Modified UNIFAC 2.0 – A Group-Contribution Method Completed with Machine Learning

Nicolas Hayer, Hans Hasse, and Fabian Jirasek*

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

E-mail: fabian.jirasek@rptu.de

Abstract

Predicting thermodynamic properties of mixtures is a cornerstone of chemical engineering, yet conventional group-contribution (GC) methods like modified UNIFAC (Dortmund) remain limited by incomplete parameter tables. To address this, we present modified UNIFAC 2.0, a hybrid model that integrates a matrix completion method from machine learning into the GC framework, allowing for the simultaneous training of all pair-interaction parameters, including those that cannot be fitted directly due to missing data. By training on more than 500,000 experimental data points for activity coefficients and excess enthalpies from the Dortmund Data Bank, modified UNIFAC 2.0 achieves improved accuracy, while significantly expanding the predictive scope compared to the latest published modified UNIFAC (Dortmund) version, which covers only 39% of all possible interactions. Its flexible design allows updates with new experimental data or customizations for specific applications. The new model can easily be implemented in established simulation software with complete parameter tables readily available.

1 Introduction

Understanding the thermodynamic properties of mixtures is essential for chemical engineering. Due to the impracticality of studying each relevant mixture experimentally, reliable prediction methods are crucial. Group-contribution (GC) methods offer an efficient solution by decomposing molecules into structural groups, significantly reducing the number of parameters and enabling extrapolations to unstudied components and mixtures. The most successful GC method in chemical engineering is probably UNIFAC,[1] which is available in different versions.[2–5] UNIFAC is a model for predicting the excess Gibbs energy of mixtures and derived properties, such as activity coefficients and excess enthalpies. It has been widely adopted for describing reaction and phase equilibria in mixtures and is implemented in all relevant process simulators.[6–8]

However, UNIFAC has important drawbacks: Firstly, the most comprehensive versions of UNIFAC, namely, original UNIFAC[3] and modified UNIFAC (Dortmund),[5] have been regularly updated, but only up to 2003[3] and 2016,[5] respectively. Since then, the work on UNIFAC updates has continued, but only commercially within the so-called UNIFAC-Consortium (TUC),[4] so the latest UNIFAC versions are not publicly available. Furthermore, the applicability of all UNIFAC versions, including the commercial ones, is limited by the availability of pair-interaction parameters between structural groups, cf. Figs. S.1 and S.2 in the Supporting Information. These parameters are derived from vapor-liquid equilibrium (VLE) and other thermodynamic data of mixtures, leaving substantial gaps when no suitable training data are available, severely hampering the applicability of UNIFAC – particularly for describing processes involving poorly studied or complex components and mixture. Furthermore, since alternative modeling approaches, such as COSMO-RS[9] and COSMO-SAC-dsp,[10] have broader applicability but are often less accurate than UNIFAC,[11] the lack of appropriate UNIFAC parameters can lead to significant inaccuracies in thermodynamic modeling, thereby compromising robust process design and optimization.

Compared to the original UNIFAC,[3] in which two parameters are used to describe the

interactions between a given pair of groups, modified UNIFAC[5] considers the temperature dependence of these parameters by a simple function, leading to up to six parameters that can be adjusted for a given pair of groups. This increased flexibility often improves accuracy in describing different mixtures, making modified UNIFAC (Dortmund) arguably the best GC method presently available.[5,12] For simplicity, we will label the latest public version of modified UNIFAC (Dortmund), which we use as the reference here, as mod. UNIFAC 1.0. Mod. UNIFAC 1.0 considers 63 *main groups*, subdivided into 125 *subgroups*. While each subgroup k has individual size parameters describing their surface area (Q_k) and volume (R_k), which are reported for all 125 defined subgroups, pair-interaction parameters are defined between main groups m and n. In the current parameterization of mod. UNIFAC 1.0, these interaction parameters are reported for only 39% of all possible pairs of main groups; Fig. S.1 in the Supporting Information illustrates this. This situation significantly hampers the applicability of mod. UNIFAC 1.0 since a single missing group pair-interaction parameter for a given mixture prevents the use of the method.

Consequently, the pair-interaction parameters of mod. UNIFAC 1.0, which are asymmetric ($a_{mn} \neq a_{nm}$, $b_{mn} \neq b_{nm}$, $c_{mn} \neq c_{nm}$), can be arranged in (sparsely filled) matrices, making the prediction of the missing parameters a matrix completion problem, for which matrix completion methods (MCMs) from machine learning (ML)[13,14] can be used. The basic idea of MCMs is to model the sparse matrix of interest with so-called features specific for each row and column, which are learned during the training step. This approach relies on correlated entries in the matrix. The MCM learns these correlations and captures them through the features. We have demonstrated the applicability of MCMs in thermodynamics in prior work, where we have developed MCMs to predict different thermodynamic properties of mixtures[15–19] and different types of pair-interaction parameters.[20,21] Most importantly, we have recently introduced UNIFAC 2.0,[22] a hybrid model that embeds an MCM into the framework of the original UNIFAC model.[3] Through this integration, the MCM predicts the missing pair-interaction parameters between the main groups of original UNIFAC. UNIFAC

2.0 was trained on experimental activity coefficients derived from binary VLE data and limiting activity coefficient data taken from the Dortmund Data Bank (DDB)[23] in an end-to-end manner, avoiding the sequential and often intuitive approaches that have characterized the traditional fitting process of UNIFAC. Our recent work demonstrates that the hybrid UNIFAC 2.0, based on a learned completed pair-interaction parameter table, outperforms the original UNIFAC method in terms of scope and accuracy.[22]

In this work, we transfer the concept of embedding an MCM in GC methods from original UNIFAC to mod. UNIFAC and introduce mod. UNIFAC 2.0. Similar to UNIFAC 2.0,[22] mod. UNIFAC 2.0 exhibits complete pair-interaction parameterizations and was trained end-to-end on an extensive database of more than 500,000 data points from the DDB. As the consideration of the temperature dependence of the group interactions makes mod. UNIFAC more flexible, we have included experimental data on the excess enthalpy besides data on activity coefficients in the training process of mod. UNIFAC 2.0.

By retaining the mod. UNIFAC equations, mod. UNIFAC 2.0 maintains the high accessibility of the original model and can easily be implemented in process simulators by simply replacing the parameter sets with the ones freely provided in the Supporting Information of this work. At the same time, mod. UNIFAC 2.0 eliminates the most significant limitation of the original model by filling all gaps in the pair-interaction parameter tables, tremendously increasing the applicability to any mixture whose components can be represented by the presently defined structural groups. The subgroup-specific size parameters R_k and Q_k for using mod. UNIFAC 2.0, which are identical to those of the published mod. UNIFAC 1.0 version, are also provided in the Supporting Information of this work.

2 Development of Mod. UNIFAC 2.0

2.1 General Framework

Fig. 1 illustrates how mod. UNIFAC 2.0 was developed by embedding an MCM into the mod. UNIFAC framework. The resulting method was trained end-to-end on experimental logarithmic activity coefficients ($\ln \gamma_i$) and excess enthalpies (h^{E}) in binary mixtures. Since mod. UNIFAC 2.0 is trained directly on the experimental data, it does not depend on the parameterization of mod. UNIFAC 1.0 but generates a completely new set of pair-interaction parameters. The $\ln \gamma_i$ were obtained from the limiting activity coefficient database of the DDB and derived from binary VLE data, cf. Section "Data" for details. Mod. UNIFAC 2.0 is compared here to mod. UNIFAC 1.0, which uses the same structural groups and physical model equations as mod. UNIFAC 2.0 but whose parameters were obtained by sequential parameter fitting on a data basis that includes only data taken before 2016. Additionally, mod. UNIFAC 1.0 was trained on additional mixture properties beyond those included here.[5,24]

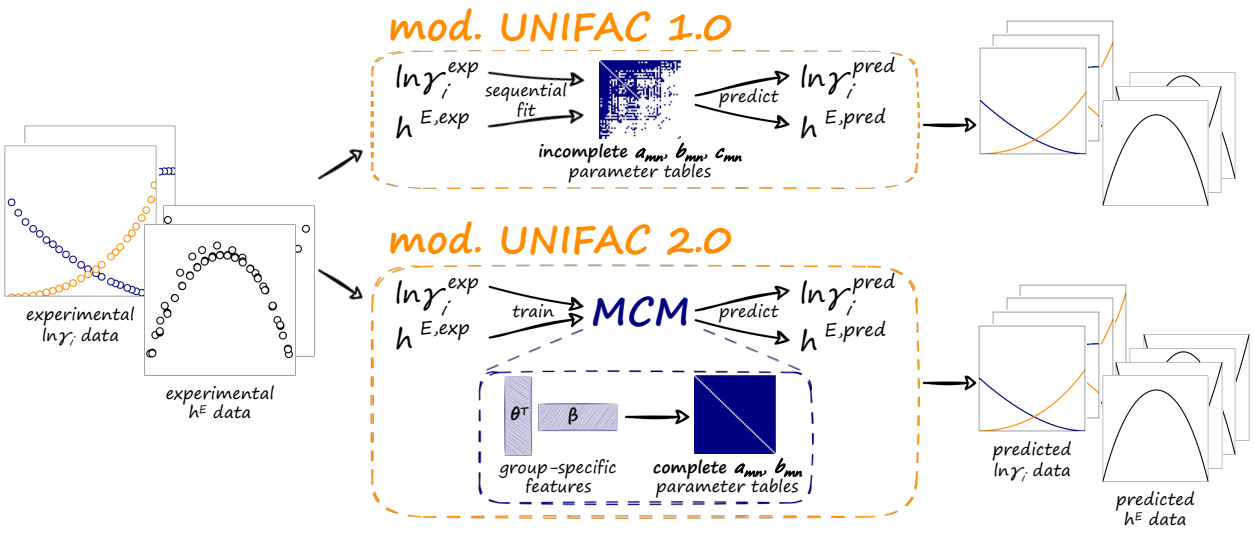

Figure 1: Comparison of mod. UNIFAC 1.0[5] and mod. UNIFAC 2.0 (this work). Mod. UNIFAC 1.0 relies on sequential parameter fitting, whereas mod. UNIFAC 2.0 integrates a matrix completion method (MCM) for predicting pair-interaction parameters into the mod. UNIFAC framework. Mod. UNIFAC 2.0 was trained end-to-end on experimental logarithmic activity coefficients ($\ln\gamma_i$) and excess enthalpy (h^{E}) data. After training, the completed pair-interaction parameter matrices facilitate predictions of thermodynamic properties for a vast range of binary and multi-component mixtures.

Mod. UNIFAC 1.0 extends the parameter Ψ_{mn} of the original UNIFAC model by introducing a temperature dependence through the additional interaction parameters b_{mn} and c_{mn}:

$$\Psi_{mn} = \exp\left(-\frac{a_{mn} + b_{mn}T + c_{mn}T^2}{T}\right) \tag{1}$$

Setting $b_{mn} = c_{mn} = 0$ results in the original UNIFAC definition of Ψ_{mn}.[3]

However, in mod. UNIFAC 1.0, c_{mn} parameters were fitted for only very few pairs of groups, and $c_{mn} = 0$ is used for most group combinations. Therefore, we have decided to only use a_{mn} and b_{mn} in mod. UNIFAC 2.0, which are modeled by two MCMs trained to decompose the two matrices containing the parameters a_{mn} and b_{mn}, respectively, into the product of two respective feature matrices. Each pair-interaction parameter is thereby

modeled as:

$$a_{mn} = C_a \cdot \boldsymbol{\theta}_m^a \cdot \boldsymbol{\beta}_n^a \tag{2}$$

$$b_{mn} = C_b \cdot \boldsymbol{\theta}_m^b \cdot \boldsymbol{\beta}_n^b \tag{3}$$

Here, $\boldsymbol{\theta}_m^a$, $\boldsymbol{\theta}_m^b$, $\boldsymbol{\beta}_n^a$, and $\boldsymbol{\beta}_n^b$ are vectors of length K, where K is called latent dimension. This hyperparameter was set to $K = 8$, following our recent work on the original UNIFAC method,[22] in which we investigated its influence on model performance and observed robust behavior across different values. Thereby, rather robust behavior towards changes in the hyperparameters could be observed. For simplicity, we collectively refer to the feature vectors as θ and β in the following. The scaling factors C_a and C_b ensure equal treatment of the two types of pair-interaction parameters. They are estimated from the mean values of a_{mn} and b_{mn} of mod. UNIFAC 1.0 and were set to $C_a = 10$ and $C_b = 0.04$.

All parameters of mod. UNIFAC 2.0 are learned *simultaneously*, which is in sharp contrast to the sequential approach used in the original model. We have trained mod. UNIFAC 2.0 within a Bayesian framework, treating each experimental data point ($\ln \gamma_i$, h^{E}), feature (θ, β), and interaction parameter (a_{mn}, b_{mn}) as random variables drawn from probability distributions. By applying Bayes' theorem, we link these variables through three key distributions: the prior, the likelihood, and the posterior.

The prior represents initial assumptions about the features before observing data. Here, the prior for all features is a standard normal distribution, $\mathcal{N}(0, 1)$, which is uninformative and introduces no bias toward specific feature values, except for discouraging very large values, thereby serving as a kind of regularization. This choice provides a simple and effective starting point for learning features from the empirical data.

The likelihood defines the probability of observing the data ($\ln \gamma_i^{\mathrm{exp}}$ and $h^{\mathrm{E,exp}}$) given the features. It is modeled using a Cauchy distribution centered around the predicted values

$\ln \gamma_i^{\text{pred}}$ and $h^{\text{E,pred}}$, respectively:

$$p(\ln \gamma_i^{\text{exp}} | \theta, \beta) = \text{Cauchy}(\ln \gamma_i^{\text{pred}}, \lambda) \tag{4}$$

$$p(h^{\text{E,exp}} | \theta, \beta) = \text{Cauchy}(h^{\text{E,pred}}, \lambda) \tag{5}$$

where λ is the scale parameter of the Cauchy distribution, which was set to $\lambda = 0.4$, consistent with our recent work on the original UNIFAC method.[22] There, we have demonstrated that the approach is rather robust towards changes in the hyperparameters, with only very small values of K or λ showing a significant deterioration in prediction accuracy. This supports the use of the same hyperparameter settings for modified UNIFAC 2.0 in this work. However, other suitable values are also possible. The Cauchy distribution's heavy tails make the model robust to outliers or flawed training data, mitigating the influence of experimental noise during training. Predicted values for $\ln \gamma_i^{\text{pred}}$ and $h^{\text{E,pred}}$ are obtained using the standard mod. UNIFAC equations, which are fully described in Refs.:[12,25]

$$\ln \gamma_i^{\text{pred}} = \text{mod. UNIFAC}(a_{mn}, b_{mn}, R_k, Q_k, \boldsymbol{x}, T) \tag{6}$$

$$h^{\text{E,pred}} = -RT^2 \sum_{i=1}^{N} x_i \left(\frac{\partial \ln \gamma_i^{\text{pred}}}{\partial T} \right)_{p,\boldsymbol{x}} \tag{7}$$

where $\boldsymbol{x}$ is the composition vector (for binary mixtures, this reduces to x_1), T is the absolute temperature, and a_{mn} and b_{mn} are the predicted pair-interaction parameters of mod. UNIFAC 2.0 calculated from the learned features according to Eqs. (2) and (3).

The goal of Bayesian inference is to find the posterior, which combines the prior and the likelihood, i.e., it encapsulates updated beliefs about the features after considering both prior information on the features and empirical data. Using Pyro, a probabilistic programming language written in Python and supported by PyTorch,[26] we have approximated the posterior using stochastic variational inference (VI) under the mean-field assumption,[27] where all features are considered independent, and a normal variational distribution approximates

each. Variational inference thereby replaces the complex true posterior with a simpler but flexible distribution and adjusts its parameters to approximate the true posterior as closely as possible. This approach ensures computational feasibility and scalability for large data sets. The quality of this approximation is measured via the evidence lower bound (ELBO). During optimization, the ELBO was maximized using the Adam optimizer[28] with a learning rate of 0.15.

The result of training mod. UNIFAC 2.0 is a learned probability density for each feature, from which we used the means to calculate the final pair-interaction parameters (cf. Eqs. (2) and (3)), which are subsequently plugged into the mod. UNIFAC equations[12,25] to give predictions for unstudied activity coefficients.

We have made the complete final set of pair-interaction parameters – derived from training mod. UNIFAC 2.0 on the entire database (see Section "Data") – freely available in the Supporting Information as Excel files. We also provide the learned group-specific features, which are normal distributions and defined by their mean and standard deviation. Additionally, we provide the subgroup-specific size parameters R_k and Q_k, which are identical to the published mod. UNIFAC 1.0 version.[5]

2.2 Data

Experimental data for activity coefficients γ_i and excess enthalpies h^{E} in binary mixtures were used for training mod. UNIFAC 2.0. All data were taken from the most extensive database for thermodynamic properties, the DDB.[6] During preprocessing, data points that were considered to be of low quality by the DDB were excluded, corresponding to only 0.8% of all raw data points. We restricted our selection to binary mixtures whose components could be decomposed into the at present publicly available mod. UNIFAC subgroups, without excluding any component for other reasons, e.g., based on their chemical family. Additionally, the VLE data were limited to pressures up to 10 bar.

After preprocessing, the h^{E} data set comprises 259,707 data points for 8,735 binary

mixtures. The data set for γ_i consists of 243,257 data points for 21,452 binary mixtures, which was obtained by combining 68,642 data points for limiting activity coefficients and 174,615 data points calculated from VLE data using the extended Raoult's law assuming an ideal gas phase[1] and neglecting the pressure dependence of the chemical potential in the liquid phase:

$$\gamma_i(T, \boldsymbol{x}) = \frac{p \cdot y_i}{p_i^{\mathrm{s}}(T) \cdot x_i} \tag{8}$$

Here, p corresponds to the total pressure and p_i^{s} to the vapor pressure of the pure component i, while x_i and y_i are the mole fractions of component i in the liquid and vapor phases, respectively.

3 Results and Discussion

3.1 Overall Performance of Mod. UNIFAC 2.0

For evaluating the performance of mod. UNIFAC 2.0 in predicting activity coefficients $\ln \gamma_i$ and excess enthalpies h^{E}, we use the mean absolute error (MAE) for each binary mixture and represent the results in box plots, as shown in Figs. 2 (for $\ln \gamma_i$) and 3 (for h^{E}). These plots also contain the corresponding results of mod. UNIFAC 1.0, evaluated on the same basis, for comparison. The results shown in these figures were obtained with a mod. UNIFAC 2.0 version trained on all available experimental data in our database. However, as detailed in the subsequent subsections, we have also performed two extrapolation tests by withholding parts of the data during the training to demonstrate and validate the predictive capacities of mod. UNIFAC 2.0.

Although the exact training set for mod. UNIFAC 1.0 has not been disclosed, it is reasonable to assume that the experimental data used in this work are similar to the data

[1] At elevated pressures of up to 10 bar, some deviations from this assumption have to be expected. The 10 bar limit was chosen as a compromise between limiting these deviations and losing interesting systems from the database. We have refrained from including fugacity coefficients to correct for the non-ideality of the gas phase for computational reasons.

used for its parameterization, which supports a fair comparison in Figs. 2 and 3. Note that the comparison between mod. UNIFAC 1.0 and mod. UNIFAC 2.0 is carried out on the "mod. UNIFAC 1.0 horizon", i.e., only those mixtures from our data set that can be modeled with the incomplete parameter set of mod. UNIFAC 1.0. Since mod. UNIFAC 2.0, with its completed parameter set, has a much larger scope, its performance is additionally evaluated on those mixtures that cannot be predicted with mod. UNIFAC 1.0, labeled as the "mod. UNIFAC 2.0 only" data set in Figs. 2 and 3.

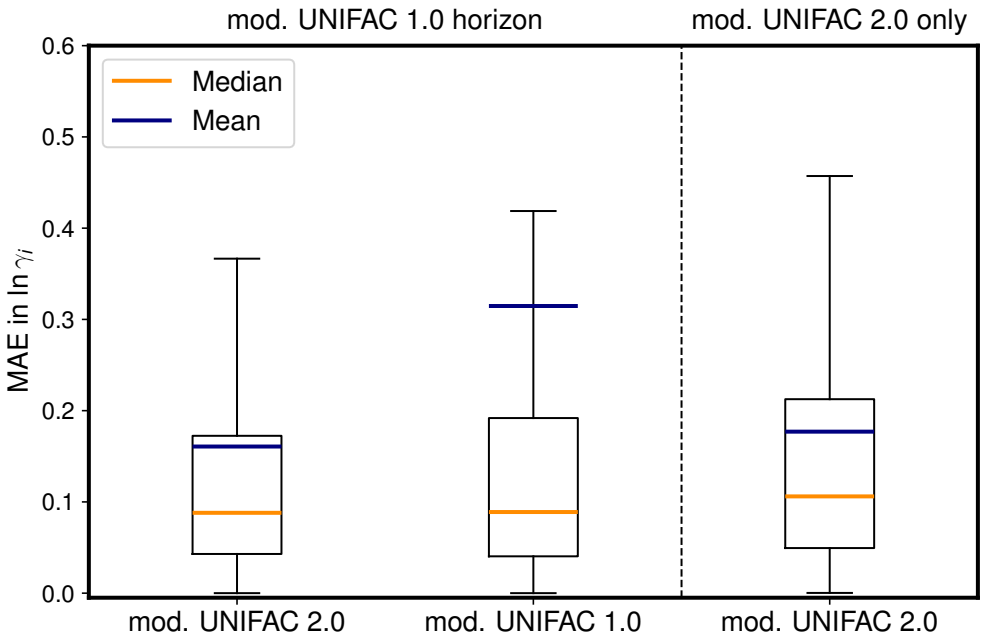

Figure 2: Mean absolute error (MAE) of the predicted $\ln \gamma_i$ with mod. UNIFAC 2.0 and comparison to mod. UNIFAC 1.0 for those mixtures that can also be predicted by the latter model ("mod. UNIFAC 1.0 horizon"). The "mod. UNIFAC 1.0 horizon" comprises 221,639 data points for 16,932 binary mixtures, while an additional 21,618 experimental data points for 4,520 binary mixtures could only be predicted with mod. UNIFAC 2.0 ("mod. UNIFAC 2.0 only"). The boxes represent the interquartile ranges (IQR), and the whiskers extend to the last data points within 1.5 times the IQR from the box edges.

The results in Fig. 2 show an improved prediction accuracy for $\ln \gamma_i$ with mod. UNIFAC 2.0 compared to mod. UNIFAC 1.0 for those mixtures that can be described with both models ("mod. UNIFAC 1.0 horizon"). This is particularly evident concerning the mean of the MAE, which is nearly halved with mod. UNIFAC 2.0, demonstrating the ability of

mod. UNIFAC 2.0 to reduce very poorly predicted data points. Regarding the median of the MAE and the interquartile range, mod. UNIFAC 2.0 also shows some improvements compared to mod. UNIFAC 1.0.

These results indicate that using the holistic end-to-end training of mod. UNIFAC 2.0 results in an improved set of pair-interaction parameters compared to the one obtained by the classical sequential fit carried out in the development of mod. UNIFAC 1.0. However, the even more significant advantage of mod. UNIFAC 2.0 is that its parameter set is complete, leading to a much broader applicability. By evaluating the results of mod. UNIFAC 2.0 for those mixtures in our data set that cannot be modeled with mod. UNIFAC 1.0 ("mod. UNIFAC 2.0 only") in Fig. 2, we find a high prediction accuracy. It is similar to that of the results obtained with mod. UNIFAC 1.0 for the mixtures to which this method can be applied.

Fig. 3 shows the results for the prediction of h^{E}, where we see a similar picture as for the prediction of $\ln \gamma_i$: we find an improved performance of mod. UNIFAC 2.0 on the "mod. UNIFAC 1.0 horizon", and still high prediction accuracy on the "mod. UNIFAC 2.0 only" data set, for which mod. UNIFAC 1.0 cannot be applied. Similar results are observed when using mean squared error (MSE) instead of the MAE, as shown in Figs. S.4 and S.5 in the Supporting Information. Furthermore, we have evaluated the performance of mod. UNIFAC 1.0 and mod. UNIFAC 2.0 for those mixtures for which mod. UNIFAC 1.0 uses c_{mn} pair-interaction parameters whereas in mod. UNIFAC 2.0, all c_{mn} were set to zero. The results are shown in Figs. S.6 and S.7 in the Supporting Information. Even in this case, mod. UNIFAC 2.0 shows superior performance, underscoring the validity to exclude c_{mn} from the hybrid model.

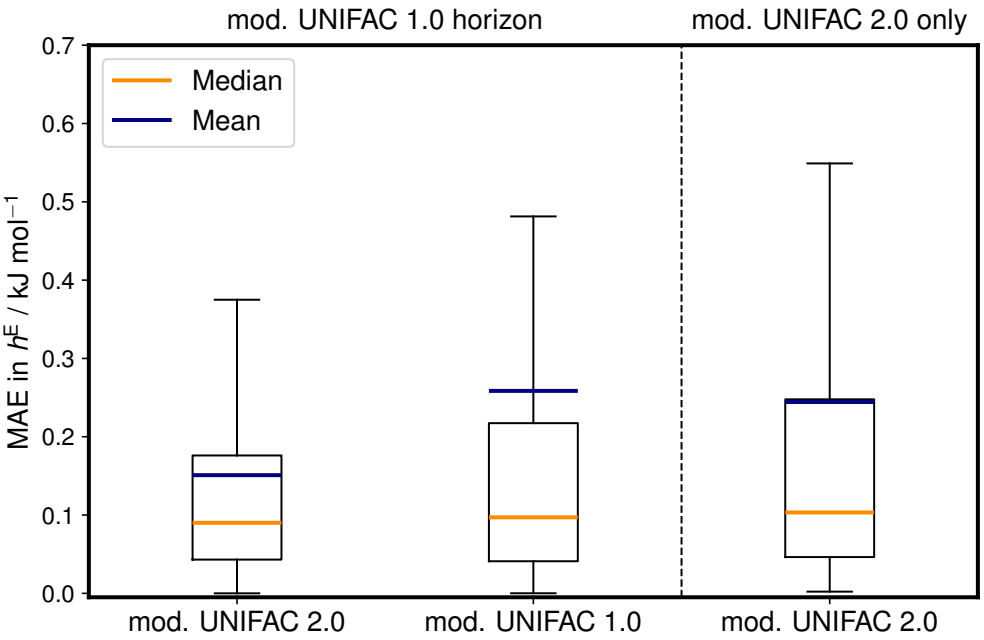

Figure 3: Mean absolute error (MAE) of the predicted h^{E} with mod. UNIFAC 2.0 and comparison to mod. UNIFAC 1.0 for those mixtures that can also be predicted by the latter model ("mod. UNIFAC 1.0 horizon"). The "mod. UNIFAC 1.0 horizon" comprises 239,770 data points for 7,776 binary mixtures, while an additional 19,937 experimental data points for 959 binary mixtures could only be predicted with mod. UNIFAC 2.0 ("mod. UNIFAC 2.0 only"). The boxes represent the interquartile ranges (IQR), and the whiskers extend to the last data points within 1.5 times the IQR from the box edges.

Fig. 4 provides a deeper insight into the overall performance of mod. UNIFAC 2.0 by assigning an MAE for predicting $\ln \gamma_i$ to each pair of main groups, visualized as heatmaps. The shown MAEs are calculated by considering the predictions for all mixtures for which the respective group combination is relevant, with the number of mixtures and data points varying significantly among the pairs of main groups, as detailed in Fig. S.1b of the Supporting Information. Panel (a) of Fig. 4 shows the MAEs calculated as described above on our complete data set, while panel (b) visualizes improvements (or deteriorations) with mod. UNIFAC 2.0 compared to mod. UNIFAC 1.0 by showing the differences in the MAEs ($\Delta\mathrm{MAE} = \mathrm{MAE}_{\text{mod. UNIFAC 2.0}} - \mathrm{MAE}_{\text{mod. UNIFAC 1.0}}$) on the "mod. UNIFAC 1.0 horizon". Missing entries indicate that no data were available to compare the given combination of groups.

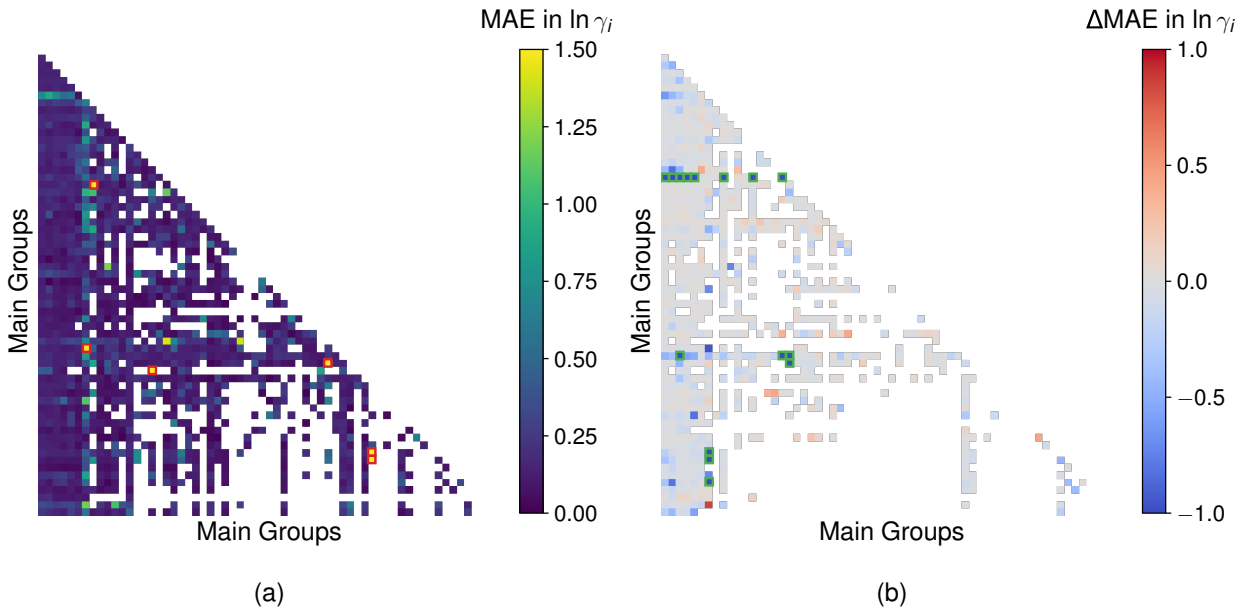

Figure 4: (a) Heatmap of the mean absolute error (MAE) of the predicted $\ln \gamma_i$ with mod. UNIFAC 2.0 calculated for each pair of main groups by considering all data points for which that particular group combination is relevant. Group combinations with an MAE above 1.5 are highlighted by red frames. (b) Difference between the MAE in $\ln \gamma_i$ with mod. UNIFAC 2.0 and the MAE of mod. UNIFAC 1.0 on the "mod UNIFAC 1.0 horizon" (ΔMAE = $\text{MAE}_{\text{mod. UNIFAC 2.0}} - \text{MAE}_{\text{mod. UNIFAC 1.0}}$) for each pair of main groups. Group combinations with a ΔMAE below -1 are highlighted by green frames, indicating the most significant improvements with mod. UNIFAC 2.0. Missing entries indicate that no data were available for the comparison of the given combination of groups.

Fig. 4a highlights the overall strong performance of mod. UNIFAC 2.0, with a small MAE for most group combinations. Note that the prediction for a particular mixture usually requires the consideration of multiple pair-interaction parameters. Hence, the MAEs in Fig. 4a, although assigned to specific group combinations, cannot be attributed to imperfections of the respective pair-interaction parameters alone, but are also affected by other pair-interaction parameters. However, despite this complexity, a clear trend can be observed. For instance, mixtures containing water (main group 7) apparently represent a particular challenge, likely because of the unique properties of water due to strong hydrogen bonding and polarity. While most group combinations yield an MAE below 0.14, which is a very good result, a few show high prediction errors. The three group combinations with an MAE greater than 2.0 are cases based on extremely small test sets, each consisting of a single

binary mixture with ten or fewer data points. This suggests that the higher errors may also be due to limited data, and caution should be taken not to over-interpret these results.

Fig. 4b shows that mod. UNIFAC 2.0 outperforms mod. UNIFAC 1.0 for most group combinations. It significantly improves the results for 461 interaction parameters, with a mean ΔMAE of -0.31, whereas for the 267 combinations mod. UNIFAC 1.0 yields better results; however, the deterioration is typically only minor with a mean ΔMAE of only 0.05. Notable improvements are observed for parameters involving main groups 7 ("H2O"), 18 ("PYRIDINE"), and 42 ("CY-CH2"), with ten group combinations showing extremely high MAE reductions with $\Delta\text{MAE} < -4$. In addition, parameters involving the most common group, main group 1 ("CH3"), also show significant improvements. For example, the mean MAE specific for the pair-interaction parameter between main group 1 ("CH3") and main group 7 ("H2O"), known to be poorly fitted in mod. UNIFAC 1.0, is nearly halved, from 1.35 to 0.71.

Fig. 5 shows an example of the practical application of mod. UNIFAC 2.0. It is used to predict vapor-liquid phase equilibria for binary mixtures, a critical task in chemical engineering. Six typical examples are shown, covering a range of phase behaviors from near-ideal mixtures to those with significant deviations, including low- and high-boiling azeotropes. All shown mixtures are part of the "mod. UNIFAC 2.0 only" set, i.e., they cannot be modeled with mod. UNIFAC 1.0. These examples illustrate the generally close agreement between predictions and experimental data observed across most mixtures, highlighting the method's ability to capture diverse phase behavior. While some mixtures exhibit larger deviations, the overall performance makes mod. UNIFAC 2.0 a valuable tool for a wide range of industrial processes, from distillation to solvent recovery.

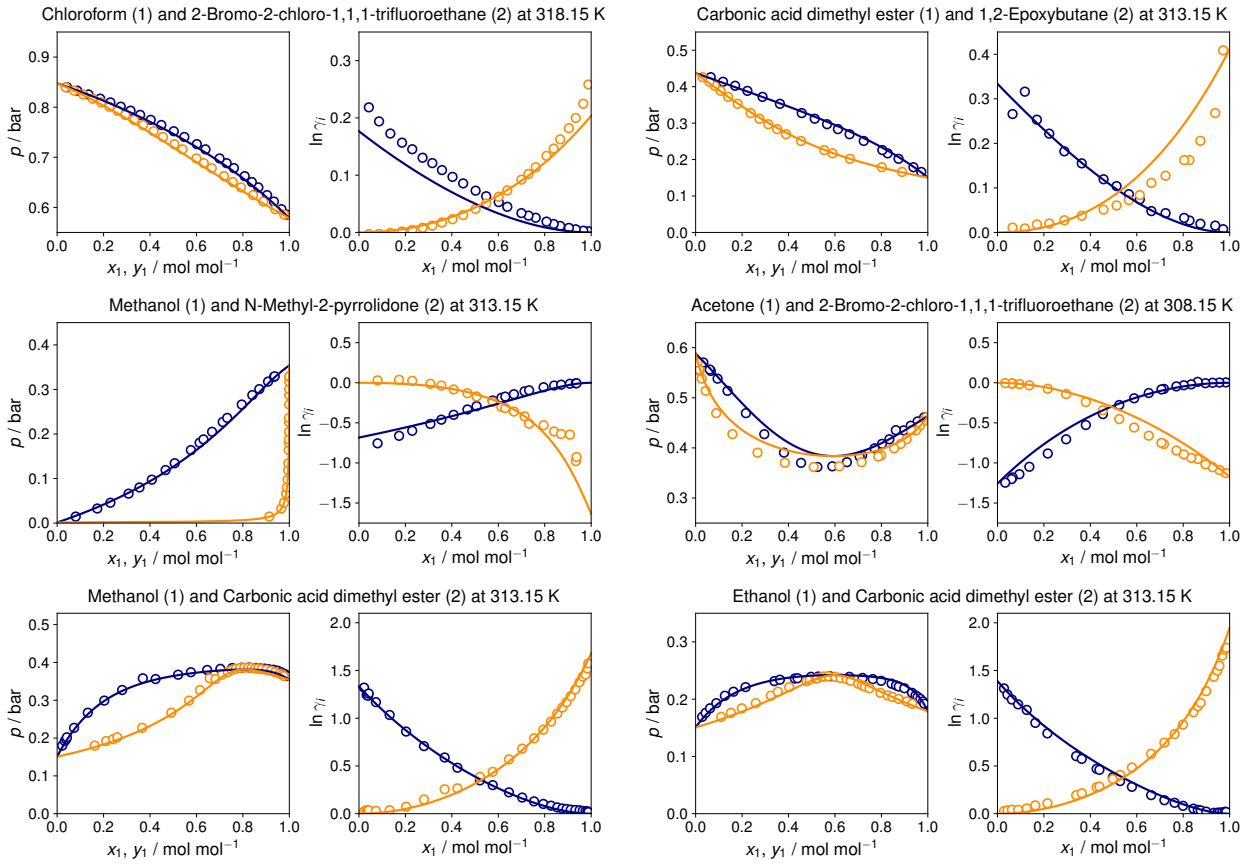

Figure 5: Prediction of $\ln \gamma_i$ and isothermal vapor–liquid phase diagrams for binary mixtures with mod. UNIFAC 2.0 (lines) and comparison to experimental data from the DDB (symbols). Mod. UNIFAC 1.0 is not applicable to the mixtures shown.

Fig. 6 demonstrates the ability of mod. UNIFAC 2.0 to predict excess enthalpies h^{E} in binary mixtures. The figure presents six representative examples from the "mod. UNIFAC 2.0 only" data set, i.e., mixtures for which mod. UNIFAC 1.0 is not applicable, cf. Fig. 3. The mixtures have been selected to highlight a variety of behaviors, ranging from nearly ideal to strongly non-ideal systems with both positive and negative deviations. The predicted excess enthalpy curves (solid lines) align closely with the experimental data (open circles), demonstrating the model's ability to accurately capture both the magnitude and the trend of h^{E} across different mixtures.

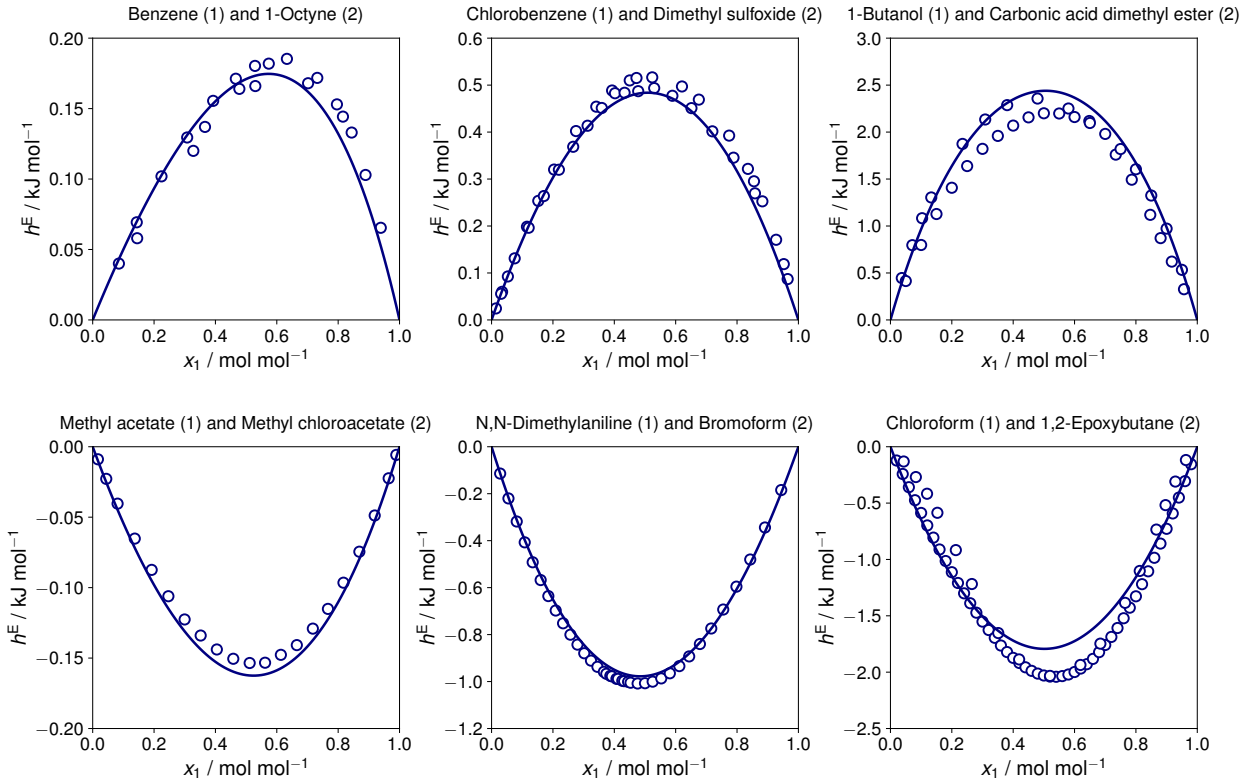

Figure 6: Prediction of excess enthalpies h^{E} at 298.15 K for binary mixtures with mod. UNIFAC 2.0 (lines) and comparison to experimental data from the DDB (symbols). Mod. UNIFAC 1.0 is not applicable to the mixtures shown.

Furthermore, since mod. UNIFAC 2.0 is based on pairwise interactions between the structural groups occurring in the mixture, for any number of components, it allows for straightforward predictions of the properties of multi-component mixtures. In Fig. S.3 in the Supporting Information, we show examples for modeling ternary mixtures with mod. UNIFAC 2.0, which demonstrate its high predictive performance although being trained only on binary data.

3.2 Extrapolation to Unseen Components

To study the ability of mod. UNIFAC 2.0 to extrapolate to mixtures involving components for which no mixture data were used in the training (termed "unseen components" in the following for simplicity), we have carried out a test in which 100 components were selected,

and all data points containing any of these components were withheld from the training set and used only for testing the predictions. This exclusion resulted in a test set comprising 34,107 data points (20,912 for $\ln \gamma_i$ and 13,195 for h^{E}), covering 1,865 different binary mixtures. Fig. 7 shows the results for the prediction of $\ln \gamma_i$ for this test set, again represented as box plots of the mixture-specific MAE. Results from mod. UNIFAC 1.0 are also shown for comparison. Additional details on component-wise error scores are provided in Tables S.1 and S.2 in the Supporting Information.

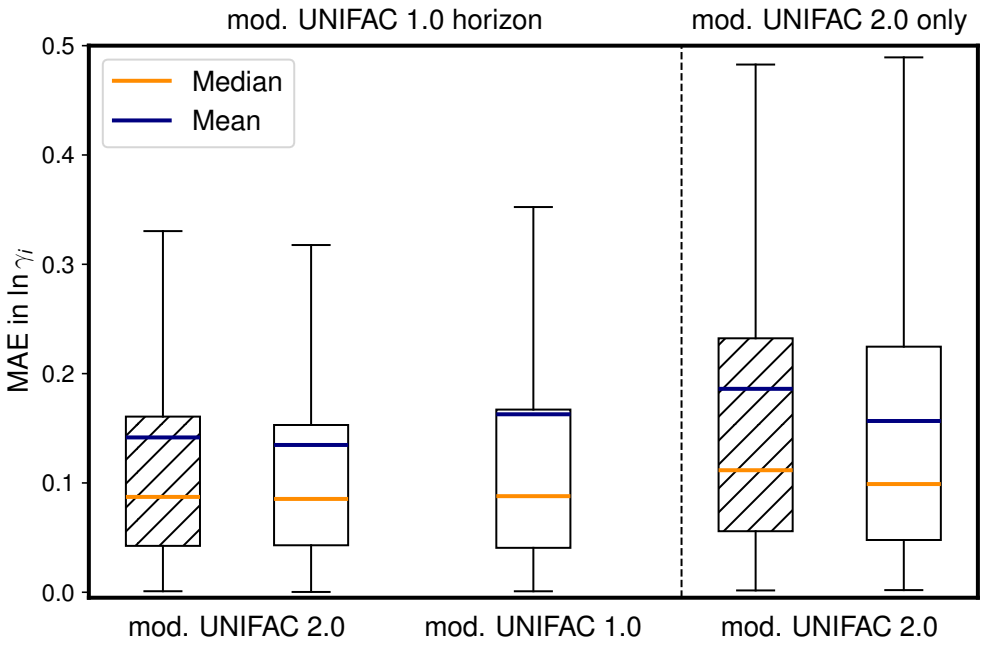

Figure 7: Mean absolute error (MAE) of the predicted $\ln \gamma_i$ of mixtures containing unseen components with mod. UNIFAC 2.0 (shaded boxes). For comparison, the results of mod. UNIFAC 2.0 trained on all experimental data and mod. UNIFAC 1.0 are also shown (plain boxes). The "mod. UNIFAC 1.0 horizon" comprises 19,015 data points for 1,254 binary mixtures, while an additional 1,897 experimental data points for 280 binary mixtures could only be predicted with mod. UNIFAC 2.0 ("mod. UNIFAC 2.0 only"). The boxes represent the interquartile ranges (IQR), and the whiskers extend to the last data points within 1.5 times the IQR from the box edges.

Fig. 7 shows that the predictive accuracy of mod. UNIFAC 2.0 for mixtures with components that were excluded from the training set (shaded boxes) is only narrowly lower than when the model is trained on the entire database (plain boxes). This consistency across

both the "mod. UNIFAC 1.0 horizon" and "mod. UNIFAC 2.0 only" data sets underscores the robustness of the hybrid approach. Moreover, even on the test data, mod. UNIFAC 2.0 outperforms mod. UNIFAC 1.0 on the "mod. UNIFAC 1.0 horizon", which is noteworthy given that mod. UNIFAC 1.0 was likely trained on many of these test data points, as discussed earlier. On the "mod. UNIFAC 2.0 only" data set, mod. UNIFAC 2.0 shows slightly reduced predictive accuracy but still maintains strong performance, while mod. UNIFAC 1.0 is not applicable. Overall, these results highlight the predictive power of mod. UNIFAC 2.0. Similar trends were observed for the prediction of h^{E}, as shown in Fig. S.8 in the Supporting Information.

3.3 Extrapolation to Unseen Pair-Interaction Parameters

Another, even more challenging, test to assess mod. UNIFAC 2.0's predictive capacities is to test its ability to extrapolate to unseen pair interactions. For such a test, we have randomly selected 100 combinations of main groups, and have withheld all experimental data for mixtures for which the respective main group combinations are relevant from the training. For each of these 100 combinations, an individual test set was created from the withheld data, while all other available data were used to train mod. UNIFAC 2.0. The selected combinations of main groups are shown in Fig. S.9 in the Supporting Information. The number of data points and binary mixtures for the 100 test sets, as well as individual error scores, are given in Tables S.3 (for $\ln \gamma_i$) and S.4 (for h^{E}).

Fig. 8 shows the results of predicting $\ln \gamma_i$ with mod. UNIFAC 2.0 from this challenging test by summarizing the MAEs for the 100 test sets in a box plot. For comparison, the performance of mod. UNIFAC 2.0 trained on all experimental data and the results of mod. UNIFAC 1.0 (on the "mod. UNIFAC 1.0 horizon") are included. Similar results were obtained for h^{E} and are summarized in Fig. S.10 in the Supporting Information.

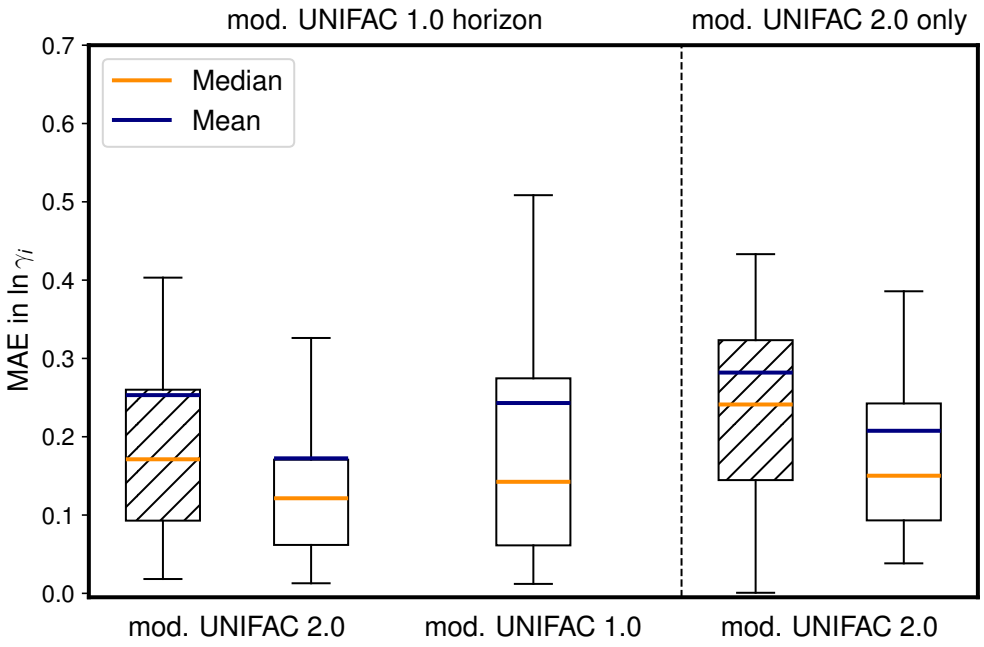

Figure 8: Mean absolute error (MAE) of the predicted $\ln \gamma_i$ with mod. UNIFAC 2.0 for 100 test sets, where all data points for which a specific main group combination is relevant were withheld during training (shaded boxes); cf. Table S.3 in the Supporting Information for numerical results. The results of mod. UNIFAC 2.0 trained on all experimental data and mod. UNIFAC 1.0 are shown for comparison (plain boxes). The boxes represent the interquartile ranges (IQR), and the whiskers extend to the last data points within 1.5 times the IQR from the box edges.

The results on the "mod. UNIFAC 1.0 horizon" demonstrate that mod. UNIFAC 2.0, even when predicting truly unseen pair-interaction parameters, which could not directly be fitted to the training data, achieves a performance comparable to mod. UNIFAC 1.0 with parameters that were likely fitted directly to the respective experimental data. Comparing mod. UNIFAC 2.0's predictions for unseen pair interactions (shaded boxes) with those trained on the entire database (plain boxes) reveals a decrease in accuracy, as expected. However, the differences are modest, highlighting the robustness and reliability of mod. UNIFAC 2.0 even in this extremely challenging test.

These tests suggest the potential of mod. UNIFAC 2.0 not only to broaden the applicability of this group-contribution method but also to improve its prediction accuracy significantly. Unlike mod. UNIFAC 1.0, which is constrained by its limited parameter tables

obtained from sequential fitting, mod. UNIFAC 2.0 excels in both scope and performance, making it a robust tool for predicting activity coefficients across a wide range of mixtures. The superior accuracy demonstrated on the shared horizon confirms that mod. UNIFAC 2.0 is not just a complementary option when mod. UNIFAC 1.0 fails but a strong candidate to become the new standard.

Its ease of implementation sets mod. UNIFAC 2.0 apart from other ML-based or hybrid models combining ML with physical modeling. Users can seamlessly adopt mod. UNIFAC 2.0 by simply replacing the original parameter tables in their existing process simulators (or similar software), in which mod. UNIFAC will most likely be implemented, with the completed parameter tables provided in the Supporting Information. This way, the tedious implementation of the ML model itself is eliminated, making mod. UNIFAC 2.0 directly accessible for practical applications.

4 Conclusions

Mod. UNIFAC[5] is currently the industrial standard for predicting activity coefficients and is implemented in basically all process simulation software packages. It is also widely used in academia and is the workhorse for calculating phase equilibria with liquid phases, such as vapor-liquid equilibria (VLE), liquid-liquid equilibria (LLE), and solid-liquid equilibria (SLE). Due to temperature-dependent parameters, it is more flexible than the original UNIFAC model[3] and often delivers better results.

However, mod. UNIFAC has several important drawbacks. Firstly, the last published version is from 2016 and has therefore been fitted only to data that were available up to then. Hence, the wealth of relevant data measured since then is not included. More recent updates of mod. UNIFAC are commercial and not publicly available. Secondly, and more importantly, as a group-contribution method, mod. UNIFAC can only be applied to make predictions for a given mixture if all pair-interaction parameters (between all groups into which all components of the mixture are decomposed) are available. If only a single pair is missing, mod. UNIFAC will not work. The latest public version of mod. UNIFAC has 63 main groups, and, hence, 1953 pairs of groups – but interaction parameters are only available for 756 of these pairs (39%), which considerably limits the method's applicability.

We have therefore developed mod. UNIFAC 2.0, which overcomes these drawbacks: It was trained on data for activity coefficients and excess enthalpies published up to 2024 that were taken from the Dortmund Data Bank (DDB). All in all, more than 500,000 data points from 27,035 binary systems were used. The equations and the groups used in mod. UNIFAC 2.0 are exactly the same as in the last published version, called mod. UNIFAC 1.0 here, but the training differs drastically. While the parameters of mod. UNIFAC 1.0 were determined in a sequential approach, without a chance to fill gaps for interactions for which no relevant data were available, mod. UNIFAC 2.0 is trained using a matrix completion method (MCM) by which the entire interaction parameter matrix is filled simultaneously. Consequently, there are no gaps in the mod. UNIFAC 2.0 parameter tables. This leads to an important extension

of the applicability of the method. However, not only was the applicability extended, but the accuracy of the predictions was also improved. This was demonstrated in tests in which mod. UNIFAC 2.0 was compared to mod. UNIFAC 1.0: In different studies, data were deliberately excluded from the training and only used for the tests. Even in these tests, mod. UNIFAC 2.0 performed consistently better than mod. UNIFAC 1.0, even though they favor mod. UNIFAC 1.0, as it must be assumed that relevant parts of the test set were used in its training. In-depth studies also reveal significant improvements for technically important classes of mixtures, such as mixtures containing water.

As a method based on the physical concept of pair interactions, mod. UNIFAC 2.0 can be used to predict thermodynamic properties not only for binary mixtures but also for multi-component mixtures. The new model can be seamlessly integrated into existing workflows, as users only need to update the parameter tables in existing implementations. Ultimately, the end-to-end training process of mod. UNIFAC 2.0 allows for straightforward updates as new experimental data become available or for tailoring the model to specific industrial needs. Mod. UNIFAC 2.0 demonstrates how combining machine learning with established physical models can significantly enhance the prediction of thermodynamic properties. Its expanded scope, improved accuracy, and ease of implementation represent a powerful and scalable solution for modern chemical engineering challenges. The complete parameter tables are freely provided in the Supporting Information as Excel files. We recommend using mod. UNIFAC 2.0 as the default in all applications where, up to now, the default was mod. UNIFAC 1.0.

Supporting Information Available

The Supporting Information contains the following:

- Parameterization details for the public and commercial versions of modified UNIFAC (Dortmund), and further results from modified UNIFAC 2.0 for the extrapolation to unseen components, pair-interaction parameters, and multi-component mixtures (PDF).
- Subgroup-specific size parameters R_k and Q_k for using modified UNIFAC 2.0 (XLSX).
- Complete set of pair-interaction parameters, i.e., the filled matrices a_{mn} and b_{mn}, for all main group combinations derived from training modified UNIFAC 2.0 on all experimental data for direct user application (XLSX).
- Complete set of features for all main groups derived from training modified UNIFAC 2.0 on all experimental data. The features are represented as normal distributions defined by their mean and standard deviation (XLSX).

Acknowledgement

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.

Literature Cited

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

(2) Magnussen, T.; Rasmussen, P.; Fredenslund, A. UNIFAC parameter table for prediction of liquid-liquid equilibriums. *Industrial & Engineering Chemistry Process Design and Development* **1981**, *20*, 331–339.

(3) Wittig, R.; Lohmann, J.; Gmehling, J. Vapor−Liquid Equilibria by UNIFAC Group Contribution. 6. Revision and Extension. *Industrial & Engineering Chemistry Research* **2003**, *42*, 183–188.

(4) The UNIFAC Consortium. 2023; `http://www.unifac.org`.

(5) Constantinescu, D.; Gmehling, J. Further Development of Modified UNIFAC (Dortmund): Revision and Extension 6. *Journal of Chemical & Engineering Data* **2016**, *61*, 2738–2748.

(6) DDBST - Dortmund Data Bank Software & Separation Technology GmbH Dortmund Data Bank. 2024; `www.ddbst.com`.

(7) Chemstations, Inc. CHEMCAD V.8. 2024; `www.chemstations.com`.

(8) Aspen Technology, Inc. Aspen Plus V14.5. 2024; `www.aspentech.com`.

(9) Klamt, A. *COSMO-RS: From quantum chemistry to fluid phase thermodynamics and drug design*, 1st ed.; Elsevier: Amsterdam, 2005.

(10) Bell, I. H.; Mickoleit, E.; Hsieh, C.-M.; Lin, S.-T.; Vrabec, J.; Breitkopf, C.; Jäger, A. A Benchmark Open-Source Implementation of COSMO-SAC. *Journal of Chemical Theory and Computation* **2020**, *16*, 2635–2646.

(11) Xue, Z.; Mu, T.; Gmehling, J. Comparison of the a priori COSMO-RS models and group contribution methods: original UNIFAC, modified UNIFAC (Do), and modified UNIFAC (Do) consortium. *Industrial & engineering chemistry research* **2012**, *51*, 11809–11817.

(12) Gmehling, J.; Li, J.; Schiller, M. A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties. *Industrial & Engineering Chemistry Research* **1993**, *32*, 178–193.

(13) A. Ramlatchan; M. Yang; Q. Liu; M. Li; J. Wang; Y. Li A survey of matrix completion methods for recommendation systems. *Big Data Mining and Analytics* **2018**, *1*, 308–323.

(14) Koren, Y.; Bell, R.; Volinsky, C. Matrix Factorization Techniques for Recommender Systems. *Computer* **2009**, *42*, 30–37.

(15) Jirasek, F.; Bamler, R.; Mandt, S. Hybridizing physical and data-driven prediction methods for physicochemical properties. *Chemical Communications* **2020**, *56*, 12407–12410.

(16) Jirasek, F.; Alves, R. A. S.; Damay, J.; Vandermeulen, R. A.; Bamler, R.; Bortz, M.; Mandt, S.; Kloft, M.; Hasse, H. Machine Learning in Thermodynamics: Prediction of Activity Coefficients by Matrix Completion. *The journal of physical chemistry letters* **2020**, *11*, 981–985.

(17) Damay, J.; Jirasek, F.; Kloft, M.; Bortz, M.; Hasse, H. Predicting Activity Coefficients at Infinite Dilution for Varying Temperatures by Matrix Completion. *Industrial & Engineering Chemistry Research* **2021**, *60*, 14564–14578.

(18) Hayer, N.; Jirasek, F.; Hasse, H. Prediction of Henry's law constants by matrix completion. *AIChE Journal* **2022**, *68*, e17753.

(19) Großmann, O.; Bellaire, D.; Hayer, N.; Jirasek, F.; Hasse, H. Database for liquid phase diffusion coefficients at infinite dilution at 298 K and matrix completion methods for their prediction. *Digital Discovery* **2022**, *1*, 886–897.

(20) Jirasek, F.; Bamler, R.; Fellenz, S.; Bortz, M.; Kloft, M.; Mandt, S.; Hasse, H. Making thermodynamic models of mixtures predictive by machine learning: matrix completion of pair interactions. *Chemical science* **2022**, *13*, 4854–4862.

(21) Jirasek, F.; Hayer, N.; Abbas, R.; Schmid, B.; Hasse, H. Prediction of parameters of group contribution models of mixtures by matrix completion. *Physical chemistry chemical physics : PCCP* **2023**, *25*, 1054–1062.

(22) Hayer, N.; Wendel, T.; Mandt, S.; Hasse, H.; Jirasek, F. Advancing thermodynamic group-contribution methods by machine learning: UNIFAC 2.0. *Chemical Engineering Journal* **2024**, *504*, 158667.

(23) DDBST - Dortmund Data Bank Software & Separation Technology GmbH Dortmund Data Bank. 2023; `www.ddbst.com`.

(24) Gmehling, J.; Kolbe, B.; Kleiber, M.; Rarey, J. R. *Chemical thermodynamics for process simulation*; Wiley-VCH-Verl.: Weinheim, 2012.

(25) Weidlich, U.; Gmehling, J. A modified UNIFAC model. 1. Prediction of VLE, hE, and γ^{∞}. *Industrial & Engineering Chemistry Research* **1987**, *26*, 1372–1381.

(26) Bingham, E.; Chen, J. P.; Jankowiak, M.; Obermeyer, F.; Pradhan, N.; Karaletsos, T.; Singh, R.; Szerlip, Paul and Horsfall, Paul; Goodman, N. D. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research* **2019**, *20*, 1–6.

(27) Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* **2017**, *112*, 859–877.

(28) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. arXiv preprint, http://arxiv.org/pdf/1412.6980.pdf.