Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

Check for updates

Cite this: *Chem. Sci.*, 2022, **13**, 4854

All publication charges for this article have been paid for by the Royal Society of Chemistry

Making thermodynamic models of mixtures predictive by machine learning: matrix completion of pair interactions†

Fabian Jirasek, *[a] Robert Bamler,[b] Sophie Fellenz,[c] Michael Bortz,[d] Marius Kloft,[c] Stephan Mandt‡[e] and Hans Hasse‡[a]

Received 26th December 2021
Accepted 1st April 2022

DOI: 10.1039/d1sc07210b

rsc.li/chemical-science

Predictive models of thermodynamic properties of mixtures are paramount in chemical engineering and chemistry. Classical thermodynamic models are successful in generalizing over (continuous) conditions like temperature and concentration. On the other hand, matrix completion methods (MCMs) from machine learning successfully generalize over (discrete) binary systems; these MCMs can make predictions without any data for a given binary system by implicitly learning commonalities across systems. In the present work, we combine the strengths from both worlds in a hybrid approach. The underlying idea is to predict the *pair-interaction energies*, as they are used in basically all physical models of liquid mixtures, by an MCM. As an example, we embed an MCM into UNIQUAC, a widely-used physical model for the Gibbs excess energy. We train the resulting hybrid model in a Bayesian machine-learning framework on experimental data for activity coefficients in binary systems of 1146 components from the Dortmund Data Bank. We thereby obtain, for the first time, a *complete* set of UNIQUAC parameters for all binary systems of these components, which allows us to predict, in principle, activity coefficients at arbitrary temperature and composition for any combination of these components, not only for binary but also for multicomponent systems. The hybrid model even outperforms the best available physical model for predicting activity coefficients, the modified UNIFAC (Dortmund) model.

Introduction

Information on thermodynamic properties of mixtures is of crucial importance in chemical engineering and chemistry. However, providing this information is hampered by a combinatorial problem: the number N of known components is increasing rapidly (it is presently in the order of $N = 10^8$, counted by entries in the CAS Registry[1]); if only binary mixtures are considered, the number of mixtures that can be formed from these components already goes with N^2. Even if only technically relevant components and their mixtures are considered, the numbers are still extremely high. Consequently, experimental data on thermodynamic properties are available only for a small fraction of the relevant mixtures, especially as the corresponding experimental investigations are tedious. Therefore, methods for the prediction of thermodynamic properties of mixtures are essential in practice.

The present work is focused on thermodynamic properties of fluid mixtures. Physical models for describing these properties are generally based on the concept of *pair interactions*, which are commonly described by *pair-interaction energies*. All types of pair interactions in a multicomponent mixture can be investigated by studying the pure components (for the like interactions) and the binary subsystems (for the unlike interactions). Therefore, physical models for thermodynamic properties of fluid mixtures are generally developed using data on pure components and binary mixtures; based on this, they can also be used to predict properties of multicomponent mixtures. In the application of models of thermodynamic properties, their predictive capabilities are of prime importance. To assess them, the relation of the data used in the model development (the *training data*) to the data to be predicted is important. Regarding thermodynamic properties of mixtures, we distinguish two categories of predictions here:

[a]*Laboratory of Engineering Thermodynamics (LTD), TU Kaiserslautern, Erwin-Schrödinger-Str. 44, 67663 Kaiserslautern, Germany. E-mail: fabian.jirasek@mv.uni-kl.de*

[b]*Data Science and Machine Learning, University of Tübingen, Maria-von-Linden-Str. 6, 72076 Tübingen, Germany*

[c]*Machine Learning Group, TU Kaiserslautern, Paul-Ehrlich-Str. 34, 67663 Kaiserslautern, Germany*

[d]*Fraunhofer Institute for Industrial Mathematics (ITWM), Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany*

[e]*Department of Computer Science, University of California, Irvine, Donald Bren Hall, 92617 Irvine, CA, USA*

† Electronic supplementary information (ESI) available: UNIQUAC equations; information on data set and data preprocessing; model details incl. source codes to run the model in Stan; additional results; complete UNIQUAC parameter set. See https://doi.org/10.1039/d1sc07210b

‡ These authors jointly supervised this work.

This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

© 2022 The Author(s). Published by the Royal Society of Chemistry

(i) Those, in which for a given system (a fixed set of components) only the conditions are changed compared to the training data (*e.g.*, temperature, pressure, or concentration of the components); we refer to this as *generalization over conditions*.

(ii) Those, in which the system itself is changed, *i.e.*, was not included in the training data; we refer to this as *generalization over systems*.

While (i) involves *continuous* variables, (ii) is *discrete*. In the nomenclature of the present paper, we follow the common usage, which says that '*mixture*' can have two meanings: in the first, it simply designates the combination of specific components (regardless of their concentrations, *e.g.*, 'water + ethanol'), in the second, also the concentrations are included in the term (*e.g.*, 'water + ethanol with $x_{\text{water}} = 0.1$ mol/mol'). In cases where this ambiguity can lead to misunderstandings, we use the term '*system*' instead of 'mixture' when we refer to the first case.

The most common types of thermodynamic models of mixtures are models of the Gibbs excess energy G^{E}, such as UNIQUAC[2,3] or NRTL,[4] and equations of state (EoS).[5] These models excel in generalizing over conditions. By virtue of the underlying thermodynamic theory, they can also be used for *generalizing over properties*, *i.e.*, if trained on data for a specific property, they can be used to predict a different but related property; this is an important issue, which, however, is not in the focus of the present work.

In their basic form, G^{E} models and EoS are only partially useful for generalizing over systems: for the reasons given above, they are usually trained on data for binary systems, but can then be used for modeling those binary systems only for which data were available during the training. In contrast, predictions of properties of unstudied multicomponent systems with G^{E} models and EoS often turn out to be sufficiently accurate, under the condition that the model was trained on data for *all constituent binary subsystems*.[6]

To overcome the lack of generalization over (binary) systems, group-contribution (GC) approaches like UNIFAC,[7–9] modified UNIFAC (Dortmund),[10,11] and the Predictive Soave–Redlich–Kwong (PSRK) EoS[12,13] have been developed. They are based on the idea that components can be split into structural groups and, instead of considering pair interactions between components, pair interactions between these groups are modeled. As the number of relevant structural groups is comparatively small (in the order of 100), the combinatorial problem described above thereby becomes tractable. GC methods contain group-interaction parameters that are usually trained on data for binary systems containing the pertinent groups. They can then be used for predicting the properties of systems for which no data are available.

Unfortunately, the applicability of these GC methods is still hampered by incomplete group-interaction parameter tables. This is due to both the elaborate procedure of fitting new group-interaction parameters, and the fact that often not enough pertinent experimental data are available for a meaningful fit. The problems resulting from an inadequate database also lead to poor performances of GC methods when applied to systems with components that contain groups for which only a few training data points are available. The most prominent alternatives to GC methods are quantum-chemical approaches, namely COSMO-RS.[14,15] In principle, COSMO-RS enables predictions for any system based on quantum-chemical calculations, which are, however, computationally costly and not trivial for complex components. Furthermore, also these methods are tuned on experimental data, but the number of parameters is typically low and common users do not change them.[16] Compared to GC methods, quantum-chemical methods are often found to be somewhat less accurate in the prediction of thermodynamic properties.[17]

We have recently introduced a completely new approach for predicting thermodynamic properties of unstudied binary systems.[18–21] This approach is based on employing matrix completion methods (MCMs) from machine learning (ML), where the MCMs are prominently associated with recommender systems.[22,23] *E.g.*, for movie recommendations, these methods can implicitly learn and quantify similarities among users and similarities among movies by observing interaction patterns (ratings or clicks) between them, allowing to predict preference scores for unseen pairs of users and movies. In a similar spirit, our previous work[18,19] employed MCMs for predicting activity coefficients γ_{ij}^{∞} of solutes i in pure solvents j at infinite dilution and constant temperature, where the MCM learns similarities among the solutes and among the solvents. Applying the MCM approach for predicting thermodynamic properties is based on the fact that properties of binary mixtures at constant conditions, such as isothermal γ_{ij}^{∞}, can be stored in matrices, where the rows and columns represent the components i and j that make up the mixtures. As these matrices are only sparsely occupied by experimental data, the prediction of the unobserved entries constitutes a matrix completion problem. In our previous work,[18,19] we employed a matrix factorization following eqn (1):

$$\gamma_{ij}^{\infty} = f(\boldsymbol{\theta}_i \cdot \boldsymbol{\beta}_j) \tag{1}$$

The value of γ_{ij}^{∞} is thereby modeled by the dot product of two vectors $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_j$, which contain the so-called latent features and capture properties of the solute i and the solvent j, respectively, and which are inferred from the sparse available data on the mixture property γ_{ij}^{∞}. The function f was chosen to be the exponential function, taking into account that the γ_{ij}^{∞} are by definition non-negative and span a wide range of values. Moreover, this choice is also a physical one since the excess chemical potential $\mu_{ij}^{\text{E},\infty}$ of the solute i at infinite dilution in the solvent j depends linearly on $\ln \gamma_{ij}^{\infty}$.

The basic idea behind this MCM approach is that *information on the pure components* (as quantified by $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_j$) is implicitly contained in the *mixture data*, *i.e.*, in the way the components interact with each other, which manifests itself in the observable mixture property γ_{ij}^{∞}. This information is captured by the MCM and stored in the latent features $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_j$ of the components i and j. Using eqn (1) in the trained model, the inferred latent features allow predicting γ_{ij}^{∞} also for previously unstudied combinations i–j. As demonstrated in our previous work,[18,19] this approach even outperforms the present state-of-the-art method for predicting

© 2022 The Author(s). Published by the Royal Society of Chemistry

activity coefficients, namely the modified UNIFAC (Dortmund) model,[10,11] if γ_{ij}^{∞} at 298 K are considered. We have also extended the approach to modeling the dependence of γ_{ij}^{∞} on the temperature T, by simply exploiting the fact that this dependence can be well described by $\gamma_{ij}^{\infty}(T) = A_{ij} + B_{ij}/T$ with system-specific but temperature-independent parameters A_{ij} and B_{ij} in many cases.[20]

In the present work, we further expand the MCM approach by combining it with the physical modeling of thermodynamic properties of mixtures. We thereby exploit the fact that all physical models of mixtures are based on the idea of pair interactions, described by pair-interaction energies as critical parameters. We propose to predict these pair-interaction energies by an MCM to obtain a new *hybrid* concept for the prediction of properties of mixtures. The feasibility and the merits of this widely applicable approach are demonstrated by using the well-known G^{E} lattice model UNIQUAC[2,3] as an example.

En passant, we come back to the derivation of the UNIQUAC equation, in which two asymmetric pair-interaction parameters $\Delta U_{ij} \neq \Delta U_{ji}$ were introduced instead of using one symmetric pair-interaction energy $U_{ij} = U_{ji}$ as it follows from the lattice theory. The two adjustable parameters ΔU_{ij} and ΔU_{ji} were introduced as a workaround to get more flexibility for fitting binary phase equilibrium data. We show that this workaround is not necessary and base our considerations on the physical symmetric interaction energies $U_{ij} = U_{ji}$, explicitly including those for the like interactions U_{ii}.

The resulting new model, which we call 'MCM-UNIQUAC' in the following, combines the capabilities of the MCM regarding the generalization over binary systems with those of the physical model UNIQUAC regarding the generalization over conditions. It is a hybrid method that takes advantage of the strengths of both worlds; in particular, it also enables predictions of properties of multicomponent systems by virtue of the physics behind UNIQUAC, a feature a data-driven MCM cannot provide.

Development of MCM-UNIQUAC

The idea behind our approach is shown in Fig. 1. At its heart, an MCM is used to predict pair-interaction energies U_{ij} of a physical model of mixtures, which is UNIQUAC in the present work. The physical model relates the pair-interaction energies to temperature- and concentration-dependent properties of the binary mixtures of the components i and j. This could, in principle, be any property of interest. In the present work, we have chosen a fundamental thermodynamic property, the activity coefficient γ_{ij} of component i in a binary mixture with component j normalized according to Raoult's law. The activity coefficients, in turn, are directly related to observable mixture properties (*e.g.*, vapor–liquid equilibria (VLE),[24] liquid–liquid equilibria (LLE),[25] and solid–liquid equilibria (SLE)[26]) by thermodynamic laws. This enables training the approach on the corresponding thermodynamic data of different types so that many data sources are accessible. The model can be written as:

$$\ln \gamma_{ij}(T,x_i) = f_{\mathrm{UNIQUAC}}(T,x_i,P_i,P_j,U_{ii},U_{jj},U_{ij}) \tag{2}$$

where the function f_{UNIQUAC} contains the UNIQUAC equation, which is defined in eqn (S.1)–(S.7) of the ESI.† Here, T is the temperature and x_i is the mole fraction of component i in the binary mixture. Eqn (2) contains two types of parameters: first, the geometric pure-component parameters P_i and P_j, which are reported for many components (*e.g.*, in the Dortmund Data Bank (DDB)[27]) or can be readily estimated (*e.g.*, with the approach described in connection with the development of the UNIFAC method[7]); and second, the pair-interaction energies, where we distinguish the like interactions U_{ii} and U_{jj}, which are also pure-component parameters, and the unlike interactions U_{ij} ($i \neq j$), which are binary parameters. It follows from the physical interpretation of U_{ij} as pair-interaction energies that $U_{ij} = U_{ji}$.

Furthermore, it follows from the derivation of UNIQUAC that all model parameters are independent of temperature and concentration. We use this assumption throughout the present work but note that some authors work with temperature-dependent parameters. Further, as already mentioned in the introduction, it is common in the literature not to use the symmetric pair-interaction energies $U_{ij} = U_{ji}$ as parameters, but rather the two parameters $\Delta U_{ij} \neq \Delta U_{ji}$, which are calculated by:

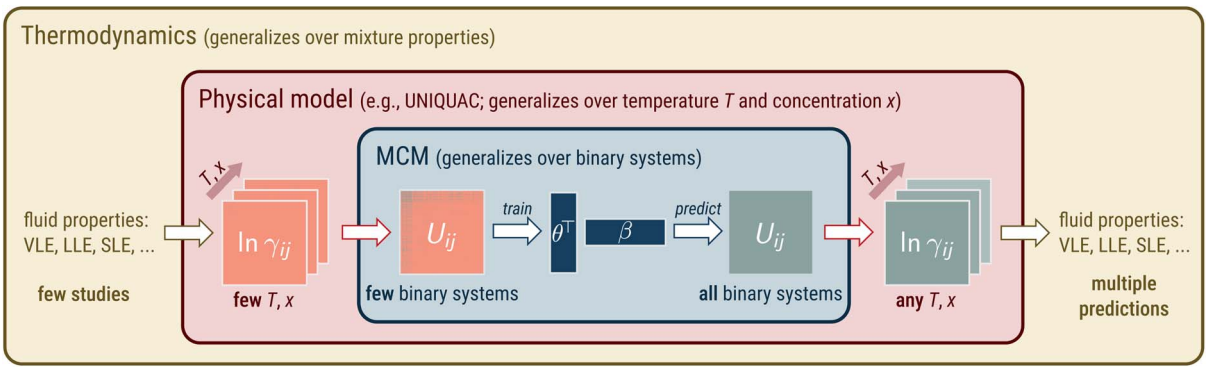

Fig. 1 Illustration of embedding an MCM into a physical model of mixtures (here: the lattice model UNIQUAC). Blue part: application of an MCM to pair-interaction energies U_{ij}. Red part: the physical model relates U_{ij} to temperature- and concentration-dependent activity coefficients $\ln \gamma_{ij}$. Yellow part: $\ln \gamma_{ij}$ are directly related to observable mixture properties (*e.g.*, vapor–liquid equilibria (VLE),[24] liquid–liquid equilibria (LLE),[25] solid–liquid equilibria (SLE)[26]) by thermodynamic laws.

 © 2022 The Author(s). Published by the Royal Society of Chemistry

$$\Delta U_{ij} = U_{ij} - U_{jj}; \quad \Delta U_{ji} = U_{ij} - U_{ii} \tag{3}$$

The correlations between ΔU_{ij} and ΔU_{ji} defined by eqn (3) are usually ignored and the two parameters are fitted *independently* for each binary system *i–j* under consideration, simply to increase the flexibility of the model.§ Fitting ΔU_{ij} and ΔU_{ji} independently to data for binary systems of $N \geq 3$ components will, in general, lead to results that cannot be reconciled with eqn (3), as demonstrated in detail in the ESI.† For MCM-UNIQUAC, we only consider the truly physical pair-interaction energies U_{ij}.

UNIQUAC allows generalizing (*i.e.*, interpolating and extrapolating) over temperatures and concentrations, but not over mixture components. Applying UNIQUAC to a mixture for which no parameters are available requires at least some mixture data points for determining the pair-interaction parameters. This is a severe restriction, as experimental data are often unavailable, especially for mixtures. To overcome this limitation, we introduce MCM-UNIQUAC to extend UNIQUAC so that it can generalize over temperatures and concentrations *and* over binary systems.

The proposed generalization over binary systems is achieved by an MCM and illustrated in the central (blue) panel in Fig. 1. We thereby model the unlike pair-interaction energies U_{ij} as a sum of dot products:

$$U_{ij} = \boldsymbol{\theta}_i\boldsymbol{\beta}_j + \boldsymbol{\theta}_j\boldsymbol{\beta}_i,\ i \neq j \tag{4}$$

where $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_i$ as well as $\boldsymbol{\theta}_j$ and $\boldsymbol{\beta}_j$ are the feature vectors for the components *i* and *j*, respectively, and the right-hand side of eqn (4) is constructed in such a way that the physical constraint $U_{ij} = U_{ji} \forall i,j$ is enforced, resulting in a *symmetric* matrix U of the pair-interaction energies. Besides the feature vectors, the like pair-interaction energies U_{ii} and U_{jj} are considered as parameters of MCM-UNIQUAC. Note that all parameters of MCM-UNIQUAC are *component-specific* (P_i, P_j, $\boldsymbol{\theta}_i$, $\boldsymbol{\beta}_i$, U_{ii}, U_{jj}), but they are, except for P_i and P_j, inferred from *mixture data*. After fitting these parameters, a *complete set* of pair-interaction energies for all conceivable binary combinations *i–j* of all considered components is obtained from eqn (4).

MCM-UNIQUAC was trained end-to-end on a set of measured logarithmic temperature- and concentration-dependent activity coefficients in binary mixtures $\ln \gamma_{ij}$ (red panel of Fig. 1). We used data from the DDB[27] here; in specific, we used $\ln \gamma_{ij}$ derived from binary vapor–liquid equilibrium (VLE) data using the extended Raoult's law, *cf.* eqn (S.14) in the ESI,† which we augmented with temperature-dependent data on binary activity coefficients at infinite dilution $\ln \gamma_{ij}^{\infty}$. In total, we obtained a set of 363 181 experimental data points for $\ln \gamma_{ij}$ for 12 199 different binary systems *i–j* involving 1146 distinct components *i*, *j* at varying concentrations and temperatures ranging from 183 K to 638 K. The considered $N = 1146$ components result in $N(N-1)/2 = 656\,085$ possible different binary systems. Experimental data are only available for 12 199 of these systems, *i.e.*, data are available for less than 2% of all systems and, consequently, only less than 2% of these systems can be modeled with UNIQUAC in the conventional way. More details on the data set are given in the ESI.†

The systems for which data are available were divided into three sets: 80% were used for training the model (training set), 10% were used for setting the model's hyperparameters (validation set), and 10% were used for testing the predictions (test set). We trained our model using the probabilistic programming language Stan[28] and resorted to Variational Inference[29–31] for performing approximate Bayesian inference. Details on the random data split, the model training (including the source code to run the model in Stan), and the hyperparameter selection are given in the ESI.†

After the training, MCM-UNIQUAC can predict temperature- and concentration-dependent activity coefficients in any binary and multicomponent mixture of the considered 1146 components. The activity coefficients can, in turn, be used for predicting observable mixture properties, such as VLE or other phase equilibria (yellow panel in Fig. 1). We demonstrate the predictive capacity of MCM-UNIQUAC by considering the data from the test set in the following section.

Results and discussion

The results that were obtained with MCM-UNIQUAC are shown as bars in Fig. 2 (left), where the mean absolute error (MAE) is reported both for the systems from the training set and those from the test set; in Fig. S2 in the ESI,† we show the respective results for the mean squared error (MSE). As expected, the error metrics for the test set are larger (worse) than those for the training set, but in both cases the overall agreement between the results of MCM-UNIQUAC and the experimental data is remarkable. We demonstrate this in Fig. 2 (right) by comparison with the best available physical method for the prediction of activity coefficients, the group-contribution model modified UNIFAC (Dortmund),[10,11] which is called 'UNIFAC' in the following for brevity. Unfortunately, in contrast to MCM-UNIQUAC, UNIFAC cannot be applied to all systems for which data are available (denoted as 'complete horizon' in Fig. 2 (left)) because multiple group-interaction parameters are missing. Hence, for a fair comparison of both methods, only those subsets of the training set and of the test set for which UNIFAC could be applied were used; this 'UNIFAC horizon' covers 7578 of 9759 systems from the training set and 961 of 1220 systems from the test set.

Furthermore, as baselines, the scores obtained for the different sets by directly fitting UNIQUAC parameters to all available data points are marked as lines in Fig. 2. We thereby considered two variants: first, UNIQUAC was used in the usual manner by fitting the two binary parameters ΔU_{ij} and ΔU_{ji} *individually* to the data for each system *i–j*; this procedure is denoted as 'UNIQUAC (ΔU)' in the following and the respective results are shown as dotted lines in Fig. 2. Second, we fitted the symmetric pair-interaction energies $U_{ij} = U_{ji}$ to all data points; this procedure is denoted as 'UNIQUAC (U)' in the following and the respective results are shown as dashed lines in Fig. 2. Note that considering the full set of 12 199 binary systems for which experimental data are available, there are 24 398 parameters in

© 2022 The Author(s). Published by the Royal Society of Chemistry

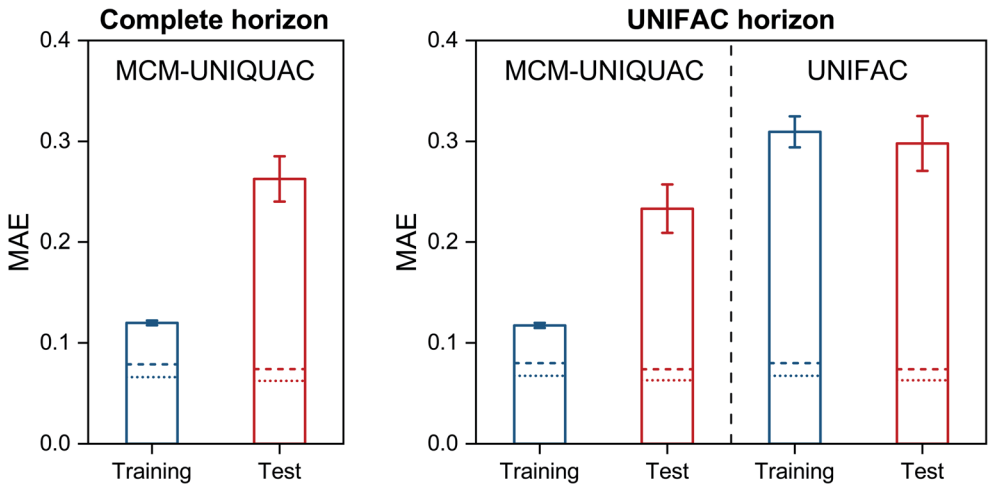

Fig. 2 Mean absolute error (MAE) of MCM-UNIQUAC on the training and test set (left) and comparison to UNIFAC based only on those systems that can also be modeled with UNIFAC (right). Bars indicate the results of MCM-UNIQUAC and UNIFAC, and lines denote the baselines obtained by directly fitting UNIQUAC pair-interaction parameters (ΔU_{ij}, dotted) or pair-interaction energies (U_{ij}, dashed) to all available data points. Error bars denote standard errors of the means.

UNIQUAC (ΔU), whereas there are only 13 345 parameters in UNIQUAC (U), of which 1146 are pure-component parameters U_{ii}.

Let us consider the scores of the baselines UNIQUAC (ΔU) and UNIQUAC (U) first. They simply indicate how well the data can, in principle, be described with the physical model, *i.e.*, UNIQUAC here. The first astonishing message is that UNIQUAC (U) works almost as well as UNIQUAC (ΔU), even though the latter has almost twice as many parameters, *cf.* above. We have therefore based our hybrid model MCM-UNIQUAC on the more physical UNIQUAC (U).

Second, we observe in Fig. 2 that MCM-UNIQUAC is in general flexible enough for describing the data well, *cf.* the relatively small differences between the scores of MCM-UNIQUAC on the training set (blue bars) and the baseline scores (lines).

Third, by comparing the scores for the test set (red bars), we find that MCM-UNIQUAC clearly outperforms UNIFAC in accurately predicting $\ln \gamma_{ij}$ in both MAE and MSE (*cf.* Fig. S.2 in the ESI†). This is particularly remarkable since UNIFAC has been fitted to most of these data points, whereas no data point from the test set was used for training or validation of MCM-UNIQUAC. An alternative representation of the results for the test set is shown in histogram plots in Fig. S.3 and S.4 in the ESI,† which underpin the superior performance of MCM-UNIQUAC compared to UNIFAC. Similar results are found by considering COSMO-SAC-dsp[32,33] as baseline, as shown in Fig. S.5 in the ESI.† Additionally, we discuss the influence of the number of training data points on the performance of MCM-UNIQUAC in Fig. S.6 and Table S.1 in the ESI.†

We note that also a version of MCM-UNIQUAC based on UNIQUAC (ΔU), *i.e.*, based on the asymmetric pair-interaction parameters, can be trained if *and only if* an end-to-end training on $\ln \gamma_{ij}$ is performed; the resulting performance is, however, slightly worse than that of the model discussed above as we demonstrate in Fig. S.7 in the ESI.†

MCM-UNIQUAC can not only be used for predicting activity coefficients, but also for predictions of phase equilibria and many other thermodynamic properties of any mixture consisting of the considered 1146 components. We demonstrate this in Fig. 3, which shows the results of the prediction of isobaric vapor–liquid equilibrium (VLE) phase diagrams for eight binary systems based on the extended Raoult's law, *cf.* eqn (S.14) in the ESI.† All eight systems were chosen randomly from the test set, *i.e.*, not a single data point for these systems was used for training MCM-UNIQUAC or setting its hyperparameters. However, the selection was carried out in such a way as to cover a wide range of different phase behaviors, ranging from high-boiling azeotrope (top left) to heteroazeotrope (bottom right). For details on the selection of the systems and the prediction of the VLE phase diagrams with MCM-UNIQUAC, we refer to the ESI.†

For all eight binary systems, excellent agreement between the predicted phase diagram and the experimental data is found, both qualitatively and quantitatively. Similarly, liquid–liquid and solid–liquid phase diagrams and other thermodynamic properties like excess enthalpies can be predicted with MCM-UNIQUAC. Moreover, the physical foundation on which the hybrid approach MCM-UNIQUAC builds, namely learning and predicting pair-interaction energies between components on the hypothetical lattice from data for binary mixtures, even allows extrapolations from binary to multicomponent mixtures; MCM-UNIQUAC thereby does not require any data of multicomponent systems for training. As an example, isobaric VLE phase diagrams for two ternary systems are shown in Fig. 4. These systems were selected such that all constituent binary subsystems were neither part of the training set nor of the validation set.

 © 2022 The Author(s). Published by the Royal Society of Chemistry

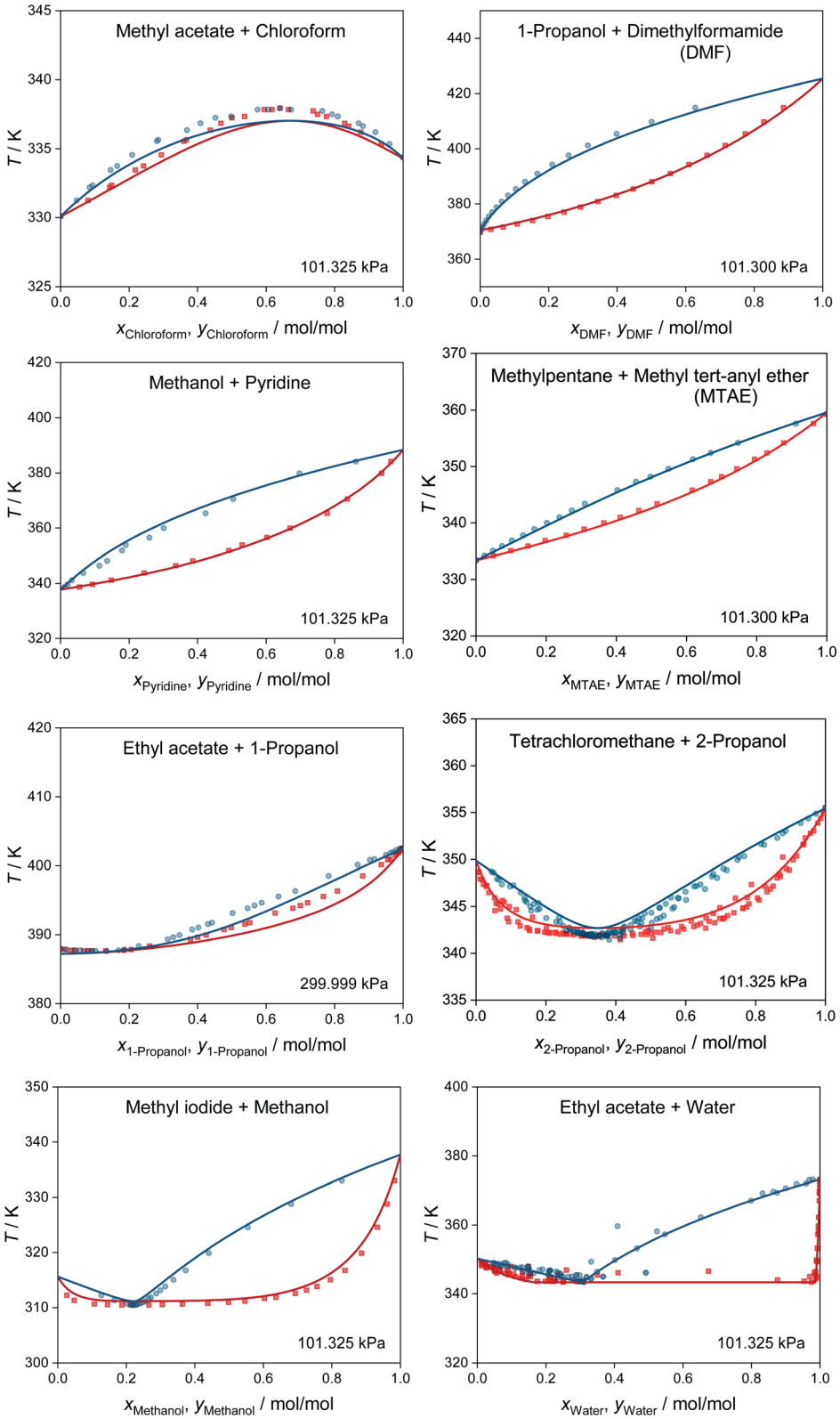

Fig. 3 Prediction of isobaric vapor–liquid phase diagrams for binary systems from the test set with MCM-UNIQUAC (lines) and comparison to experimental data from the DDB[27] (symbols). No data on any of the depicted systems were used for training MCM-UNIQUAC or setting the hyperparameters. Blue: dew point curves. Red: bubble point curves.

© 2022 The Author(s). Published by the Royal Society of Chemistry

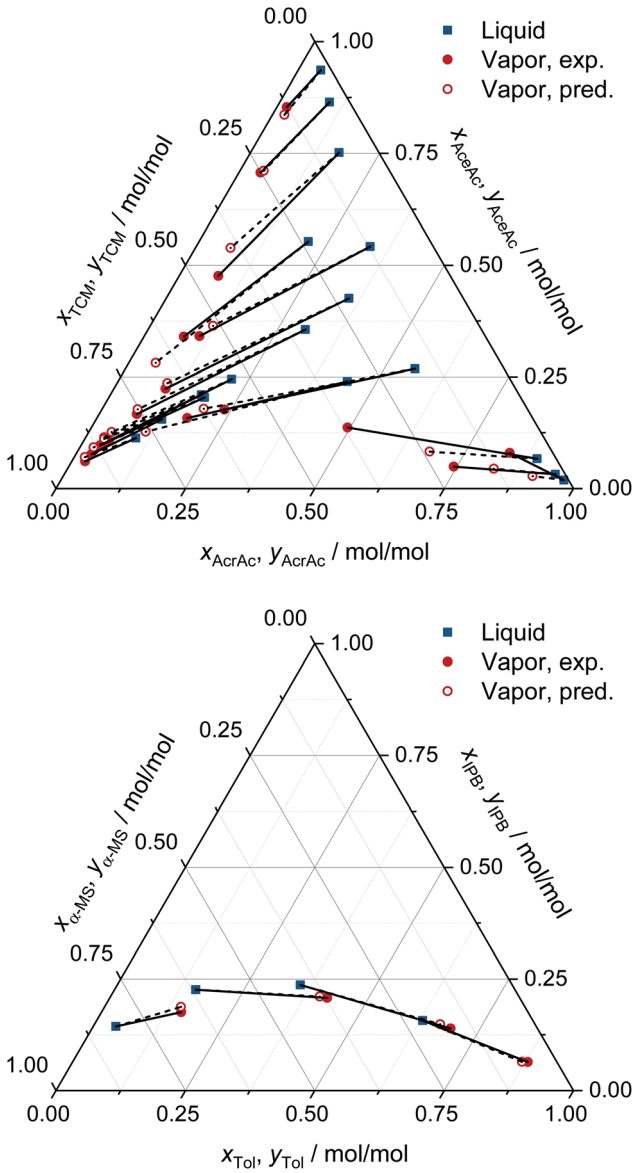

Fig. 4 Prediction of the VLE in ternary systems at constant pressure with MCM-UNIQUAC and comparison to experimental data (exp.) from the DDB.[27] The pressure and the composition of the liquid phase were specified, the composition of the corresponding vapor phase was predicted (pred.). Top: acrylic acid (AcrAc) + acetic acid (AceAc) + tetrachloromethane (TCM) at 100 kPa. Bottom: toluene (Tol) + isopropylbenzene (IPB) + α-methyl styrene (α-MS) at 101 kPa. No data on any of the systems and any of the binary subsystems were part of the training set or the validation set of MCM-UNIQUAC.

For each data point, the pressure and the composition of the liquid phase (blue symbols in Fig. 4) were specified. The associated composition of the vapor phase in equilibrium was predicted with MCM-UNIQUAC (open red symbols) and compared to the experimental composition of the vapor phase as reported in the DDB[27] (filled red symbols). We observe an excellent agreement.

In the ESI,† we additionally provide a complete set of the pair-interaction energies U_{ij}, including the respective model uncertainties in the form of standard deviations, for all 656 085 binary systems of the components considered here (the pair-interaction parameters ΔU_{ij} of the commonly used version of UNIQUAC can easily be calculated from these by applying eqn (3)). This set of U_{ij} was obtained from an end-to-end training of MCM-UNIQUAC on the *entire* available data set of all 12 199 systems. If predictions for any system of the components considered in this work are required, this parameter set should be used.

Conclusions

In the present work, we describe a novel hybrid approach for predicting thermodynamic properties of mixtures, which combines methods from machine learning (ML) with physical modeling. The basic idea is to predict the pair interactions between components in mixtures using matrix completion methods (MCMs). The approach is generic, it can be applied to any mixture property, and any physical model based on pair interactions can be used. As an example, we combined an MCM with UNIQUAC, a well-known lattice model of the Gibbs excess energy G^E.

We trained our hybrid approach, MCM-UNIQUAC, on experimental data on activity coefficients in binary mixtures of 1146 components from the Dortmund Data Bank (DDB). Out of the possible 656 085 binary systems that can be formed from these components, suitable experimental data were only available for 12 119 systems, corresponding to only 2%. In its basic form, UNIQUAC can only be applied to this subset, as pair-interaction parameters need to be fitted to data of the respective systems. In contrast, MCM-UNIQUAC yields, in principle, predictions for all 656 085 binary systems; we demonstrate the predictive capacity of the model based on the systems for which experimental data are available by withholding the test data points during the training of the model. Moreover, by virtue of being a model based on pair interactions, MCM-UNIQUAC can be applied not only to binary systems but also to any multi-component system that can be formed from the considered 1146 components; and it also yields predictions for any composition and temperature.

We compared the quality of these predictions to those from the best available physical model for this purpose, the group-contribution model modified UNIFAC (Dortmund). However, due to missing group-interaction parameters, the public version of UNIFAC can at present only be applied to 9502 of the 12 119 binary mixtures for which data were available. We show that MCM-UNIQUAC outperforms UNIFAC, even on a test set whose data were not used for training MCM-UNIQUAC, whereas most of these data were probably used for training UNIFAC.

In its commonly used version, UNIQUAC has two binary parameters $\Delta U_{ij} \neq \Delta U_{ji}$ (with $i \neq j$). It is known that they are highly correlated and hard to interpret physically. In this work, we have discovered that going back to the basic idea of the lattice model and working directly with symmetric pair-interaction energies $U_{ij} = U_{ij}$, *i.e.*, with only a single binary parameter for each binary system, gives almost the same quality of the description of the phase equilibria in the studied binary systems as working with the common version of UNIQUAC with

 © 2022 The Author(s). Published by the Royal Society of Chemistry

two binary parameters, which is highly remarkable. It can be assumed that the pair-interaction energies U_{ij}, which are now available for 656 085 binary systems and 1146 pure components, can be interpreted physically. As they are basically mean-field energies, this is probably only a first step, and deeper insights may be expected from applying the MCM in connection with more sophisticated mixture models that account for different types of interactions.

MCM-UNIQUAC combines the strengths of ML and physical modeling – the MCM enables generalization over (discrete) components, allowing predictions for unstudied binary systems, whereas UNIQUAC generalizes over (continuous) conditions and enables the extrapolation to multicomponent mixtures. As demonstrated in examples, MCM-UNIQUAC allows the direct prediction of phase diagrams as exemplified for binary and ternary vapor–liquid phase diagrams. Furthermore, our approach can be retrained easily whenever additional data points become available; this is in contrast to established group-contribution methods, which require extensive parameter tuning. Also, MCM-UNIQUAC does not rely on expensive calculations as required by established quantum-chemical prediction methods. Finally, by exploiting the fact that basically all established thermodynamic models for mixture properties rely on the description of pair interactions, our approach can be further developed in many directions, *e.g.*, by combining MCMs with equations of state and even group-contribution methods.

Data availability

All data used in the present work are available in the Dortmund Data Bank.[27] The source code to run the model in Stan as well as the final parameter set are available in the ESI.†

Author contributions

FJ, SM, and HH have conceptualized this work. FJ and RB developed, trained, and evaluated the model. All authors contributed to developing the methodology that was applied. FJ, SM, and HH wrote the draft of the manuscript and all authors finalized it.

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 Bundesministerium für Wirtschaft und Energie (BMWi) in the frame of the project 'KEEN'. This material is in part based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0021. Stephan Mandt furthermore acknowledges support by the National Science Foundation under the NSF CAREER award 2047418 and Grants 1928718, 2003237, and 2007719, the Department of Energy under grant DE-SC0022331, as well as unrestricted gifts from Intel, Disney, and Qualcomm. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or NSF.

Notes and references

§ For a set of N components, there are $N(N-1)/2$ binary systems and, following the common usage, $N(N-1)$ binary UNIQUAC parameters ΔU_{ij}. Using the pair-interaction energies U_{ij} directly as parameters, there are $N(N+1)/2$ parameters, of which N are pure-component parameters. Hereby $N(N-1) > N(N+1)/2$ for $N \geq 3$.

1 *CAS Registry*, 2021, **https://www.cas.org/cas-data/cas-registry**.
2 D. S. Abrams and J. M. Prausnitz, *AIChE J.*, 1975, **21**, 116–128.
3 G. Maurer and J. Prausnitz, *Fluid Phase Equilib.*, 1978, **2**, 91–99.
4 H. Renon and J. M. Prausnitz, *AIChE J.*, 1968, **14**, 135–144.
5 B. Poling, J. Prausnitz and J. O'Connell, *The Properties of Gases and Liquids*, McGraw-Hill, 2001.
6 H. C. Carlson and A. P. Colburn, *Ind. Eng. Chem. Res.*, 1942, **34**, 581–589.
7 A. Fredenslund, R. L. Jones and J. M. Prausnitz, *AIChE J.*, 1975, **21**, 1086–1099.
8 A. Fredenslund, J. Gmehling and P. Rasmussen, *Vapor-Liquid Equilibria using UNIFAC, a Group-Contribution Method*, Elsevier, Amsterdam, The Netherlands, 1977.
9 A. Fredenslund, *Fluid Phase Equilib.*, 1989, 135–150.
10 U. Weidlich and J. Gmehling, *Ind. Eng. Chem. Res.*, 1987, **26**, 1372–1381.
11 D. Constantinescu and J. Gmehling, *J. Chem. Eng. Data*, 2016, **61**, 2738–2748.
12 T. Holderbaum and J. Gmehling, *Fluid Phase Equilib.*, 1991, **70**, 251–265.
13 S. Horstmann, A. Jabłoniec, J. Krafczyk, K. Fischer and J. Gmehling, *Fluid Phase Equilib.*, 2005, **227**, 157–164.
14 A. Klamt, *J. Phys. Chem.*, 1995, **99**, 2224–2235.
15 A. Klamt, V. Jonas, T. Bürger and J. C. Lohrenz, *J. Phys. Chem. A*, 1998, **102**, 5074–5085.
16 A. Klamt, F. Eckert and W. Arlt, *Annu. Rev. Chem. Biomol. Eng.*, 2010, **1**, 101–122.
17 H. Grensemann and J. Gmehling, *Ind. Eng. Chem. Res.*, 2005, **44**, 1610–1624.
18 F. Jirasek, R. A. S. Alves, J. Damay, R. A. Vandermeulen, R. Bamler, M. Bortz, S. Mandt, M. Kloft and H. Hasse, *J. Phys. Chem. Lett.*, 2020, **11**, 981–985.
19 F. Jirasek, R. Bamler and S. Mandt, *Chem. Commun.*, 2020, **56**, 12407–12410.
20 J. Damay, F. Jirasek, M. Kloft, M. Bortz and H. Hasse, *Ind. Eng. Chem. Res.*, 2021, **60**, 14564–14578.
21 F. Jirasek and H. Hasse, *Fluid Phase Equilib.*, 2021, **549**, 113206.
22 Y. Koren, R. Bell and C. Volinsky, *Computer*, 2009, **42**, 30–37.
23 G. Takács, I. Pilászy, B. Németh, D. Tikk, Investigation of various matrix factorization methods for large

© 2022 The Author(s). Published by the Royal Society of Chemistry

recommender systems, *IEEE International Conference on Data Mining Workshops*, 2008, pp. 553–562.
24 A. Fredenslund, *Vapor-liquid Equilibria using UNIFAC: A Group-contribution Method*, Elsevier, 2012.
25 T. Anderson and J. Prausnitz, *Ind. Eng. Chem. Process Des. Dev.*, 1978, **17**, 561–567.
26 J. G. Gmehling, T. F. Anderson and J. M. Prausnitz, *Ind. Eng. Chem. Fundam.*, 1978, **17**, 269–273.
27 U. Onken, J. Rarey-Nies and J. Gmehling, *Int. J. Thermophys.*, 1989, 739–747.
28 B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li and A. Riddell, *J. Stat. Software*, 2017, 1–32.
29 D. M. Blei, A. Kucukelbir and J. D. McAuliffe, *J. Am. Stat. Assoc.*, 2017, **112**, 859–877.
30 C. Zhang, J. Bütepage, H. Kjellström and S. Mandt, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, **41**, 2008–2026.
31 A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman and D. M. Blei, *J. Mach. Learn. Res.*, 2017, 1–45.
32 C.-M. Hsieh, S.-T. Lin and J. Vrabec, *Fluid Phase Equilib.*, 2014, **367**, 109–116.
33 I. H. Bell, E. Mickoleit, C.-M. Hsieh, S.-T. Lin, J. Vrabec, C. Breitkopf and A. Jäger, *J. Chem. Theory Comput.*, 2020, **16**, 2635–2646.

© 2022 The Author(s). Published by the Royal Society of Chemistry