Similarity-Informed Matrix Completion Method for Predicting Activity Coefficients

Nicolas Hayer, Thomas Specht, Justus Arweiler, 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

Accurate prediction of thermodynamic properties of mixtures, like activity coefficients, is essential for designing and optimizing chemical processes. While established physics-based methods face limitations in prediction accuracy and scope, emerging machine learning approaches, such as matrix completion methods (MCMs), offer promising alternatives. However, their performance can suffer in data-sparse regions. To address this issue, we propose a novel hybrid MCM for predicting activity coefficients at infinite dilution at 298 K that uses not only experimental training data but also includes synthetic training data from two sources: predictions obtained from the physics-based modified UNIFAC (Dortmund) and from a similarity-based approach developed in previous work. The resulting hybrid method combines the broad applicability of MCMs with the precision of the similarity-based approach, resulting in a more robust prediction framework that excels even in regions with limited data. Additionally, our analysis provides valuable insights into how different types of training data affect prediction accuracy. When experimental data are sparse, incorporating synthetic training data from modified UNIFAC (Dortmund) and the similarity-based approach significantly improves the performance of the MCMs. Conversely, even with abundant experimental

data, high accuracy is only achieved if the training set includes mixtures similar to those of interest.

Introduction

Reliable prediction methods for thermodynamic properties of mixtures are essential for the design and optimization of many processes in chemistry and chemical engineering. A particularly important property is the activity coefficient, which describes the deviation from the ideal mixture and is widely used for modeling reaction and phase equilibria in mixtures. Common physical prediction methods for activity coefficients include group-contribution (GC) models, such as UNIFAC[1,2] and modified UNIFAC (Dortmund) (mod. UNIFAC),[3,4] as well as models based on quantum chemistry like COSMO-RS[5–7] and COSMO-SAC-dsp.[8]

While these physical methods are well established, numerous alternative methods for predicting activity coefficients have recently been developed based on machine learning (ML). Some of them learn exclusively from available experimental data,[9–12] others are hybrid methods that combine the strengths of ML with those of physics.[13–27] An example for such a hybrid model is the so-called Whisky method,[13] which belongs to the class of matrix completion methods (MCMs). It was developed for the prediction of activity coefficients of solutes i at infinite dilution in solvents j γ_{ij}^{∞} at 298 K in unstudied binary mixtures. The MCM exploits the fact that the properties of binary mixtures can be stored conveniently in matrices, which are only sparsely occupied by experimental data in all relevant cases. In the Whisky method, a Bayesian approach is used to complete the matrix, exploiting similarities between components that are learned in the training. This training involves two steps, named in analogy to Whisky production: a distillation step, in which information from mod. UNIFAC predictions is distilled into prior knowledge, followed by a maturation step, in which the model is refined using experimental data. This hybrid approach, which combines the physics-based mod. UNIFAC with a data-driven MCM, has demonstrated superior performance compared

to a purely data-driven MCM and the mod. UNIFAC model alone.[13] For more technical details on the Whisky method, see Ref.[13]

As an alternative to the Whisky method, we have recently developed the so-called similarity-based method (SBM)[28] for predicting activity coefficients at infinite dilution. The SBM is based on the idea that similar mixtures should have similar properties, following the ancient alchemistic knowledge "similia similibus solvuntur". Consequently, the SBM uses the available experimental data for activity coefficients in similar mixtures to predict the activity coefficients in an unstudied mixture of interest by simply averaging the data for similar mixtures. At its core, the SBM calculates similarities between mixture components to determine mixtures for which data are available and which are sufficiently similar to the (unstudied) mixture of interest so that the data can be used for the prediction.

It is clear that the accuracy and the range of applicability of the SBM are inversely correlated: the higher the demanded accuracy, the stricter the required similarity, and the lower the chance of finding sufficiently similar mixtures in a given data set. In the SBM, the same chemical descriptors as they are used in the COSMO models[5,29] are used for defining a similarity score. The SBM achieves a high prediction accuracy, often within typical experimental uncertainties, whenever sufficient similar data are available. For more details on the SBM, we refer to Ref.[28]

In this work, we propose a novel model combining the Whisky approach with the SBM. Specifically, we integrate synthetic data from the SBM in the training process of the Whisky method, basically doubling the amount of training data in Whisky's maturation step. Therefore, we have chosen to apply a strict similarity criterion in the SBM, leading to precise predictions (at the cost of the number of mixtures for which the SBM yields predictions). The novel model, which we call Blended Whisky, thereby combines the strengths of both underlying methods: the Whisky method enables a broad scope by filling the entire matrix of missing data, while the SBM contributes precise predictions, which act as a powerful substitute for actual experimental data, increasing the prediction accuracy. This synergy

leads to a more robust and accurate predictive framework that consistently outperforms its predecessors.

In addition to introducing the Blended Whisky method, this work discusses and emphasizes the implications of model assessment and training data design for ML models, particularly MCMs. By analyzing the correlations between the training data quantity and type (synthetic or experimental), the similarity between the mixtures of interest and those in the training set, and the overall model performance, we obtain fundamental insights for the efficient training of MCMs, laying the foundation for their advancement, particularly in data-sparse regions.

Development of the Blended Whisky Method

The Blended Whisky method developed here is a probabilistic model that calculates logarithmic activity coefficients at infinite dilution γ_{ij}^{∞} as follows:

$$\ln \gamma_{ij}^{\infty} = \boldsymbol{u}_i \cdot \boldsymbol{v}_j + \varepsilon_{ij} \tag{1}$$

where $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are feature vectors of the solute i and solvent j, respectively, and ε_{ij} is a random variable that captures experimental noise and model inaccuracies. The length of the feature vectors, i.e., the number of considered features per component is a hyperparameter and was set to $K = 4$ as in our previous work.[13] The solute- and solvent-specific feature vectors can be aggregated into two feature matrices, $\boldsymbol{U}$ and $\boldsymbol{V}$, representing the learned characteristics of all solutes and all solvents, respectively, in the data set.

Blended Whisky is a Bayesian method that, as such, incorporates three key probability distributions: the prior, the likelihood, and the posterior. The prior captures initial beliefs or assumptions about the solute and solvent features ($\boldsymbol{u}_i$ and $\boldsymbol{v}_j$) before observing any data. The likelihood represents the probability of observing the data ($\ln \gamma_{ij}^{\infty}$ here) given the component features, describing how these hidden features manifest themselves in the observable quantities. The goal of Bayesian inference is to compute the posterior distribution, which updates our beliefs about the features by combining prior information with the observed data. Consequently, the posterior is a probability distribution over the component features, which captures both information from the prior and the likelihood and from which final model parameters used for making predictions can be inferred. For more details on the Bayesian approach, we refer to our work on the Whisky method.[13]

In contrast to simple MCMs,[9] which infer the component features ($\boldsymbol{u}_i$ and $\boldsymbol{v}_j$) only from the available experimental data in a single step, the Blended Whisky method (analogously to the Whisky method[13]) is based on a two-step approach with different data sources. Fig. 1 shows a schematic of the training process of Blended Whisky.

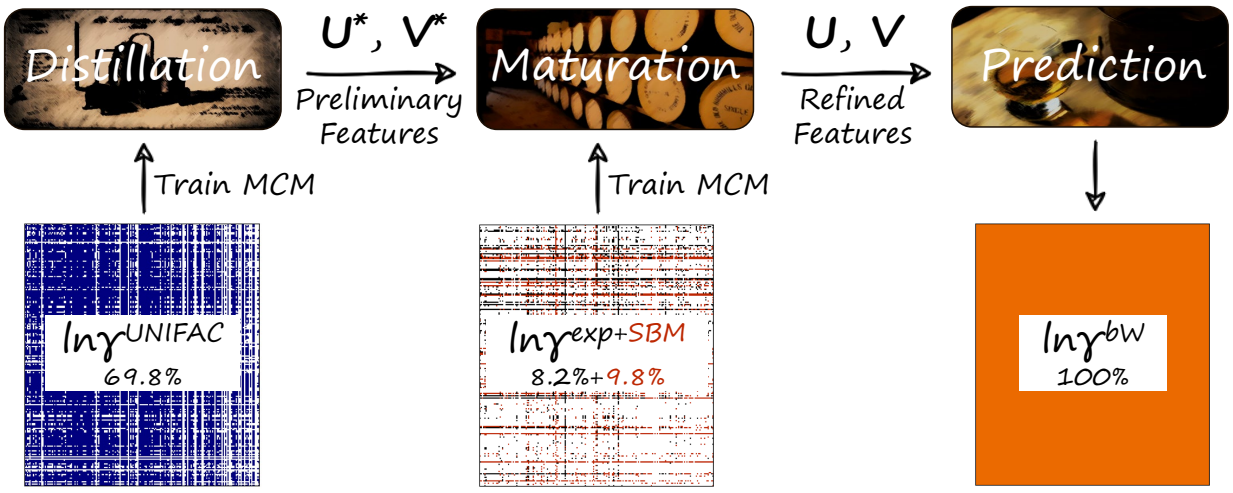

Figure 1: Schematic illustration of the Blended Whisky method. In the distillation step, an MCM is trained on 30,597 synthetic data points for $\ln \gamma_{ij}^{\infty}$ at 298 K obtained with mod. UNIFAC, constituting 69.8% of of all possible combinations of the considered 221 solutes and 198 solvents. The thereby fitted MCM parameters, stored in component feature matrices $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$, are used as informative prior for the maturation step, where a second MCM is trained on the available experimental data (3,568 data points, 8.2% of all possible matrix entries) augmented with SBM predictions (4,277 data points, 9.8% of all possible matrix entries). The final MCM parameters, stored in $\boldsymbol{U}$ and $\boldsymbol{V}$, are used for making predictions for unstudied $\ln \gamma_{ij}^{\infty}$.

In the first step, the distillation step, we train on synthetic data obtained from mod. UNIFAC, thereby distilling the knowledge captured in mod. UNIFAC's predictions and storing it in a first set of component features. These preliminary features will be the starting point for the second training step. In the distillation step, we use a rather broad, uninformative prior for each feature, specifically, a normal distribution with a mean of $\mu = 0$ and a standard deviation of $\sigma = 0.8$. As likelihood, we use a Cauchy distribution with scale parameter $\lambda = 0.15$ centered around the product of preliminary feature vectors:

$$p(\ln \gamma_{ij}^{\infty,\text{mod. UNIFAC}} \mid \boldsymbol{u}_i^*, \boldsymbol{v}_j^*) = \text{Cauchy}(\boldsymbol{u}_i^* \cdot \boldsymbol{v}_j^*,\ \lambda) \tag{2}$$

For all components for which mod. UNIFAC predictions were available during the distillation step, the resulting posterior mean is retained and is, in combination with a standard deviation of $\sigma = 0.5$, used as informative prior for the subsequent maturation step. For all other

components, i.e., those for which mod. UNIFAC can not predict the activity coefficients in the distillation step, a broader normal distribution with a mean of $\mu = 0$ and a standard deviation of $\sigma = 3$ is used as the prior for the maturation step.

In the maturation step, the features obtained from the distillation step are refined by training on experimental data ($\ln \gamma_{ij}^{\infty,\mathrm{exp}}$) and, in contrast to the Whisky method, synthetic training data obtained from the SBM, as described below. This way, the otherwise sparse experimental training set is substantially augmented with synthetic data of high quality.[28] The likelihood in the maturation step follows a Cauchy distribution with a scale parameter of $\lambda = 0.15$ for the experimental data and $\lambda = 0.2$ for the SBM data.

The synthetic data used for augmenting the training data in the maturation step of the Blended Whisky method were obtained using the SBM approach from our prior work,[28] which is based on a similarity score S derived from quantum-chemically calculated σ-profiles. The SBM makes predictions for $\ln \gamma_{ij}^{\infty}$ by averaging the experimental data from mixtures that are similar to the one of interest. Thereby, a *similar mixture* is defined as one with the same solute i (solvent j) and a different solvent n (solute m) that has a similarity score with the solvent of interest j (solute of interest i) higher than a threshold, which was set to 0.93, i.e., $S_{nj} > 0.93$ ($S_{mi} > 0.93$). The similarity score has values between 0 (no similarity) and 1 (full similarity). The choice of the threshold value $\xi = 0.93$ indicates that we require a high degree of similarity, leading to reliable predictions of the γ_{ij}^{∞} with the SBM. The downside is that choosing a high value of ξ leads to the fact that there will be only a few mixtures for which sufficiently similar mixtures for which data exist can be found. In our case, the SBM with $\xi = 0.93$ yielded only additional results for 9.3% of the entries of the matrix. However, since only for 8.6% of the entries experimental data were available, the database for the maturation step of the Blended Whisky method could be more than doubled. Whenever an experimental value and an SBM prediction were available, the experimental value was used. For more technical details on the SBM, see Ref.[28]

The Blended Whisky method was trained on 30,597 synthetic data points from mod. UNI-

FAC in the distillation step, and 3,568 experimental data points along with 4,277 synthetic data points from the SBM in the maturation step. All experimental $\ln \gamma_{ij}^{\infty,\mathrm{exp}}$ data were taken from the Dortmund Data Bank (DDB);[30] in total, 221 different solutes and 198 different solvents were considered. These experimental data and the synthetic training data obtained from the SBM are identical to the ones in our previous work on the SBM; further details can be found in Ref.[28]

Variational inference (VI) was used to train Blended Whisky, transforming the inference problem into an optimization problem, since exact Bayesian inference is computationally infeasible due to the complexity of computing the posterior.[31] Specifically, VI was implemented using the probabilistic programming language Stan[32] and Gaussian mean-field VI.[33] The hyperparameters, namely, the standard deviations of the prior and the scale parameters of the likelihood, were determined through preliminary studies.

The performances of the studied MCMs (MCM-data, Whisky, and Blended Whisky) and the SBM were evaluated using a leave-one-out analysis. Thereby, for each binary mixture with available experimental data ($\ln \gamma_{ij}^{\infty,\mathrm{exp}}$), the corresponding data point was excluded from the training set. The remaining data were then used to generate predictions for the excluded mixture. This approach ensures that the predictions are independent of the experimental data for that particular mixture, providing a more rigorous test of the predictive performance of the method. All calculations were performed using Matlab.[34]

Results and Discussion

Overall Performance of Blended Whisky

In Fig. 2, the performance of the Blended Whisky method for predicting $\ln \gamma_{ij}^{\infty}$ at 298 K obtained by a leave-one-out analysis is shown in terms of the mean absolute error (MAE) and the mean squared error (MSE). It is compared to the performances of the building blocks of Blended Whisky, the SBM[28] and the Whisky method,[13] as well as that of a purely data-driven MCM (MCM-data)[9] and the physical benchmarks mod. UNIFAC,[4] COSMO-SAC,[35] and COSMO-SAC-dsp.[8] The results are plotted as a function of the number N of predictable data points in our experimental data set, containing $N_{\max} = 3,568$ data points. For COSMO-SAC and COSMO-SAC-dsp, the implementations by Bell et al.[29] were used. We provide the predictions of $\ln \gamma_{ij}^{\infty}$ with all considered methods that were used to create Fig. 2 as Excel file in the Supporting Information.

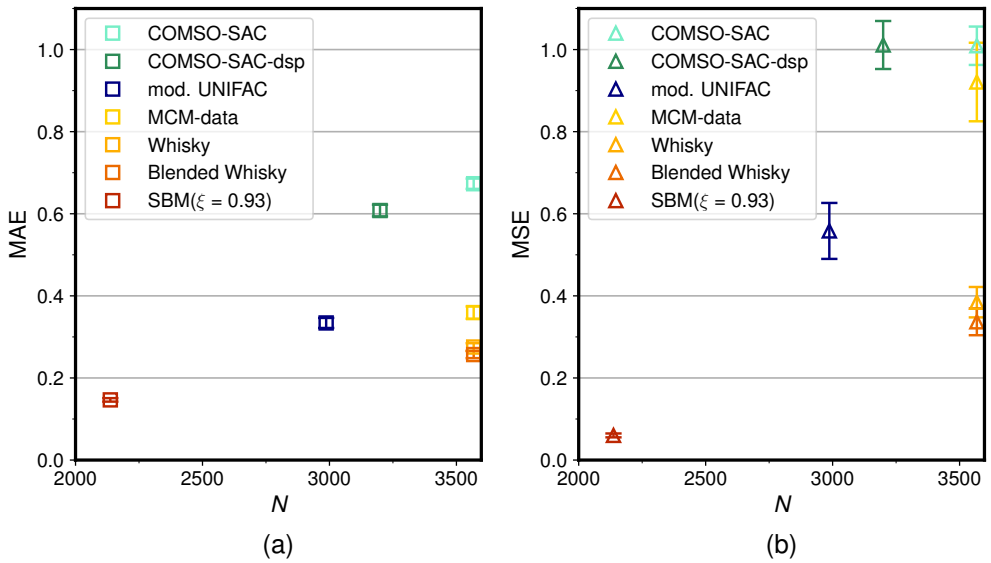

Figure 2: Mean absolute error (MAE, panel a) and mean squared error (MSE, panel b) of the predictions of $\ln \gamma_{ij}^{\infty}$ for the Blended Whisky method as a function of the number N of predictable data points in the data set. For comparison, the results of mod. UNIFAC, COSMO-SAC, COSMO-SAC-dsp, MCM-data, the Whisky method, and the SBM($\xi = 0.93$) are shown. Error bars denote standard errors of the means.

The SBM($\xi = 0.93$) achieves the highest prediction accuracy in both error scores; how-

ever, its scope is the smallest of all methods compared, as it is limited by the requirement for training data on mixtures similar to those of interest. Mod. UNIFAC offers broader applicability but at the cost of reduced accuracy. COSMO-SAC-dsp extends the scope further, albeit with even lower prediction accuracy. COSMO-SAC, MCM-data, Whisky, and Blended Whisky can predict all test data points. Among them, COSMO-SAC exhibits the poorest accuracy, while the Blended Whisky method achieves the highest accuracy (MAE = 0.26, MSE = 0.34), slightly outperforming the Whisky method (MAE = 0.28, MSE = 0.38). Although this overall reduction of the error scores may seem small, it was achieved without additional data compared to its predecessors, as both the SBM and Whisky are trained on the same database. Fig. S.1 in the Supporting Information further illustrates the predictive performance for all MCMs and physical benchmark methods via parity plots, highlighting the robust performance of the Blended Whisky method, which achieves accurate predictions even for strongly non-ideal mixtures.

Notably, both hybrid MCMs (Whisky and Blended Whisky) significantly reduce outliers in their predictions, as evidenced by their lower MSE values than MCM-data. In contrast, mod. UNIFAC exhibits some extreme outliers, which have even been excluded from the error score calculations in Fig. 2. Ref.[28] provides a detailed analysis of these outliers. Remarkably, despite including these outliers during the training of both the Whisky and Blended Whisky methods in their distillation steps, the MCMs still achieve high prediction accuracies, outperforming mod. UNIFAC, demonstrating the robustness of the hybrid models. Of the benchmark methods, only the SBM($\xi = 0.93$) achieves a lower MAE than the Blended Whisky method; however, this score is based on only the 60% of test data that the SBM($\xi = 0.93$) can predict.

Although this work focuses exclusively on γ_{ij}^{∞} at 298 K, the Blended Whisky method illustrates how data from multiple sources, each with its own uncertainties, can be combined into a unified training procedure. This lays the groundwork for extending the approach to more complex model architectures that incorporate temperature and concentration depen-

dencies (e.g., by embedding MCMs in conventional models of the Gibbs excess energy G^{E} as in our previous work[15,26]).

Influence of Training Data on Predictive Performance

In the following, we systematically examine two factors that influence the predictive performance of the MCMs. The first factor is the amount of experimental training data for each solute i and solvent j, corresponding to the number of data points in each row and column of the experimental data matrix. A binary mixture $i + j$ is thus characterized by the number of data points available for solute i and the number of data points available for solvent j. We only use the smaller of these two values, representing the less frequently measured component, and denote it as N_{min} in the following.

In Figure 3, we analyze how the prediction accuracy correlates with N_{min}. It shows the MAE and MSE of all studied methods (excluding the SBM($\xi = 0.93$) due to its limited scope) on a shared test data set containing only mixtures predictable by all methods. We differentiate between mixtures with rarely measured components ($N_{\mathrm{min}} \leq N_{\mathrm{cutoff}}$) and frequently measured components ($N_{\mathrm{min}} > N_{\mathrm{cutoff}}$), using a cutoff value of $N_{\mathrm{cutoff}} = 5$.

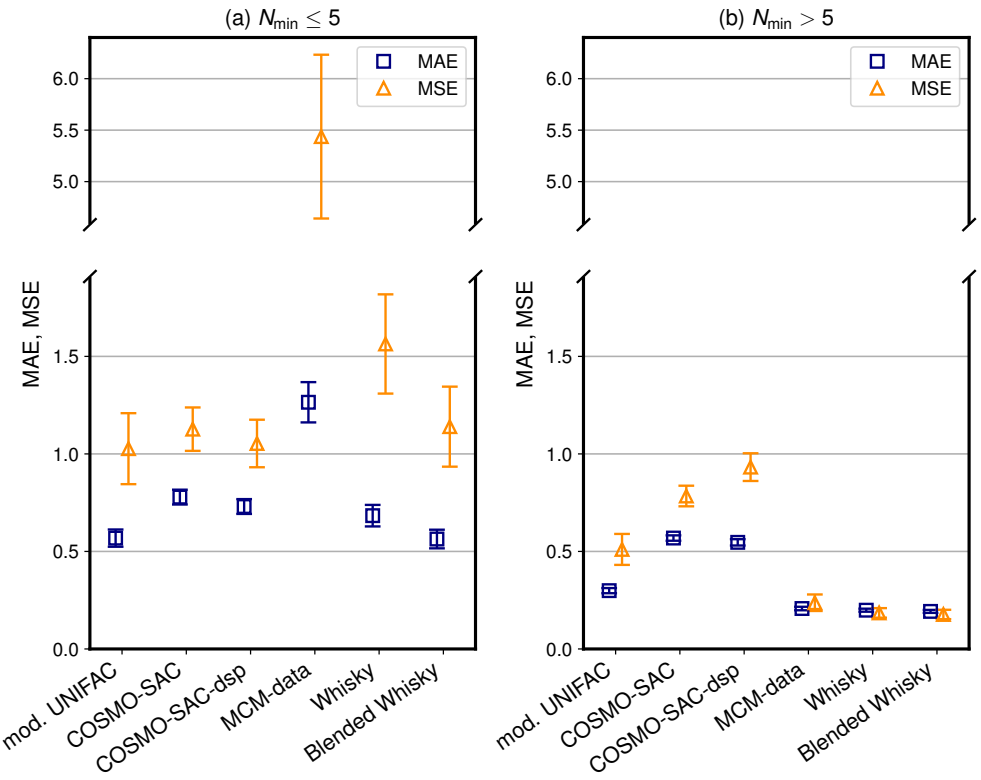

Figure 3: Mean absolute error (MAE) and mean squared error (MSE) of predictions of $\ln \gamma_{ij}^{\infty}$ for the Blended Whisky method in comparison to Whisky, MCM-data, mod. UNIFAC, COSMO-SAC, and COSMO-SAC-dsp. Error bars denote standard errors of the means. (a) Binary mixtures containing components with five or less studied mixtures in our data set ($N_{\text{min}} \leq 5$; 361 data points). (b) Binary mixtures where both components were studied in more than five mixtures in our data set ($N_{\text{min}} > 5$; 2,418 data points).

All methods depicted in Fig. 3 show higher average accuracy for mixtures with $N_{\text{min}} > 5$ as for mixtures with $N_{\text{min}} \leq 5$. The observed improvement in accuracy for the MCMs with increasing N_{min} is expected, given that these models are explicitly trained on the available experimental data in our database and rely on learning the component-specific features from these data; hence, more available data for a specific component can be expected to increase the quality of the learned features. In contrast, the physical benchmark methods (mod. UNIFAC, COSMO-SAC, and COSMO-SAC-dsp) have been evaluated 'as-is', using their published parameters, without any additional training or fine-tuning on our data set. Since the training sets for these methods are not fully disclosed, it is plausible that they may have been optimized or validated using data sets that overlap with frequently measured

components in our database, which could explain their improved performance for those mixtures. Alternatively, the mixtures in Fig. 3 may pose greater challenges for all methods due to higher molecular complexity or less predictable interaction effects.

For mixtures with only frequently studied components ($N_{\text{min}} > 5$), all three MCMs significantly outperform the three physical benchmark models. In contrast, for the mixtures with $N_{\text{min}} \leq 5$, mod. UNIFAC and the Blended Whisky method prove to be the best-performing approaches. Especially when comparing the performance of Blended Whisky to Whisky and MCM-data, the positive impact of integrating synthetic data from mod. UNIFAC *and* the SBM($\xi = 0.93$) into the training process becomes evident, particularly reducing the MSE.

While so far, we have only investigated the influence of data quantity for specific solutes or solvents, we now investigate the influence of whether data for *similar* mixtures are available on the performance of the MCMs. This distinction is not relevant to the physical benchmarks, as their performance does not depend on the presence of similar mixtures in the training data; hence, they are not further discussed in the following. We first explore the impact of similar mixtures in the training data on the performance of the different MCMs for mixtures with $N_{\text{min}} > 5$, as shown in Fig. 4. For this, we split the data set from Fig. 3b by categorizing each mixture according to the availability of at least one similar mixture (defined by a similarity score above a threshold of $\xi = 0.93$) in the training set. All mixtures meeting this condition can also be predicted using the SBM($\xi = 0.93$), which also requires at least one data point of a similar mixture for prediction. Consequently, the error scores of the SBM($\xi = 0.93$) are included in Fig. 4a.

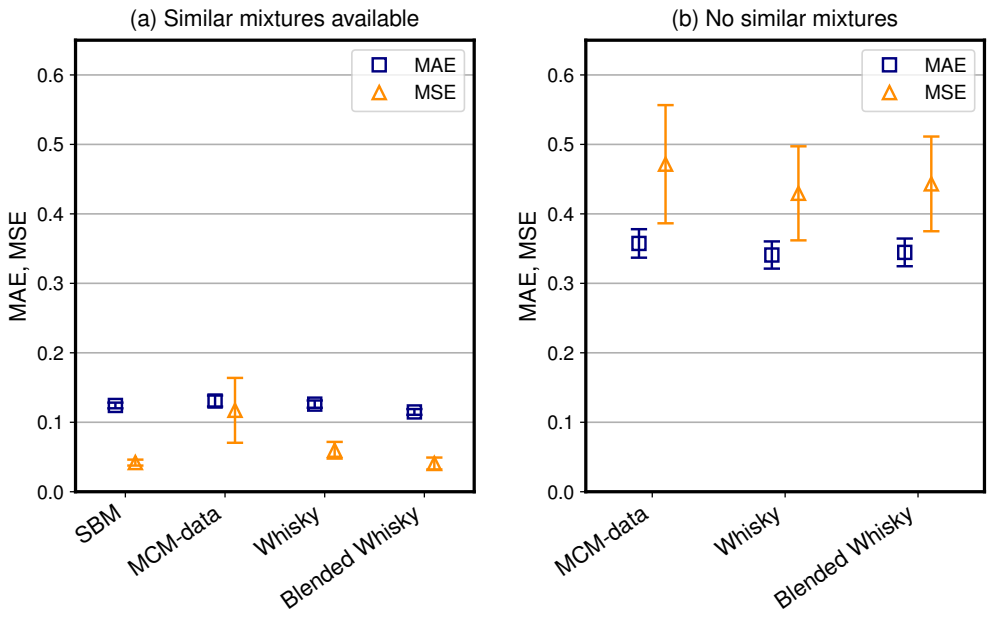

Figure 4: Mean absolute error (MAE) and mean squared error (MSE) of predictions of $\ln \gamma_{ij}^{\infty}$ for the Blended Whisky method in comparison to Whisky and MCM-data, focusing only on mixtures where both components were studied in more than five mixtures in our data set ($N_{\min} > 5$; 2,418 data points). Error bars denote standard errors of the means. (a) Binary mixtures for which experimental training data of similar mixtures are available (1,600 data points); similar mixtures are defined by a similarity score above a threshold of $\xi = 0.93$. An additional comparison with the SBM($\xi = 0.93$) is performed here. (b) Binary mixtures for which no experimental data with similar mixtures are available (818 data points).

Fig. 4 shows that similar mixtures in the training set significantly enhance the performance of all MCMs, emphasizing that a large amount of data alone is not sufficient for efficient training of data-driven methods; the similarity between the unstudied mixture of interest and the studied ones in the training set is, hence, a crucial factor for prediction accuracy. Notably, this similarity is derived here solely from the quantum-chemical descriptors of the pure components, rather than from the mixture data itself, making it an unbiased and powerful tool for categorizing mixtures and thereby revealing the hidden structure within the mixture property matrix.

The performance of the different MCMs is generally similar here, although MCM-data performs slightly worse than the hybrid MCMs (Whisky and Blended Whisky), especially in terms of MSE, for both cases in Fig. 4.

Fig. 5 shows the influence of available similar mixtures in the training set on the prediction accuracy for mixtures with rarely measured components ($N_{\min} \leq 5$; cf. Fig. 3a).

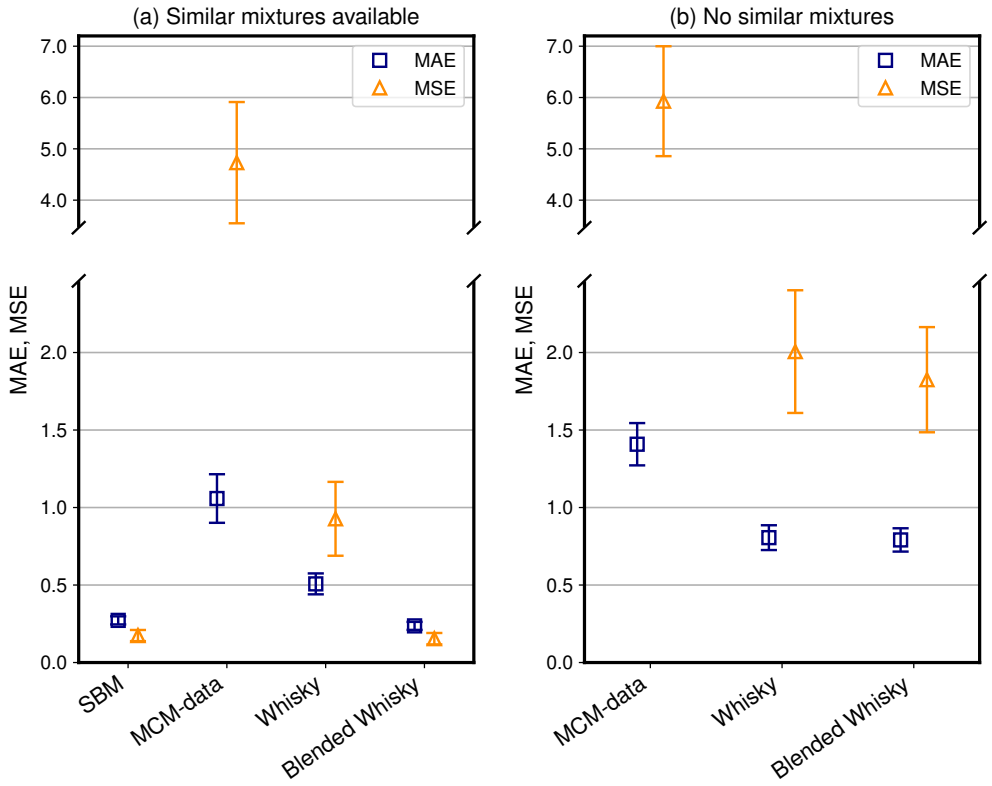

Figure 5: Mean absolute error (MAE) and mean squared error (MSE) of predictions of $\ln \gamma_{ij}^{\infty}$ for the Blended Whisky method in comparison to Whisky and MCM-data, focusing only on mixtures containing components with five or less studied mixtures in our data set ($N_{\min} \leq 5$; 361 data points). Error bars denote standard errors of the means. (a) Binary mixtures for which experimental training data of similar mixtures are available (148 data points); similar mixtures are defined by a similarity score above a threshold of $\xi = 0.93$. An additional comparison with the SBM($\xi = 0.93$) is performed here. (b) Binary mixtures for which no experimental data with similar mixtures are available (213 data points).

Fig. 5 shows again that the MCMs benefit strongly from the availability of training data for similar mixtures. Again, MCM-data performs worse than the hybrid methods and Blended Whisky yields better results than Whisky. This is especially true if similar mixtures are available, because in that case, the SBM($\xi = 0.93$) yields excellent results and effectively supports Blended Whisky via the accurate synthetic data in the distillation step. The Blended Whisky method (MAE=0.24, MSE=0.15) even slightly surpasses the

SBM($\xi = 0.93$) (MAE=0.27, MSE=0.17) while covering a significantly broader scope.

Conclusions

In this work, we have developed the Blended Whisky method, a hybrid matrix completion method (MCM) that successfully combines the strengths of two previously developed approaches, the Whisky method and the similarity-based method (SBM), to predict γ_{ij}^{∞} with high accuracy and broad scope. By incorporating synthetic data from the SBM as supplementary training data in the Whisky method's framework, the Blended Whisky method achieves superior performance compared to physical benchmarks and its predecessors, especially in data-sparse regions that previously challenged the Whisky method.

Furthermore, we have carried out a detailed analysis of how the training data affects the accuracy of different MCMs. When only limited experimental training data are available for the components that make up the mixtures of interest, the prediction accuracy of the MCMs suffers but can be significantly improved by pre-training on predictions from mod. UNIFAC. Additionally augmenting the experimental training set with synthetic data from the SBM, as used in the proposed Blended Whisky method, leads to further improvements. On the other hand, the sheer amount of training data is not everything that is important to achieve very high prediction accuracy. The training data must contain information on mixtures that are similar to the target mixtures. These insights are valuable for selecting training data for MCMs and other data-driven prediction methods and pave the way for developing targeted design of experiments (DOE) strategies. The study also shows that using similarity measures is helpful for ML studies of pure components and mixtures in different ways: in analyzing and selecting training data as well as in assessing uncertainties of the predictions.

Acknowledgement

The authors gratefully acknowledge financial support by Carl Zeiss Foundation in the frame of the project 'Process Engineering 4.0' and by DFG in the frame of the Priority Program SPP 2363 'Molecular Machine Learning' (grant number 497201843). Furthermore, FJ gratefully

acknowledges financial support by DFG in the frame of the Emmy-Noether program (grant number 528649696).

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) 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.

(3) Weidlich, U.; Gmehling, J. A modified UNIFAC model. 1. Prediction of VLE, hE, and .gamma..infin. *Industrial & Engineering Chemistry Research* **1987**, *26*, 1372–1381.

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

(5) Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *The Journal of Physical Chemistry* **1995**, *99*, 2224–2235.

(6) Klamt, A.; Eckert, F. COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. *Fluid Phase Equilibria* **2000**, *172*, 43–72.

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

(8) Hsieh, C.-M.; Lin, S.-T.; Vrabec, J. Considering the dispersive interactions in the COSMO-SAC model for more accurate predictions of fluid phase behavior. *Fluid Phase Equilibria* **2014**, *367*, 109–116.

(9) 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.

(10) Jirasek, F.; Hasse, H. Perspective: Machine Learning of Thermophysical Properties. *Fluid Phase Equilibria* **2021**, *549*, 113206.

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

(12) Sanchez Medina, E. I.; Linke, S.; Stoll, M.; Sundmacher, K. Graph neural networks for the prediction of infinite dilution activity coefficients. *Digital Discovery* **2022**, *1*, 216–225.

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

(14) 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.

(15) 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.

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

(17) 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.

(18) 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.

(19) Jirasek, F.; Hasse, H. Combining Machine Learning with Physical Knowledge in Thermodynamic Modeling of Fluid Mixtures. *Annual review of chemical and biomolecular engineering* **2023**, *14*, 31–51.

(20) Rittig, J. G.; Felton, K. C.; Lapkin, A. A.; Mitsos, A. Gibbs–Duhem-informed neural networks for binary activity coefficient prediction. *Digital Discovery* **2023**, *2*, 1752–1767.

(21) Sanchez Medina, E. I.; Linke, S.; Stoll, M.; Sundmacher, K. Gibbs–Helmholtz graph neural network: capturing the temperature dependency of activity coefficients at infinite dilution. *Digital Discovery* **2023**, *2*, 781–798.

(22) Winter, B.; Winter, C.; Schilling, J.; Bardow, A. A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing. *Digital Discovery* **2022**, *1*, 859–869.

(23) Winter, B.; Winter, C.; Esper, T.; Schilling, J.; Bardow, A. SPT-NRTL: A physics-guided machine learning model to predict thermodynamically consistent activity coefficients. *Fluid Phase Equilibria* **2023**, *568*, 113731.

(24) Hoffmann, M.; Hayer, N.; Kohns, M.; Jirasek, F.; Hasse, H. Prediction of pair interactions in mixtures by matrix completion. *Physical Chemistry Chemical Physics* **2024**, *26*, 19390–19397.

(25) Gond, D.; Sohns, J.-T.; Leitte, H.; Hasse, H.; Jirasek, F. Hierarchical Matrix Completion for the Prediction of Properties of Binary Mixtures. arXiv preprint, `http://arxiv.org/pdf/2410.06060v1`, arXiv:2410.06060 [physics.chem-ph].

(26) 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.

(27) Specht, T.; Nagda, M.; Fellenz, S.; Mandt, S.; Hasse, H.; Jirasek, F. HANNA: Hard-constraint Neural Network for Consistent Activity Coefficient Prediction. *Chemical science* **2024**, *15*, 19777–19786.

(28) Hayer, N.; Specht, T.; Arweiler, J.; Gond, D.; Hasse, H.; Jirasek, F. Prediction of Activity Coefficients by Similarity-Based Imputation using Quantum-Chemical Descriptors. arXiv preprint, `http://arxiv.org/pdf/2412.04993`, arXiv:2412.04993 [physics.chem-ph].

(29) 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.

(30) DDBST - Dortmund Data Bank Software & Separation Technology GmbH Dortmund Data Bank. 2023; `https://www.ddbst.com`.

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

(32) Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M. A.; Guo, J.; Li, P.; Riddell, A. Stan: A Probabilistic Programming Language. *Grantee Submission* **2017**, *76*, 1–32.

(33) Kucukelbir, A.; Tran, D.; Ranganath, R.; Gelman, A.; Blei, D. M. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research* **2017**, *18*, 1–45.

(34) The MathWorks Inc. MATLAB version: 9.13.0 (R2022b). 2022; `https://www.mathworks.com`.

(35) Hsieh, C.-M.; Sandler, S. I.; Lin, S.-T. Improvements of COSMO-SAC for vapor–liquid and liquid–liquid equilibrium predictions. *Fluid Phase Equilibria* **2010**, *297*, 90–97.