 26 (1987) 159–161.

[6] H.K. Hansen, P. Rasmussen, A. Fredenslund, M. Schiller, J. Gmehling, Vapor–liquid equilibria by UNIFAC group contribution. 5. Revision and extension, Ind. Eng. Chem. Res. 30 (1991) 2352–2355.

[7] R. Wittig, J. Lohmann, J. Gmehling, Vapor−liquid equilibria by UNIFAC group contribution. 6. Revision and extension, Ind. Eng. Chem. Res. 42 (2003) 183–188.

[8] The UNIFAC Consortium, 2023, http://www.unifac.org.

[9] T. Magnussen, P. Rasmussen, A. Fredenslund, UNIFAC parameter table for prediction of liquid-liquid equilibriums, Ind. Eng. Chem. Process Des. Dev. 20 (1981) 331–339.

[10] Dortmund Data Bank, 2023, www.ddbst.com.

[11] A. Ramlatchan, M. Yang, Q. Liu, M. Li, J. Wang, Y. Li, A survey of matrix completion methods for recommendation systems, Big Data Min. Anal. 1 (2018) 308–323.

[12] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42 (2009) 30–37.

[13] R. Salakhutdinov, A. Mnih, Bayesian probabilistic matrix factorization using Markov chain Monte Carlo, in: Proceedings of the 25th International Conference on Machine Learning, New York, NY, 2008.

[14] F. Jirasek, R. Bamler, S. Mandt, Hybridizing physical and data-driven prediction methods for physicochemical properties, Chem. Commun. 56 (2020) 12407–12410.

[15] F. Jirasek, R.A.S. Alves, J. Damay, R.A. Vandermeulen, R. Bamler, M. Bortz, S. Mandt, M. Kloft, H. Hasse, Machine learning in thermodynamics: Prediction of activity coefficients by matrix completion, J. Phys. Chem. Lett. 11 (2020) 981–985.

[16] J. Damay, F. Jirasek, M. Kloft, M. Bortz, H. Hasse, Predicting activity coefficients at infinite dilution for varying temperatures by matrix completion, Ind. Eng. Chem. Res. 60 (2021) 14564–14578.

[17] N. Hayer, F. Jirasek, H. Hasse, Prediction of Henry’s law constants by matrix completion, AIChE J. 68 (2022) e17753.

[18] O. Großmann, D. Bellaire, N. Hayer, F. Jirasek, H. Hasse, Database for liquid phase diffusion coefficients at infinite dilution at 298 K and matrix completion methods for their prediction, Digit. Discov. 1 (2022) 886–897.

[19] F. Jirasek, R. Bamler, S. Fellenz, M. Bortz, M. Kloft, S. Mandt, H. Hasse, Making thermodynamic models of mixtures predictive by machine learning: matrix completion of pair interactions, Chem. Sci. 13 (2022) 4854–4862.

[20] F. Jirasek, N. Hayer, R. Abbas, B. Schmid, H. Hasse, Prediction of parameters of group contribution models of mixtures by matrix completion, Phys. Chem. Chem. Phys. : PCCP 25 (2023) 1054–1062.

[21] E. Bingham, J.P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, N.D. Goodman, Pyro: Deep universal probabilistic programming, J. Mach. Learn. Res. 20 (2019) 1–6.

[22] D.M. Blei, A. Kucukelbir, J.D. McAuliffe, Variational inference: A review for statisticians, J. Amer. Statist. Assoc. 112 (2017) 859–877.

[23] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization. http://arxiv.org/pdf/1412.6980.pdf.

[24] T. Specht, M. Nagda, S. Fellenz, S. Mandt, H. Hasse, F. Jirasek, Hanna: Hard-constraint neural network for consistent activity coefficient prediction, Chem. Sci. 15 (2024) 19777–19786.