PREDICTION OF PARAMETERS OF GROUP CONTRIBUTION MODELS OF MIXTURES BY MATRIX COMPLETION *

Fabian Jirasek
Laboratory of Engineering Thermodynamics (LTD)
TU Kaiserslautern
Kaiserslautern, Germany
fabian.jirasek@mv.uni-kl.de

Nicolas Hayer
Laboratory of Engineering Thermodynamics (LTD)
TU Kaiserslautern
Kaiserslautern, Germany

Rima Abbas
DDBST GmbH
Oldenburg, Germany

Bastian Schmid
DDBST GmbH
Oldenburg, Germany

Hans Hasse
Laboratory of Engineering Thermodynamics (LTD)
TU Kaiserslautern
Kaiserslautern, Germany

ABSTRACT

Group contribution (GC) methods are widely used for predicting the thermodynamic properties of mixtures. They divide components into structural groups, which can be combined freely so that the applicability of a GC method is only limited by the availability of its parameters for the groups of interest. For describing mixtures, pairwise interaction parameters between the groups are of prime importance. Finding suitable numbers for these parameters is a demanding task and is often impeded by the lack of suitable experimental data. We address this problem in the present work by using matrix completion methods (MCMs) from machine learning to predict missing group-interaction parameters. This new approach is elaborated by applying it to a practically highly relevant case, the group contribution method UNIFAC for predicting activity coefficients in mixtures. An MCM is developed that yields the complete set of parameters for the first 50 main groups of UNIFAC, which substantially extends the scope and applicability of UNIFAC. The quality of the predicted parameter set is evaluated using vapor-liquid equilibrium data of binary mixtures from the Dortmund Data Bank: our approach yields comparable prediction accuracies as UNIFAC for data sets to which UNIFAC was fitted, and only slightly lower accuracies for data sets to which UNIFAC is not applicable.

1 Introduction

Methods for predicting thermodynamic properties are of paramount importance in chemical engineering, simply because there are too many relevant substances to study them all in experiments. The scale of this problem soars when going from pure components to mixtures, for simple combinatorial reasons. Also methodologically, predicting properties of mixtures is a demanding task. It can be tackled basically from two sides: on the one hand, one can look for similarities between substances (which is basically a data-driven approach), on the other hand, one can try to base predictions on physical theory.

The most successful methods in the field combine these two aspects. Among these, methods that rely on the concept of *group contributions* (GC) play an important role. They are based on the idea that components can be characterized by the *structural groups* they contain and take advantage of the fact that the number of relevant structural groups is many orders of magnitude smaller than the number of relevant components. As a consequence, GC methods can be used for describing a very large number of components based on a relatively small number of *group-specific parameters*: any component that can be built from groups, for which parameters are available, can be modeled.

* *Citation*: F. Jirasek, N. Hayer,R. Rima Abbas, B. Schmid, H. Hasse: Prediction of Parameters of Group Contribution Models of Mixtures by Matrix Completion, Physical Chemistry Chemical Physics Science 25 (2023) 1054-1062. DOI:10.1039/D2CP04478A.

Basically all thermodynamic models of mixtures rely on describing *pair interactions*. Component-specific models, like UNIQUAC [1, 2] or NRTL [3], thereby describe the pairwise interactions between components using *component-specific* pair-interaction parameters, which need to be fitted to experimental data. Usually, data for binary mixtures are used for this purpose, which means that for modeling multi-component mixtures, binary mixture data are needed for all binary subsystems of the studied mixture. Unfortunately, due to the combinatorial problem, even data for binary mixtures are often missing, which strongly limits the applicability of the component-specific models.

GC methods circumvent this problem. By dividing components into structural groups, GC methods only rely on *group-specific* pair-interaction parameters, namely *group-interaction* parameters, which are fitted to experimental mixture data, whereby the amount of required training data compared to component-specific models is significantly reduced.

One of the most successful thermodynamic group contribution method for mixtures is UNIFAC, which was first introduced in 1975 [4] and has been significantly extended and refined since then [5, 6, 7, 8, 9, 10]. Also, several tailored versions of UNIFAC fitted for specific applications are available [11, 12, 13]. And there is also a commercial version of UNIFAC, provided within the UNIFAC-Consortium, which is based on the same model equations as the public versions of UNIFAC, but whose parameter tables have been revised and extended on a regular basis since 1996 [14] using both public data and non-public data provided or generated within the consortium. The scope of the commercial version is therefore larger than that of the public versions of UNIFAC. Since the commercial version is not freely accessible, we focus here on the most recent public version of UNIFAC [10], to which we refer simply as UNIFAC in the following for brevity. The authors have also access to the commercial version of UNIFAC, called UNIFAC-TUC in the following, but this version is used for comparisons only.

UNIFAC was derived from the component-specific lattice model UNIQUAC [1, 2] and describes the molar excess Gibbs energy g^{E} of a mixture as a function of temperature T and composition x. Both energetic and entropic contributions to g^{E} are considered in the model. All versions of UNIFAC use geometric parameters for the individual structural groups, which describe their volume and surface and determine the entropic contribution. Furthermore, parameters describing the pairwise energetic interactions between the different structural groups in the mixture are used. These group-interaction parameters play the central role in the model.

From the Gibbs excess energy g^{E}, many properties that are essential in chemical engineering can be determined, most importantly the activity coefficients γ_i of the components i in the mixture, based on which phase equilibria can be predicted [15]. Over the years, many structural groups have been included in the UNIFAC parameter tables, so that a huge number of components of practical interest can be modeled. UNIFAC presently considers 54 *main groups*, which are further divided into 113 *sub groups* [10]. The difference between main and sub groups is that each sub group g has individual geometric parameters, namely the group volume R_g and group surface area Q_g [16], while all sub groups that belong to the same main group G share the same group-interaction parameters. There are two distinct group-interaction parameters for each binary combination of different main groups (G, G'); they are generally labeled as $A_{GG'}$ and $A_{G'G}$, and have, as a result of the fit, usually different values, i.e., $A_{GG'} \neq A_{G'G}$.

While Q_g and R_g are reported for 113 individual sub groups, there are still significant gaps regarding the group-interaction parameters $A_{GG'}$ and $A_{G'G}$ between the 54 main groups: there are 1,431 distinct binary combinations of unlike main groups ($G \neq G'$), for which only for 635 (44%) group-interaction parameters have been reported yet. Figure 1 schematically shows the publicly available set of group-interaction parameters between the first 50 main groups of UNIFAC [10]. The first 50 main groups were chosen here since for all of these, group-interaction parameters with at least five other main groups are publicly available to date. This threshold was chosen since, as described in detail below, the missing group-interaction parameters were predicted based on information from the *available* parameters only. For the sake of completeness, Figure S.1 in the ESI shows for which of the group combinations parameters are available in the commercial UNIFAC-TUC.

Hence, the availability of the parameters describing the individual groups R_g and Q_g generally poses no problem, whereas missing group-interaction parameters $A_{GG'}$ and $A_{G'G}$ significantly limit the applicability of all versions of UNIFAC. The main reason why these gaps still persist, after so many years of work on the development of UNIFAC, is that the data base for their determination is simply too narrow. There are structural groups that occur in many molecules, such as the methyl group or the hydroxyl group, and there are less common groups. It is particularly these less common groups for which the parameters are lacking. This is not to say that these groups do not occur in interesting components, but there are simply less data on binary mixtures containing components with these groups. It is evident that this causes problems in the parameterization of UNIFAC.

A further drawback is that fitting group-interaction parameters is still not a routine, but rather artwork, in particular regarding the selection of the considered data sets, including their initial evaluation and consistency checking, and

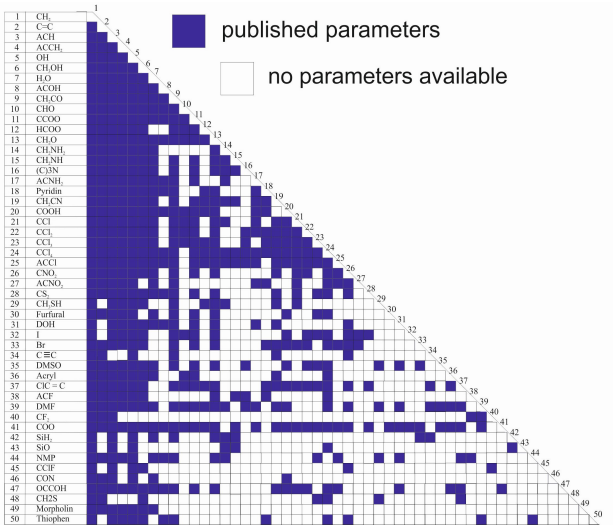

Figure 1: Matrix representing the availability of group-interaction parameters of UNIFAC [10] up to main group 50. Blue: parameters available.

regarding the selection of a suitable objective function to be minimized during the fitting procedure. For a more detailed description of the fitting procedure of UNIFAC group-interaction parameters, we refer to the literature[15, 17, 18, 19].

In this work, we present a method for the *prediction of the complete set* of the group-interaction parameters of group contribution methods based on an existing parameter set, without requiring new experimental data. The basic idea is to consider the group-interaction parameters as entries of a squared matrix (which is only partially filled, as several parameters are missing), and to use a matrix completion method (MCM) [20, 21] to estimate the missing entries. To demonstrate the applicability of our approach, it is applied to UNIFAC [10], for which the complete set of the group-interaction parameters for the first 50 main groups is predicted. Figure S.2 in the ESI gives an overview of our approach.

Following an idea developed in a recent paper [22], in which we have applied an MCM for estimating the component-specific pair-interaction parameters of UNIQUAC, we do not use the *asymmetric* group-interaction parameters ($A_{GG'} \neq A_{G'G}$) directly, but rather the *symmetric* group-interaction *energies* $U_{GG'} = U_{G'G}$. The parameters of the two types (A and U) are connected by:

$$\begin{aligned} A_{GG'} &= U_{GG'} - U_{G'G'} \\ A_{G'G} &= U_{G'G} - U_{GG} \end{aligned} \tag{1}$$

Hence, according to Eq. 1, $A_{GG'}$ and $A_{G'G}$ are not independent but correlated.[3] Despite this, for parameterizing UNIFAC, $A_{GG'}$ and $A_{G'G}$ are usually considered to be uncorrelated. The fitting then results in a parameter set that does not comply with Eq. 1, cf. Ref [22]. Our approach overcomes this inconsistency.

In a series of recent papers, we have demonstrated the capabilities of MCMs for predicting different types of thermodynamic data of mixtures using various component-based approaches [23, 24, 25, 26, 27, 22]. However, these component-based approaches are inherently limited regarding the number of components that are covered; the respective models complete a matrix spanned by the components that are part of the mixtures in the training set. This is not the case for the group contribution methods, which we consider in the present work: as the groups form building blocks from which components can be created flexibly, the scope of the group contribution methods for mixture properties is inherently extremely large – and it can now be extended substantially by using an MCM to complete the set of group-interaction parameters.

The approach we propose here should also be applicable to any other version of UNIFAC, and to other group contribution models for predicting thermodynamic properties of mixtures that are based on pair interactions. One advantage of our approach is that it can be put into practice, e.g., be integrated into existing process simulators, in a very simple

[3] For an N-component mixture, there are $N^2 - N$ asymmetric pair-interaction parameters of the A-type (the diagonal remains empty or is filled with zeros), while there are $(N^2 - N)/2 + N$ symmetric pair-interaction energies of the U-type (the diagonal is occupied by the pure-component energies, but only one of the triangular matrices has to be filled due to the symmetry). It is always possible to determine the A-parameters from the U-parameters, but not vice versa.

and straightforward manner: one only has to replace the existing UNIFAC parameter set of the model implementation by the predicted one provided with our approach. For other machine-learning approaches, like artificial neural networks operating on molecular graphs [28, 29] or SMILES representations of the components[30], this might be more complicated in practice.

2 Method

We demonstrate the applicability of using MCMs for the prediction of group-interaction parameters of thermodynamic group contribution methods by applying it to UNIFAC [10]. The resulting new version of UNIFAC (in which the predicted new parameters are used) is called *UNIFAC-MCM* in the following.

The MCM that was used in the present work is based on Bayesian matrix factorization [31] and similar to the ones used in our previous works[23, 24, 25, 27, 22]. In principle, we could have applied the MCM directly to the matrix of the A-type parameters, i.e., the matrix containing the group-interaction parameters $A_{GG'}$ and $A_{G'G}$. However, this option was discarded for the following reasons: firstly, the available values for $A_{GG'}$ and $A_{G'G}$ are inconsistent with Eq. 1. Also, fitting $A_{GG'}$ and $A_{G'G}$ to mixture data can give different combinations of these parameters yielding basically equivalent results for the physical properties to which they were fitted [32]. This hinders an interpretation of these parameters and makes them poor candidates for applying an MCM. These problems were overcome by working with the group-interaction *energies* $U_{GG'}$ as explained below. Furthermore, in applying the MCM to the A matrix, the target function would have been to achieve an optimal representation of the A-type parameters. However, with UNIFAC-MCM, we are rather interested in an optimal description of activity coefficients than in a representation of model parameters. UNIFAC-MCM was therefore trained on pseudo-data for activity coefficients as described in the next section.

2.1 Training Data

As training data for UNIFAC-MCM, we have generated pseudo-data for the logarithmic activity coefficients $\ln \gamma_{GG'}$ in hypothetical binary mixtures of the 'pure main groups' of UNIFAC (G and G') at different temperatures and group mole fractions. Here, $\ln \gamma_{GG'}$ represents the logarithmic activity coefficient of G in the binary mixture with G'. For any given temperature and mole fraction, there are two distinct values $\ln \gamma_{GG'}$ and $\ln \gamma_{G'G}$, respectively, which can be represented in a matrix. The diagonal elements of this matrix are occupied with ones by definition and were not considered here. For simplicity, we will simply speak of $\ln \gamma_{GG'}$ in the following referring to that matrix, which includes the values from both triangular matrices, $\ln \gamma_{GG'}$ and $\ln \gamma_{G'G}$.

Specifically, we have calculated $\ln \gamma_{GG'}$ for all binary combinations of the first 50 main groups of UNIFAC for which the required parameters were available, which holds for 619 combinations (or 50.5% of all possible binary combinations of these main groups). The grid was spanned by $T \in \{250, 300, 350, 400, 450\}$ K for the temperature, which covers the temperature of most of the available experimental data, and $x_G \in \{0.01, 0.2, 0.4, 0.6, 0.8, 0.99\}$ mol/mol for the composition.

For generating the pseudo-data for $\ln \gamma_{GG'}$, the UNIFAC equations (cf. Eqs. S.1 - S.11 in the ESI) were used in the common manner for hypothetical components that were composed of a single main group in all cases. For main groups G with several sub groups g (with individual geometric parameters Q_g and R_g), the values of Q_g and R_g for one of the respective sub groups were selected, for details see Table S.1 in the ESI. In principle, UNIFAC-MCM could also be trained on data for the residual part of the activity coefficients alone, which describes the energetic interactions (cf. Eq. S.7 in the ESI), because the interaction parameters only occur in this term. We have also tested this option and found the results very similar to those reported here, as expected.

2.2 Matrix Factorization

At its heart, UNIFAC-MCM factorizes the matrix of group-interaction energies $U_{GG'}$ between UNIFAC main groups G and G'. The *unlike* $U_{GG'}$ ($G \neq G'$) are modeled as the sum of two dot products:

$$U_{GG'} = U_{G'G} = \theta_G \cdot \beta_{G'} + \theta_{G'} \cdot \beta_G \tag{2}$$

where θ_G and β_G as well as $\theta_{G'}$ and $\beta_{G'}$ are vectors of length K containing a-priori unknown (latent) features of the UNIFAC main groups G and G', respectively. θ_G, β_G, $\theta_{G'}$, and $\beta_{G'}$ are parameters of UNIFAC-MCM, while K is a hyperparameter that controls the number of features considered per main group and thereby determines the flexibility of the model. Based on results of our prior work [22], K was set to $K = 3$ here. The form of Eq. 2 was chosen to ensure that all resulting group-interaction energies are symmetric, as required by the lattice model. Besides the unlike

interaction energies, also *like* group-interaction energies U_{GG} are needed, cf. Eq. 1. They were not included in the factorization (Eq. 2) but determined directly in the fit.

For training UNIFAC-MCM on the pseudo-data for $\ln \gamma_{GG'}$, cf. Section 'Training Data', the matrix factorization of the group-interaction energies $U_{GG'}$, cf. Eq. 2, as well as Eq. 1, which relates the $U_{GG'}$ to the group-interaction parameters $A_{GG'}$, were embedded in the UNIFAC equations, cf. Eqs. S.1-S.11 in the ESI. This establishes a generative probabilistic model for the $\ln \gamma_{GG'}$. The training data were hence modeled by:

$$\ln \gamma_{GG'}(T, x_G) = \mathrm{UNIFAC}(T, x_G, \theta_G, \theta_{G'}, \beta_G, \beta_{G'}, U_{GG}, U_{G'G'}) + \varepsilon_{GG'} \tag{3}$$

where $\varepsilon_{GG'}$ is the deviation between the modeled $\ln \gamma_{GG'}$ and the training data. The model parameters $\theta_G, \theta_{G'}, \beta_G, \beta_{G'}, U_{GG}$, and $U_{G'G'}$ were fitted in a Bayesian framework to minimize these deviations. For more details on the implementation of the model and the training procedure, we refer to the ESI.

2.3 Prediction of UNIFAC Group-Interaction Parameters

UNIFAC-MCM only contains parameters for the 'pure' main groups, namely $\theta_G, \beta_G, \theta_{G'}, \beta_{G'}$, U_{GG}, and $U_{G'G'}$, which were fitted to the 'group-mixture' data, namely the pseudo-data for $\ln \gamma_{GG'}$, during the training of the model as described above. Based on the learned parameters, the group-interaction energies $U_{GG'}$ of all combinations of the considered main groups can be calculated based on Eq. 2, from which, in turn, the commonly used group-interaction parameters of UNIFAC $A_{GG'}$ and $A_{G'G}$ can be predicted from Eq. 1. Hence, a *complete* parameterization of UNIFAC regarding the first 50 main groups is obtained by this procedure, which can be used for predicting temperature- and concentration-dependent activity coefficients $\ln \gamma_i$ of all components i in any (binary or multi-component) mixture, if all components that make up the mixture can be segmented using the first 50 main groups of UNIFAC. We report the predicted complete set of $A_{GG'}$ (and of $U_{GG'}$) as .CSV file in the ESI. Note that this set of $A_{GG'}$ is consistent in terms of fulfilling Eq. 1 as demanded by the lattice theory, which is in contrast to the previously available UNIFAC parameter tables that were obtained by fitting $A_{GG'}$ individually.

The latter also explains why a direct matrix factorization of the reported $A_{GG'}$ is not expedient, and instead the pseudo-data for $\ln \gamma_{GG'}$ were used for training UNIFAC-MCM; the reported $A_{GG'}$ matrix simply lacks structure that could be exploited by the MCM.

3 Results and Discussion

In the following, we evaluate the quality of UNIFAC-MCM by considering predictions of vapor-liquid equilibria (VLE), which is probably the most important field in which activity coefficients are applied. As basis for this evaluation, we have used all VLE data sets for binary mixtures from the Dortmund Data Bank (DDB) [33, 34, 35] that comply with the following conditions:

- both components of the mixture can be built from the first 50 main groups of UNIFAC [10];
- the data set contains information on temperature, pressure, and composition of the liquid and vapor phase;
- the data set is labeled as 'thermodynamically consistent' in the DDB, i.e., it fulfills area and point-to-point test [36, 37, 38];
- Antoine parameters for calculating the pure-component vapor pressure at the temperature of the VLE are available in the DDB for both components;
- the pressure is not higher than 10 bar to justify the assumption of an ideal gas phase.

In the present version of the DDB, such VLE data are available for 2,246 distinct binary systems. We will call this complete set of binary systems 'complete horizon' in the following.

The VLE were predicted using extended Raoult's law assuming an ideal vapor phase and a pressure independence of the chemical potentials in the liquid phase:

$$p_i^{\mathrm{s}}(T)\, x_i\, \gamma_i(T, x_i) = p\, y_i \tag{4}$$

For the calculations, the mole fractions x_i in the liquid phase as well as either the pressure p (for isobaric data sets) or the temperature T (for isothermal data sets) were specified, the pure component vapor pressure p_i^{s} was calculated with the Antoine equation using the parameters from the DDB, and the activity coefficients γ_i of the components in the liquid phase were predicted with UNIFAC-MCM. The mole fractions y_i in the vapor phase and the pressure p (for isothermal data sets) or the temperature T (for isobaric data sets) were then calculated from the system of equations resulting from applying Eqn. 4 to both components. The results were compared to the experimental data from the DDB, with a focus on the gas phase mole fractions of the low-boiling component.

For comparison, the same calculations were also carried out with UNIFAC [10]; albeit, this is only possible for a subset of 2,068 systems from the complete horizon ('UNIFAC horizon'). At a first glance, it may look disappointing that by using UNIFAC-MCM, with its substantially enlarged parameter table, only 178 additional systems for which data are available can be modeled. However, this is as expected: the lack of data on these systems has hindered the extension of the UNIFAC parameter table so far. Furthermore, we have also used the commercial version UNIFAC-TUC for comparison, which enabled predictions of VLE for 2,237 of the studied systems ('UNIFAC-TUC horizon'). We have included the results from UNIFAC-TUC in the comparison (even though it is not publicly available) for two reasons: firstly, it is the best available benchmark method and, secondly, it allows to evaluate the predictive performance UNIFAC-MCM also on systems that can not be modeled by UNIFAC, which is the basis of UNIFAC-MCM.

The results are shown in Figure 2, where the horizons in the three panels differ: in the left panel, it is the complete horizon, in the middle panel, it is the UNIFAC-TUC horizon, and in the right one, it is the smallest horizon, that of UNIFAC[10].

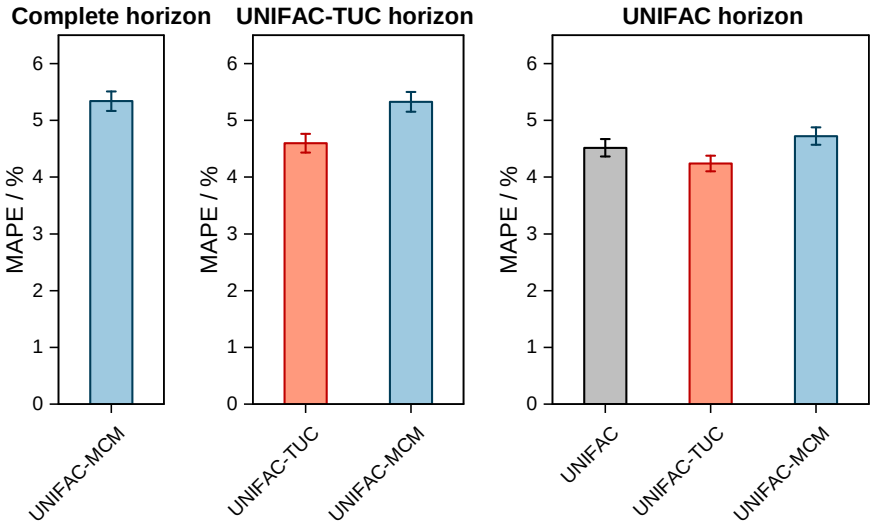

Figure 2: Mean Absolute Percentage Error (MAPE) of the predicted vapor-phase mole fractions of the low-boiling component in VLE with UNIFAC-MCM for the 'complete horizon' (2,246 systems, left) and comparison to the commercial UNIFAC-TUC for the 'UNIFAC-TUC horizon' (2,237 systems, middle), and to the public UNIFAC [10] for the 'UNIFAC horizon' (2,068 systems, right). Error bars denote standard errors of the means.

The results obtained with UNIFAC-MCM on the complete horizon are shown in Figure 2 (left), where the Mean Absolute Percentage Error (MAPE) in y_i of the low-boiling component of the predictions with UNIFAC-MCM averaged over all 2,246 systems is plotted, which was calculated by comparing the UNIFAC-MCM predictions system-wise to the respective experimental data from the DDB. As the results indicate, UNIFAC-MCM predicts the vapor-phase mole fractions for all 2,246 studied binary systems with an average error of 5.3%, which is not much larger than the typical uncertainty of experimental data for vapor-phase mole fractions. The MAPE of UNIFAC-MCM in the pressure p, averaged over all isothermal data sets from the complete horizon, is $5.0\pm0.2\%$; the MAPE in the absolute temperature T in K, averaged over all isobaric data sets from the complete horizon, is $0.48\pm0.02\%$.

In the middle panel of Figure 2, the performance of MCM-UNIFAC is compared to that of UNIFAC-TUC, and in the right panel, it is compared to UNIFAC [10] as well as to UNIFAC-TUC. The highest accuracy among the three models is found for the commercial UNIFAC-TUC (MAPE of 4.6% on the UNIFAC-TUC horizon, cf. middle panel, and 4.2% on the UNIFAC horizon, cf. right panel), which is not surprising since a lot of effort has been put into refining its parameterization during the last decades. However, the scores of UNIFAC-MCM (MAPE of 5.3% on the UNIFAC-TUC horizon, cf. middle panel, and 4.7% on the UNIFAC horizon, cf. right panel) are only slightly worse than that of UNIFAC-TUC.

On the UNIFAC horizon, cf. Figure 2 (right), the scores of UNIFAC-MCM (MAPE of 4.7%) and of the public UNIFAC (MAPE of 4.5%) are very similar. This demonstrates two things: first, that the additional flexibility of the UNIFAC model achieved by the inconsistent *individual* fitting of group-interaction parameters $A_{GG'}$ and $A_{G'G}$ compared to the sole physical consideration of group-interaction energies $U_{GG'}$ (including the like group-interaction energies U_{GG} and $U_{G'G'}$) is unnecessary; for the complete matrix of the considered 50 main groups of UNIFAC, there are 2,450 distinct group-interaction parameters $A_{GG'}$ and $A_{G'G}$, but only 1,275 distinct group-interaction energies $U_{GG'}$ (including 50

like energies U_{GG}). And second, the MCM, which is at the heart of UNIFAC-MCM, is able to capture the structure within the unlike group-interaction energies using six latent parameters for each main group.

It is interesting to also study the performance of UNIFAC-MCM and UNIFAC-TUC only for those systems that *cannot* be modeled with UNIFAC [10]; this gives an impression of the performance of UNIFAC-MCM when applied for true predictions, namely for systems containing combinations of main groups for which no interaction parameters of UNIFAC are available, as it is unlikely that data on any of these systems were used in the development of UNIFAC [10], on which UNIFAC-MCM is based. In contrast, it may be assumed that basically all these additional VLE data were used for the development of UNIFAC-TUC, so that for UNIFAC-TUC, such a comparison shows basically only if the correlation of these additional data was successful. The respective results are presented in Figure 3. Most of the systems within the complete horizon can be modeled not only with UNIFAC-MCM but also with UNIFAC-TUC. The few systems for which this is not the case, are treated separately in Figure 3 (left panel).

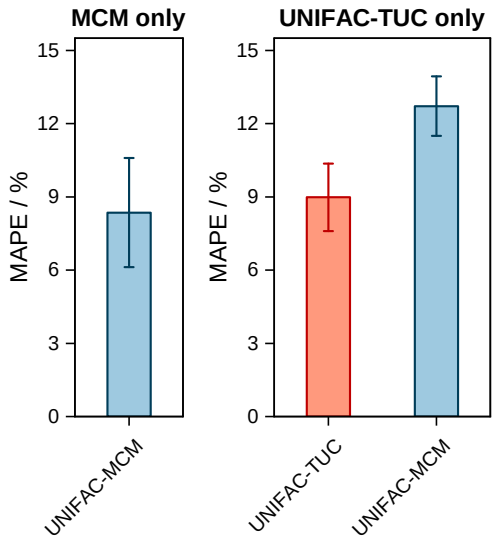

Figure 3: Mean Absolute Percentage Error (MAPE) of the predicted mole fraction of the low-boiling component in the vapor phase in VLE with UNIFAC-MCM for the systems that can only be modeled by UNIFAC-MCM (left, 'MCM only', 9 systems), and those systems that can also be predicted with UNIFAC-TUC but not with UNIFAC (middle, 'UNIFAC-TUC only', 169 systems). Error bars denote standard errors of the means.

The first message from Figure 3 is that the deviations increase compared to the ones shown in Figure 2, which holds both for UNIFAC-TUC and UNIFAC-MCM. Averaged over all systems that can be modeled by both models (but not by UNIFAC), cf. Figure 3 (right), the MAPE for UNIFAC-TUC is now 9.0%, that for UNIFAC-MCM is 12.7%. However, considering that the results from Figure 3 obtained with UNIFAC-MCM are bold predictions, while those from UNIFAC-TUC are basically only correlations, the difference between both methods is unexpectedly small.

Comparing the results from Figure 3 with those from Figure 2 is most informative when referring to Figure 2 (right), where the UNIFAC horizon is shown, because it then gives an impression on the changes when carrying out the comparison for complementary data sets: the UNIFAC horizon, for which the results are shown in Figure 2 (right), covers all systems that can also be modeled by the public UNIFAC; Figure 3, on the other hand, shows the results for all remaining systems from our data set, i.e., for the ones that *cannot* be modeled by the public UNIFAC.

Carrying out this comparison for UNIFAC-TUC (for which the results are correlations in both cases) clearly shows that the systems studied in Figure 3 are more difficult to describe than those studied in Figure 2 (right). We are not going into the details of these additional difficulties, which can be related to different factors, including spotty and uncertain data (cf. also Figure S.3 in the ESI) as well as to the fact that many of the respective systems contain components with special properties (highly halogenated or reactive components), which substantially complicates the accurate modeling with UNIFAC.

Hence, the results for UNIFAC-TUC indicate that most of the increase of the MAPE scores observed also for UNIFAC-MCM when going from Figure 2 (right) to Figure 3 is simply due to the increased difficulties in describing the data considered in Figure 3, and, thus, cannot be attributed to a lack of predictive power.

We only note here that the scope of the developed UNIFAC-MCM is much larger than we can demonstrate here, simply due to the fact that for many of the group-interaction parameters that can now be predicted, no experimental data for testing are available, cf. Figure S.3 in the ESI. An alternative representation of the results of UNIFAC-MCM in the form of histograms is given in Figure S.4 in the ESI.

In Figure 4, we show some typical examples for the prediction of vapor-liquid phase diagrams with UNIFAC-MCM and compare the results to those obtaiend with UNIFAC-TUC. Only systems that *cannot* be modeled by the public UNIFAC version were therefore chosen, such that the results of UNIFAC-MCM are true predictions. This is, again, not the case for UNIFAC-TUC, as the data shown in Figure 4 were available for the development of the method. In all cases, UNIFAC-MCM represents the different types of phase behavior well.

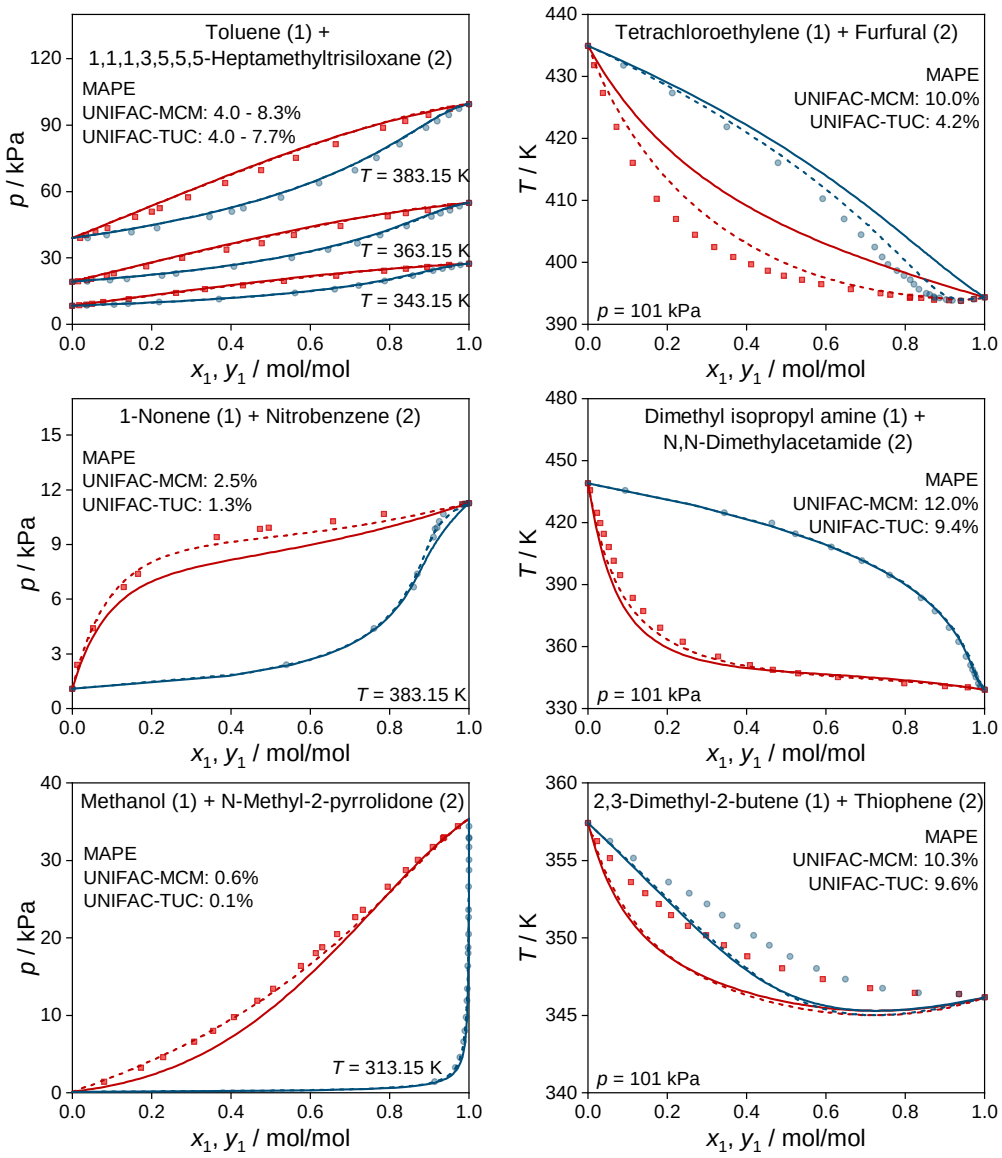

Figure 4: Prediction of vapor-liquid phase diagrams for binary systems with UNIFAC-MCM (solid lines) and UNIFAC-TUC (dashed lines) and comparison to experimental data from the DDB (symbols). For each system, the MAPE in the predicted vapor-phase mole fraction of the low-boiling component is given for both models. All shown systems can not be predicted with the public UNIFAC version. Blue: dew point curves. Red: bubble point curves.

In Figure 5, we show two further examples for the prediction of VLE phase diagrams with UNIFAC-MCM. The chosen systems can neither be modeled by the public UNIFAC, nor with the commercial UNIFAC-TUC due to missing group-

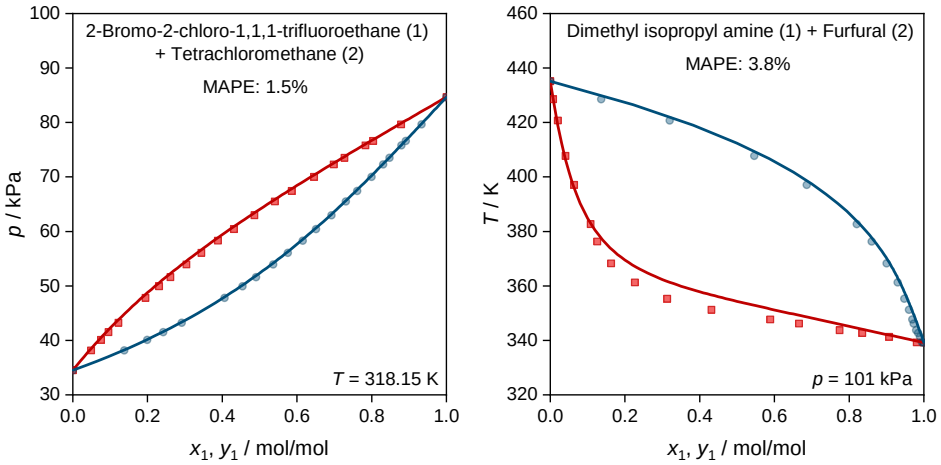

Figure 5: Prediction of vapor-liquid phase diagrams for binary systems with UNIFAC-MCM (lines) and comparison to experimental data from the DDB (symbols). For both system, the MAPE in the predicted vapor-phase mole fraction of the low-boiling component is given. Both systems can neither be predicted with the public UNIFAC version, nor with the commercial UNIFAC-TUC. Blue: dew point curves. Red: bubble point curves.

interaction parameters in both models. We observe an almost perfect agreement of the predictions with UNIFAC-MCM and the experimental data, but note that we also find systems with poorer agreement, cf. Figure S.4 in the ESI.

UNIFAC-MCM should in general be used in cases in which required group-interaction parameters of UNIFAC are missing, while in cases in which all parameters are available, we recommend using these. The reason is that UNIFAC-MCM is basically a derivate of UNIFAC, i.e., based on the available parameter tables, and it would only be by chance were it better than its basis for certain systems. However, we emphasize that the differences between UNIFAC and UNIFAC-MCM are not expected to be large, as shown in Figure 2.

4 Conclusions

Group contribution methods for the prediction of thermophysical properties are highly important in chemical engineering. One of the most successful of these methods is UNIFAC. However, the applicability of UNIFAC is still substantially hampered by missing group-interaction parameters, which is in particular due to the lack of suitable mixture data for fitting the parameters. As a consequence, there are still significant gaps in the matrix in which these UNIFAC parameters are usually represented.

In the present work, we present an approach to *complete* the group-interaction parameter set of UNIFAC using a matrix completion method (MCM) from machine learning. Our approach, called UNIFAC-MCM, was trained in a purely data-based manner solely on pseudo-data generated with UNIFAC, and approximately doubles the number of available group-interaction parameters.

We have evaluated the performance of UNIFAC-MCM for the prediction of vapor-liquid equilibria (VLE) of 2,246 binary systems from the Dortmund Data Bank. This set can be divided into data that can be predicted with the public UNIFAC (2,068 systems) and data for which this is not the case, but which can be predicted with the developed UNIFAC-MCM (169 systems). The latter set is comparatively small, as the missing groups in UNIFAC are rather uncommon ones, i.e., only present in components for which only few data have been measured.

Where a direct comparison is possible, UNIFAC and UNIFAC-MCM show a similar performance. This alone is astonishing since UNIFAC-MCM is based only on *consistent group-interaction* energies, whereas in UNIFAC the number of the parameters to describe the pairwise interactions has almost been doubled, simply to increase the flexibility, which is, however, not well founded in the physical lattice theory from which UNIFAC was derived. For the systems for which UNIFAC can not be applied, the performance of UNIFAC-MCM is poorer but still acceptable, especially given the fact the this set contains basically only demanding systems, as also the commercial version UNIFAC-TUC, which we used for comparison here, shows significantly larger error scores.

This work has shown that working with consistent group-interaction energies is not only a feasible alternative to the common procedure of fitting UNIFAC parameters, but also a highly attractive one: a similar quality is obtained by a significantly smaller (approx. 50%) number of parameters, which promises a higher predictive performance and could

be useful also for the fitting of new UNIFAC parameters in the future. The predicted parameters provided in this work might in general serve as valuable starting points for the future fitting of UNIFAC parameters to experimental data. In future work, it will be interesting to include further structural groups in the model, to transfer our approach to other group contribution methods for mixture properties, and to consider an end-to-end training, directly on experimental VLE data.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors gratefully acknowledge financial support by Carl Zeiss Foundation in the frame of the project 'Process Engineering 4.0' and by Germany's Bundesministerium für Wirtschaft und Klimaschutz (BMWK) in the frame of the project 'KEEN'.

References

[1] Denis S Abrams and John M Prausnitz. Statistical thermodynamics of liquid mixtures: a new expression for the excess gibbs energy of partly or completely miscible systems. *AIChE Journal*, 21(1):116–128, 1975.

[2] G. Maurer and J.M. Prausnitz. On the derivation and extension of the uniquac equation. *Fluid Phase Equilibria*, 2(2):91–99, 1978.

[3] Henri Renon and J. M. Prausnitz. Local compositions in thermodynamic excess functions for liquid mixtures. *AIChE Journal*, 14(1):135–144, 1968.

[4] Aage Fredenslund, Russell L Jones, and John M Prausnitz. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AIChE Journal*, 21(6):1086–1099, 1975.

[5] Steen Skjold-Jorgensen, Barbel Kolbe, Jurgen Gmehling, and Peter Rasmussen. Vapor-liquid equilibria by unifac group contribution. revision and extension. *Industrial & Engineering Chemistry Process Design and Development*, 18(4):714–722, 1979.

[6] Jürgen Gmehling, Peter Rasmussen, and Aage Fredenslund. Vapor-liquid equilibriums by unifac group contribution. revision and extension. 2. *Industrial & Engineering Chemistry Process Design and Development*, 21(1):118–127, 1982.

[7] Eugenia Almeida Macedo, Ulrich Weidlich, Juergen Gmehling, and Peter Rasmussen. Vapor-liquid equilibriums by unifac group contribution. revision and extension. 3. *Industrial & Engineering Chemistry Process Design and Development*, 22(4):676–678, 1983.

[8] Detlef Tiegs, Peter Rasmussen, Juergen Gmehling, and Aage Fredenslund. Vapor-liquid equilibria by unifac group contribution. 4. revision and extension. *Industrial & Engineering Chemistry Research*, 26(1):159–161, 1987.

[9] Henrik K Hansen, Peter Rasmussen, Aage Fredenslund, Martin Schiller, and Jürgen Gmehling. Vapor-liquid equilibria by unifac group contribution. 5. revision and extension. *Industrial & Engineering Chemistry Research*, 30(10):2352–2355, 1991.

[10] Roland Wittig, Juergen Lohmann, and Juergen Gmehling. Vapor- liquid equilibria by unifac group contribution. 6. revision and extension. *Industrial & Engineering Chemistry Research*, 42(1):183–188, 2003.

[11] Thomas Magnussen, Peter Rasmussen, and Aage Fredenslund. Unifac parameter table for prediction of liquid-liquid equilibriums. *Industrial & Engineering Chemistry Process Design and Development*, 20(2):331–339, 1981.

[12] G Wienke and J Gmehling. Prediction of octanol-water partition coefficients, henry coefficients and water solubilities using unifac. *Toxicological & Environmental Chemistry*, 65(1-4):57–86, 1998.

[13] Weidong Yan, Magnus Topphoff, Christian Rose, and Jürgen Gmehling. Prediction of vapor–liquid equilibria in mixed-solvent electrolyte systems using the group contribution concept. *Fluid Phase Equilibria*, 162(1-2):97–113, 1999.

[14] The unifac consortium, 2022.

[15] Aage Fredenslund, Jürgen Gmehling, and Peter Rasmussen. *Vapor-liquid equilibria using UNIFAC: a group-contribution method*. Elsevier, 1977.

[16] Arnold Aaron Bondi et al. *Physical Properties of Molecular Crystals Liquids, and Glasses*. Wiley, 1968.

[17] Aage Fredenslund. *Vapor-liquid equilibria using UNIFAC: a group-contribution method*. Elsevier, 2012.

[18] Jürgen Gmehling, Roland Wittig, Jürgen Lohmann, and Ralph Joh. A modified unifac (dortmund) model. 4. revision and extension. *Industrial & engineering chemistry research*, 41(6):1678–1688, 2002.

[19] Bastian Schmid, Andre Schedemann, and Jürgen Gmehling. Extension of the vtpr group contribution equation of state: Group interaction parameters for additional 192 group combinations and typical results. *Industrial & Engineering Chemistry Research*, 53(8):3393–3405, 2014.

[20] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[21] Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Investigation of various matrix factorization methods for large recommender systems. In *2008 IEEE International Conference on Data Mining Workshops*, pages 553–562. IEEE, 2008.

[22] Fabian Jirasek, Robert Bamler, Sophie Fellenz, Michael Bortz, Marius Kloft, Stephan Mandt, and Hans Hasse. Making thermodynamic models of mixtures predictive by machine learning: matrix completion of pair interactions. *Chemical Science*, 13:4854–4862, 2022.

[23] Fabian Jirasek, Rodrigo A. S. Alves, Julie Damay, Robert A. Vandermeulen, Robert Bamler, Michael Bortz, Stephan Mandt, Marius Kloft, and Hans Hasse. Machine learning in thermodynamics: Prediction of activity coefficients by matrix completion. *The Journal of Physical Chemistry Letters*, 11(3):981–985, 2020. PMID: 31964142.

[24] Fabian Jirasek, Robert Bamler, and Stephan Mandt. Hybridizing physical and data-driven prediction methods for physicochemical properties. *Chemical Communications.*, 56:12407–12410, 2020.

[25] Fabian Jirasek and Hans Hasse. Perspective: Machine learning of thermophysical properties. *Fluid Phase Equilibria*, 549:113206, 2021.

[26] Julie Damay, Fabian Jirasek, Marius Kloft, Michael Bortz, and Hans Hasse. Predicting activity coefficients at infinite dilution for varying temperatures by matrix completion. *Industrial & Engineering Chemistry Research*, 60(40):14564–14578, 2021.

[27] Nicolas Hayer, Fabian Jirasek, and Hans Hasse. Prediction of henry's law constants by matrix completion. *AIChE Journal*, 68:e17753, 2022.

[28] Jan G Rittig, Karim Ben Hicham, Artur M Schweidtmann, Manuel Dahmen, and Alexander Mitsos. Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids. *arXiv preprint arXiv:2206.11776*, 2022.

[29] Edgar Ivan Sanchez Medina, Steffen Linke, Martin Stoll, and Kai Sundmacher. Graph neural networks for the prediction of infinite dilution activity coefficients. *Digital Discovery*, 2022.

[30] Benedikt Winter, Clemens Winter, Johannes Schilling, and André Bardow. A smile is all you need: Predicting limiting activity coefficients from smiles with natural language processing. *Digital Discovery*, 2022.

[31] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th International Conference on Machine Learning*, pages 880–887, 2008.

[32] John M Prausnitz, FF Anderson, and TF Anderson. *Computer calculations for multicomponent vapor-liquid and liquid-liquid equilibria*. Prentice Hall, 1980.

[33] U Onken, J Rarey-Nies, and J Gmehling. The dortmund data bank: A computerized system for retrieval, correlation, and prediction of thermodynamic properties of mixtures. *International Journal of Thermophysics*, 10(3):739–747, 1989.

[34] Jürgen Gmehling. Sophisticated thermodynamic models and dortmund data bank. *Vakuum in Forschung und Praxis*, 14(5):272–279, 2002.

[35] Dortmund data bank (ddb), 2022.

[36] Otto Redlich and AT Kister. Algebraic representation of thermodynamic properties and the classification of solutions. *Industrial & Engineering Chemistry*, 40(2):345–348, 1948.

[37] EFG Herington. A thermodynamic test for the internal consistency of experimental data on volatility ratios. *Nature*, 160(4070):610–611, 1947.

[38] Hendrick C Van Ness, Stanley M Byer, and Richard E Gibbs. Vapor-liquid equilibrium: Part i. an appraisal of data reduction methods. *AIChE Journal*, 19(2):238–244, 1973.