Prediction of Activity Coefficients by Similarity-Based Imputation using Quantum-Chemical Descriptors

Nicolas Hayer, Thomas Specht, Justus Arweiler, Dominik Gond, 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

In this work, we introduce a novel approach for predicting thermodynamic properties of binary mixtures, which we call the similarity-based method (SBM). The method is based on quantifying the pairwise similarity of components, which we achieve by comparing quantum-chemical descriptors of the components, namely σ-profiles. The basic idea behind the approach is that mixtures with similar pairs of components will have similar thermodynamic properties. The SBM is trained on a matrix that contains some data for a given property for different binary mixtures; the missing entries are then predicted by the SBM. As an example, we consider the prediction of isothermal activity coefficients at infinite dilution (γ_{ij}^{∞}) and show that the SBM outperforms the well-established physical methods modified UNIFAC (Dortmund) and COSMO-SAC-dsp. In this case, the matrix is only sparsely occupied, and it is shown that the SBM works also if only a limited number of data for similar mixtures is available. The SBM

idea can be transferred to any mixture property and is a powerful tool for generating essential data for many applications.

Introduction

Thermodynamic properties of mixtures are fundamental for the design and optimization of processes. In this work, we describe a novel approach for predicting properties of binary mixtures based on *similarities* between components. This novel similarity-based method (SBM) is built on the fundamental assumption that similar components exhibit similar properties (*similia similibus solvuntur*), making component similarities highly informative inputs for predictive thermodynamic models.

Molecular similarity is commonly used in computational chemistry and pharmaceutical research for database searching and component selection in high-throughput screening. The goal of these applications is to find components that exhibit a behavior that is similar to that of a reference component with desired properties. This is achieved by identifying similar substructures or calculating overall similarity measures, resulting in a list of the most similar molecules in the database and, ultimately, guiding drug discovery and optimization. To perform these pairwise molecular comparisons, a molecular representation of the components and a method to evaluate the similarity based on these representations are required. Various approaches have been proposed for this purpose in the literature, each with its own merits and limitations.[1,2]

The most common molecular representations for similarity searches are molecular fingerprints, which encode structural information into bit vectors, such as the presence of specific functional groups.[2,3] Analyzing fingerprint similarities is computationally efficient, as it only involves comparing bit strings. The Tanimoto coefficient is the most popular metric for assessing fingerprint similarity.[3–5] Other molecular representations for assessing similarity include molecular graphs, molecular descriptor vectors, SMILES, SMARTS, and pharma-

cophores.[1,2,6] Molecular descriptors based on quantum-chemical charge distribution calculations, such as σ-profiles,[7] are rarely used to assess similarities in pharmaceutical research, despite their potential.[8,9]

While the idea of using similarities is implicitly at the heart of many models for predicting thermodynamic properties for unstudied systems, our similarity-based method (SBM) exploits that idea based on a measure of similarity directly.

Among the thermodynamic properties of mixtures, the activity coefficient is particularly significant since it quantifies the non-ideality of liquid mixtures, which is essential for accurately modeling reaction and phase equilibria.[10] A highly informative limiting case is the activity coefficient γ_{ij}^{∞} of a solute i infinitely diluted in a solvent j, as many mixture properties can be predicted based on the knowledge of the limiting activity coefficients. However, despite their importance, experimental data for γ_{ij}^{∞} are scarce, even in comprehensive databases for thermophysical properties such as the Dortmund Data Bank,[11] due to the high cost and time required for their measurement.[12,13] Consequently, reliable prediction methods are essential.

Activity coefficients are usually calculated from models of the Gibbs excess energy G^{E}. Predictions for binary mixtures for which no data are available can be obtained from group-contribution methods, e.g., UNIFAC[14,15] and modified UNIFAC (Dortmund)[16,17]. These models decompose molecules into structural groups and estimate interactions between these groups based on pre-tabulated parameters derived from experimental data. These methods enable predictions for mixtures or components that were not part of the training data set; the only prerequisite is that all relevant pairs of structural groups need to be parameterized. An alternative to group-contribution methods is COSMO-RS[7,18,19], which models intermolecular interactions based on quantum-chemical component descriptors, the σ-profiles. The σ-profile represents the probability distribution $p(\sigma)$ of the screening charge density σ over the surface segments of a molecule embedded in an electrically conductive continuum, and thus indicates how the surface area of the molecule is partitioned among different σ[7]. Open-source versions

of COSMO-RS include COSMO-SAC[20,21] and COSMO-SAC-dsp[22], which rely on similar principles but introduce additional modifications to improve prediction accuracy.

In addition to these physical prediction methods, new machine learning (ML) methods and hybrid models that combine physics with ML have been developed recently.[23,24] These methods include graph neural networks (GNN),[25] transformer models,[26] and matrix completion methods (MCM).[27–29] Additionally, many ML methods have been developed to predict activity coefficients over the entire concentration range, which could also be applied to the special case of activity coefficients at infinite dilution.[30–34]

We apply the SBM here to predict activity coefficients at infinite dilution γ_{ij}^{∞} in binary mixtures. The SBM thereby relies on two sources of information: a novel similarity measure S_{mn} between two components m and n and available experimental data for γ_{ij}^{∞}. The similarity measure S_{mn} is based on a comparison of σ-profiles of the pair of components and used to screen the experimental database, identifying γ_{ij}^{∞} values from similar mixtures that are then used for predictions by imputation. We benchmark the developed SBM with modified UNIFAC (Dortmund),[17] COSMO-SAC,[21] and COSMO-SAC-dsp[22] as three well-established physics-based methods for predicting γ_{ij}^{∞}. Modified UNIFAC (Dortmund) is limited by the availability of pair-interaction parameters that must be fitted to suitable experimental data, which is not always feasible due to sparse data. In contrast, COSMO-SAC and COSMO-SAC-dsp rely primarily on quantum-chemical calculations with only a few adjustable parameters, enabling a larger scope. However, these methods typically exhibit lower prediction accuracy. The SBM aims to provide an alternative to all of these methods by allowing users to tailor the balance between a larger scope and higher prediction accuracy to their specific needs. This is achieved by specifying how high the similarity score must be for a mixture to be included in the prediction process, enabling users to prioritize either accuracy or applicability depending on the chosen threshold. We emphasize that the SBM for predicting γ_{ij}^{∞} is an example; the approach is generic and can be transferred to any other binary property.

Database

Experimental data on activity coefficients at infinite dilution in binary mixtures, γ_{ij}^{∞}, were obtained from the Dortmund Data Bank (DDB).[11] In the preprocessing step, all data sets containing undefined components or labeled as "poor quality" by the DDB were discarded. The focus was restricted to binary mixtures at a temperature of $T = 298.15 \pm 1$ K. If multiple measurements existed for the same binary mixture, the median of these values was adopted. For scaling purposes, the logarithmic activity coefficients, $\ln \gamma_{ij}^{\infty}$, were used throughout this study.

The proposed SBM uses σ-profiles obtained from quantum-chemical COSMO calculations to calculate the similarity between two components. In this work, the σ-profiles were taken from the open-source database provided by Bell et al.,[35] which features results for 2,261 different components. Components not available in this database were excluded from our data set.

Finally, the SBM makes predictions using experimental data from mixtures with the same solute (in a different solvent) or the same solvent (with a different solute), cf. Section "Prediction of Activity Coefficients". To ensure that predictions are always possible, at least two experimental data points were required for each solute and solvent. Data for which this condition was violated were removed. The final data set is visualized in Fig. 1 and comprises 3,568 data points for γ_{ij}^{∞}, covering 221 solutes and 198 solvents.

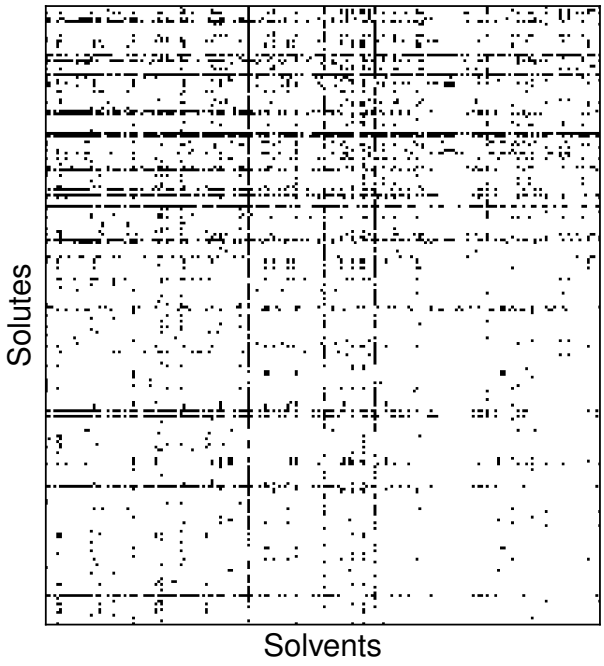

Figure 1: Matrix representing the experimental data on logarithmic activity coefficients at infinite dilution $\ln \gamma_{ij}^{\infty}$ for binary mixtures at 298.15±1 K from the DDB[11] after preprocessing (see text). Experimental data are available for 3,568 binary mixtures, constituting about 8% of all possible combinations of the considered 221 solutes and 198 solvents.

Similarity-Based Method

Similarity Score

Here, we introduce a novel similarity score S_{mn} between two components m and n based on quantum-chemical COSMO calculations. The score S_{mn} is scaled such that its values range from 0 (highly dissimilar components) to 1 (highly similar components) and consists of two contributions, as also indicated in Fig. 2: the similarity based on surface charge distributions S^{σ}_{mn} and the similarity of the surface area S^{A}_{mn} as it is also used in the COSMO method; S^{σ}_{mn} and S^{A}_{mn}, which are described in detail in the following, are also defined to range from 0 to 1. The final similarity score S_{mn} is obtained from a weighted sum of S^{σ}_{mn} and S^{A}_{mn}:

$$S_{mn} = w_{\sigma} \cdot S^{\sigma}_{mn} + (1 - w_{\sigma}) \cdot S^{A}_{mn} \tag{1}$$

where w_{σ} is the weighting factor that controls the relative importance of the surface charge distribution similarity compared to the surface area similarity.

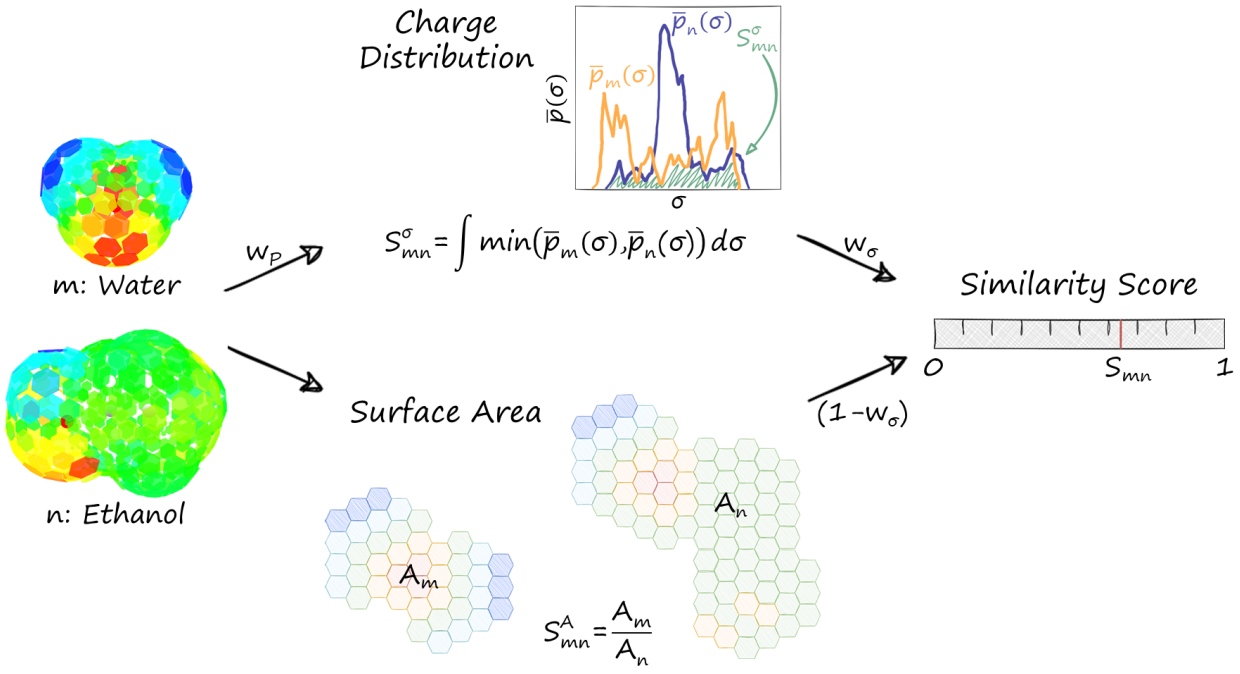

Figure 2: Schematic depiction of calculating the similarity between two components (water and ethanol in this example) as proposed in this work. The final similarity score S_{mn} is composed of two contributions: a similarity based on charge distribution S^{σ}_{mn} (cf. Eq. (3)) and a size similarity derived from the surface areas S^{A}_{mn} (cf. Eq. (2)), which are combined in a weighted sum (cf. Eq. (1)).

The size similarity S^{A}_{mn} is defined as the cavity surface area A of the smaller molecule divided by the one of the larger molecule:

$$S^{A}_{mn} = \begin{cases} \frac{A_m}{A_n}, & \text{if } A_m < A_n \\ \frac{A_n}{A_m}, & \text{if } A_m > A_n \end{cases} \tag{2}$$

For the similarity of the surface charge distributions S^{σ}_{mn}, the overlapping proportion of the σ-profiles of the two components is used, which is calculated using discrete bins for σ via:

$$S^{\sigma}_{mn} = \sum_{k=1}^{N_\sigma} \min\left(\bar{p}_m(\sigma_k), \bar{p}_n(\sigma_k)\right) \tag{3}$$

where $\bar{p}_m(\sigma_k)$ and $\bar{p}_n(\sigma_k)$ are modified σ-profiles, preprocessed as described in the following. All σ-profiles are given here in a discretized version with σ being divided into $N_\sigma = 51$ bins

ranging from -0.025 eÅ^{-2} to 0.025 eÅ^{-2} with a constant step size of 0.001 eÅ^{-2}. We will refer to these values as σ_k for $k = 1, \ldots, 51$. Thus, $p_m(\sigma_k)$ is the fraction of the surface area of the component m associated with the screening charge density σ_k. For all solutes and solvents considered in this work, there is no surface area with a screening charge density outside the range of -0.021 to 0.023 eÅ^{-2}. Accordingly, $p_m(\sigma_k)$ is zero for σ_k values below -0.021 or above 0.023 eÅ^{-2}.

We modify the σ-profiles by introducing w_{P}, which is applied to control the weight on the polar regions in the σ-profiles by being either 0 (no influence) or 2 (more focus on polar regions):

$$p_m^*(\sigma_k) = p_m(\sigma_k) \cdot \sigma_k^{w_{\mathrm{P}}} \tag{4}$$

By setting $w_{\mathrm{P}} = 2$, the similarity calculation emphasizes charge-dense regions, which can be crucial in cases where the behavior of the components is mainly determined by polar interactions.

In the case of $w_{\mathrm{P}} = 2$, the resulting $p_m^*(\sigma_k)$ does not integrate to 1. Therefore, it is again normalized:

$$p_m^{**}(\sigma_k) = \frac{p_m^*(\sigma_k)}{\sum_{k=1}^{N_\sigma} p_m^*(\sigma_k)} \tag{5}$$

In the final processing step, we address a potential issue associated with discretized σ-profiles. Specifically, when calculating the similarity score by comparing the σ-profiles of two molecules bin-wise, small shifts in σ can prevent the detection of structurally similar molecules. Therefore, a moving average with a sliding window of width 2 (corresponding to 0.002 eÅ^{-2}) is applied to all profiles to increase the robustness:

$$\bar{p}_m(\sigma_k) = \frac{p_m^{**}(\sigma_{k-1}) + p_m^{**}(\sigma_k)}{2} \tag{6}$$

The resulting σ-profiles $\bar{p}_m(\sigma_k)$ are used for calculating the similarity of the surface charge distributions S^{σ}_{mn} (see Eq. (3)). Together with the similarity of the surface area S^{A}_{mn} (see Eq. (2)), the final similarity score S_{mn} is calculated (see Eq. (1)).

The two introduced weights w_σ (in Eq. (1)), and w_P (in Eq. (4)) are hyperparameters, which were determined by a grid search. The value ranges of the hyperparameters explored in the grid search are detailed in the "Studied Model Variants" section. In addition to these two weights, other modifications to the calculation of S^{σ}_{mn} (e.g., emphasizing hydrogen-bonding surface segments) and of S^{A}_{mn} (e.g., including component volume) were tested in preliminary studies, but showed no significant impact on the performance of the SBM and were, therefore, discarded.

Prediction of Activity Coefficients

In this section, we explain how the similarity score defined in the previous section is applied for predicting activity coefficients at infinite dilution $\ln \gamma^{\infty}_{ij}$ in unstudied mixtures, where, basically, the $\ln \gamma^{\infty}_{ij}$ is just an example for a property of a binary mixture. The respective method introduced is called the similarity-based method (SBM). The central idea of the SBM is to find mixtures similar to the unstudied mixture that is of interest but for which experimental data on $\ln \gamma^{\infty}_{ij}$ are available. The activity coefficient in the unstudied mixture, $\ln \gamma^{\infty,\mathrm{pred}}_{ij}$, is then predicted simply by arithmetically averaging the corresponding experimental values $\ln \gamma^{\infty,\mathrm{exp}}_{ij}$ of all similar mixtures. As a result, this simple procedure requires no adjustable parameters and yields unbiased predictions.

Here, a *similar mixture* is defined as one with the same solute i (or the same solvent j) but a different solvent n (a different solute m) for which the similarity score S_{nj} (S_{mi}) is higher than a predefined threshold ξ, i.e., $S_{nj} > \xi$ ($S_{mi} > \xi$). Consequently, at least one similar mixture for which an experimental data point is available must be in the database to make a prediction. As a result, there will always be a trade-off when applying the SBM: increasing the threshold value ξ will improve accuracy by restricting the mixtures used as proxies for

the mixture of interest to those with a high degree of similarity, but this reduces the number of mixtures that qualify as similar enough. This can prohibit predictions for certain mixtures if no data points for sufficiently similar mixtures exist. Conversely, lowering ξ expands the range of applicability by allowing more mixtures to be considered similar, but at the cost of reduced accuracy.

The SBM does not require adjustable parameters or training on experimental data. Instead, it relies on a database search driven by the predefined similarity score to generate predictions. This approach ensures that true predictions are obtained, which can then be compared with the corresponding experimental value to assess the model's accuracy. These predictions are also used in comparing the SBM results with the physical benchmark models, resulting in a bias in favor of the physical models, as they were very likely trained on at least some of the data considered here.

A necessary consequence of this direct prediction strategy is that the SBM is constrained by the availability of experimental data for mixtures involving the same solute (in a different solvent) or the same solvent (with a different solute). Consequently, we only considered solutes and solvents that occur in at least two mixtures with experimental data points, ensuring the presence of at least one data point per component to support the prediction process. All calculations of the present study were carried out using Matlab.[36]

Studied Model Variants

The SBM described in the previous sections uses two weights, w_σ and w_P, in calculating the similarity score S_{mn}. These weights were varied in a grid search to explore their effects on model performance. Specifically, w_σ was varied from 0 to 1 in increments of 0.1, while w_P was set to either 0 or 2. This setup resulted in 22 distinct SBM configurations, each representing a different approach to the S_{mn} calculation. The goal of this grid search was to identify the SBM (i.e., weight combination) that performs best for two, often conflicting, objectives:

optimizing the accuracy in predicting $\ln \gamma_{ij}^{\infty}$ in terms of mean absolute error (MAE) and maximizing the scope, i.e., the number of predictable mixtures.

The best-performing SBM, according to these objectives, retains one further adjustable hyperparameter: the threshold ξ, which allows users to balance the trade-off between accuracy and scope. Increasing ξ typically results in more accurate predictions but limits the number of predictable data points since higher similarities are demanded for making predictions. Conversely, lowering ξ increases the number of predictable points but reduces the predictive accuracy since data for less similar components are used for the predictions. To assess the impact of ξ, it was varied from 0.5 to 1 in increments of 0.01 for each of the 22 SBM configurations.

Results and Discussion

Overall Performance of Different Similarity-Based Methods

Fig. 3 shows the predictive accuracy in terms of the MAE of the predicted $\ln \gamma_{ij}^{\infty}$ over the number of predictable data points N from our data set for all tested SBM variants (by varying the weights and ξ). The maximum possible value for N is 3,568, representing the total number of data points for γ_{ij}^{∞} considered in this work (see Section "Data").

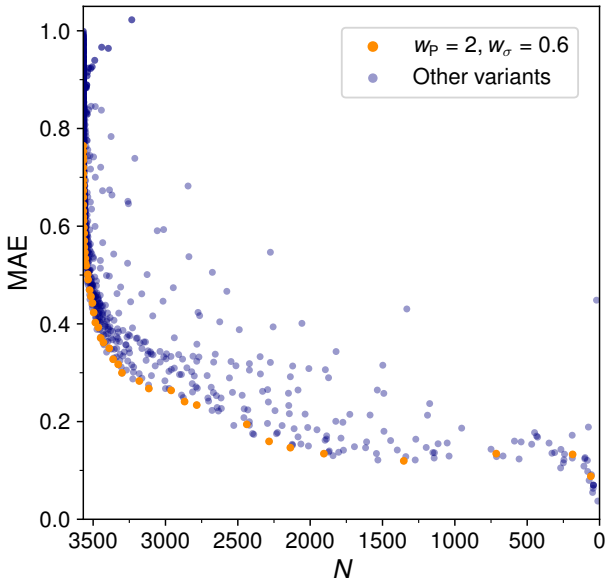

Figure 3: Mean absolute error (MAE) of the predicted $\ln \gamma_{ij}^{\infty}$ over the number of predictable experimental data points N for all tested SBM variants. The results of the best-performing SBM (as specified with the weights w) are highlighted in orange.

The model variants in Fig. 3 scatter across a broad range of MAE and N, underscoring the substantial impact of the selected hyperparameters on model performance. This range highlights the inherent trade-off between predictive accuracy and scope, representing a Pareto optimization problem. In such cases, a solution is considered Pareto-optimal if no feasible solution improves at least one objective without worsening another. Here, certain hyperparameter combinations yield Pareto-optimal SBM variants that achieve maximum ac-

curacy for a given scope and vice versa. The set representing all Pareto-optimal solutions is called the Pareto front.

One particular SBM (with variable ξ) consistently lies on or near the Pareto front, highlighted in orange in Fig. 3. This "best" SBM, defined by $w_\sigma = 0.6$ and $w_\mathrm{P} = 2$, requires only the final tuning of ξ by users to achieve a near-optimal solution tailored to their specific preferences.

The balanced value of $w_\sigma = 0.6$ in the best SBM indicates that components must share similarities in both surface charge distribution and surface area to exhibit comparable values of $\ln \gamma_{ij}^{\infty}$. Furthermore, $w_\mathrm{P} = 2$ emphasizes the importance of a similar surface charge distribution in the polar regions of the components for similar $\ln \gamma_{ij}^{\infty}$.

Figs. S.1 and S.2 in the Supporting Information show further analysis of specific hyperparameter choices. The similarity scores calculated by the best SBM can be used to identify the most similar components for a target component, as exemplified in the Supporting Information (cf. Tables S.2 and S.3).

Comparison to Physical Benchmark Models

The best-performing SBM ($w_\sigma = 0.6$, and $w_\mathrm{P} = 2$) selected in the grid search is further evaluated in the following by comparison against the state-of-the-art physical benchmark methods for predicting activity coefficients: modified UNIFAC (Dortmund),[17] COSMO-SAC,[21] and COSMO-SAC-dsp.[22] As shown in Fig. 4, the methods are compared using the MAE and the scope regarding the number of predictable data points N in our data set. Additionally, the deviations of the predictions from the experimental data are plotted in histograms for the SBM with $\xi = 0.93$, modified UNIFAC (Dortmund), and COSMO-SAC-dsp. Modified UNIFAC (Dortmund) has some extreme outliers, which were excluded from the MAE calculations in Fig. 4. A detailed analysis of these outliers can be found in the Supporting Information, cf. Table S.1.

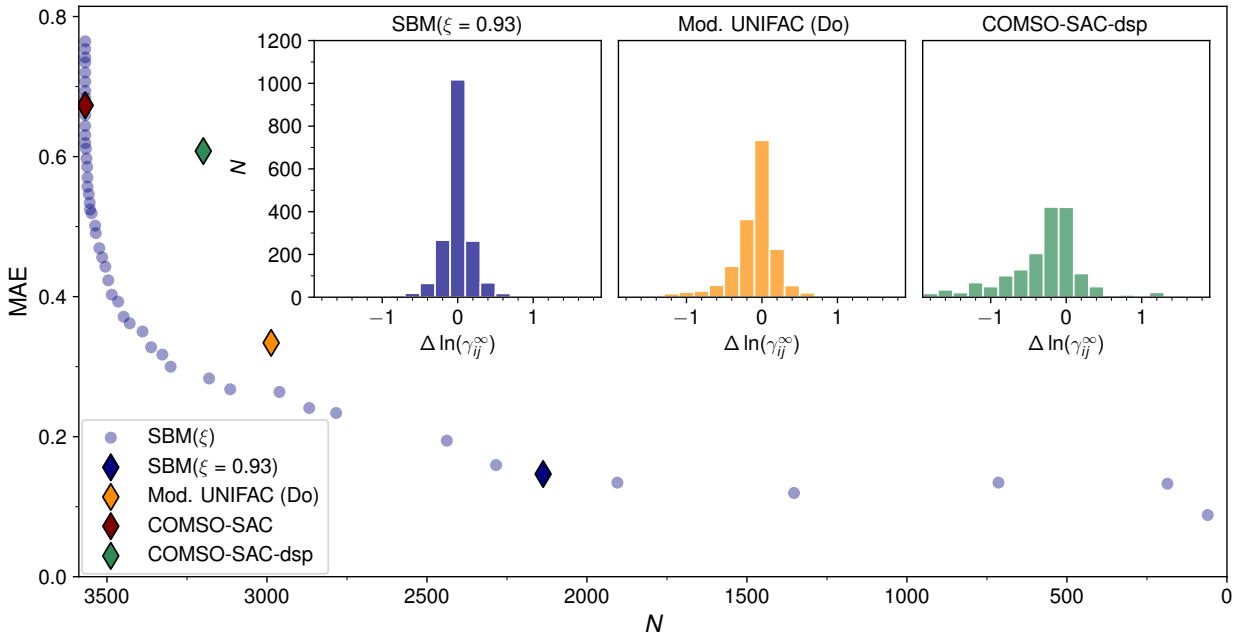

Figure 4: Mean absolute error (MAE) of the best-performing SBM (with varied thresholds ξ), modified UNIFAC (Dortmund), COSMO-SAC, and COSMO-SAC-dsp for the prediction of $\ln \gamma_{ij}^{\infty}$ over the number of predictable experimental data points N. Insets provide histograms of the deviations of the predictions with the SBM with $\xi = 0.93$, modified UNIFAC (Dortmund), and COSMO-SAC-dsp from the experimental data, considering only mixtures that all three methods can describe. The shown interval in the histograms contains 99.9 % (SBM), 96.7 % (modified UNIFAC (Dortmund)), and 96.9 % (COSMO-SAC-dsp) of the relevant 1,748 data points.

In Fig. 4 and throughout the following text, we use "scope" as a synonym for the number of predictable data points from our experimental database (N). However, other definitions of scope are possible, and we discuss one such example in the Supporting Information. This alternative considers all possible mixtures of the solutes and solvents considered in this work, i.e., the number of predictable matrix entries as shown in Fig. 1.

First of all, it is evident from Fig. 4 that for the physical models, there is also a trade-off between the scope of the method and its accuracy. COSMO-SAC-dsp is more accurate than COSMO-SAC, but in its current parameterization,[35] it is not applicable to components containing certain halogens due to missing parameters for the dispersion part, resulting in a slightly smaller scope. Both COSMO variants have a larger scope than modified UNIFAC

(Dortmund), which is constrained by its incomplete tables of pair-interaction parameters that were fitted to experimental data, but they yield less accurate results.

Compared to each physical benchmark method, one can always find an SBM variant (by varying ξ) that outperforms it in terms of prediction accuracy and scope by selecting an appropriate threshold. Specifically, at $\xi = 0.62$, the SBM can, like COSMO-SAC, predict all binary systems in our database but achieves a better MAE (0.62 compared to 0.67). At $\xi = 0.85$, the SBM has a broader scope than COSMO-SAC-dsp ($N = 3,301$ compared to $N = 3,199$) and achieves a better MAE (0.30 compared to 0.61). Similarly, the SBM with $\xi = 0.87$ has a broader scope than modified UNIFAC (Dortmund) ($N = 3,115$ compared to $N = 2,987$) and achieves a better MAE (0.27 compared to 0.33).

For the following analysis, we fix the threshold to $\xi = 0.93$. While this value is, in principle, arbitrary, the resulting model can predict more than half of the available experimental data in our database with relatively high predictive accuracy, achieving an MAE of approx. 0.15. The deviations of the predictions from the experimental data for each method are also represented as histograms in Fig. 4. Most of the predictions of the SBM with $\xi = 0.93$ show deviations from experimental data smaller than ± 0.1 (center bar in the histogram in Fig. 4), which is within the typical range of experimental uncertainty of $\ln \gamma_{ij}^{\infty}$, underscoring the high quality of the predictions that can be obtained with the proposed model.

To further analyze the performance of the best-performing SBM, we plot the respective objectives (MAE and N) over the threshold ξ, as shown in Fig. 5.

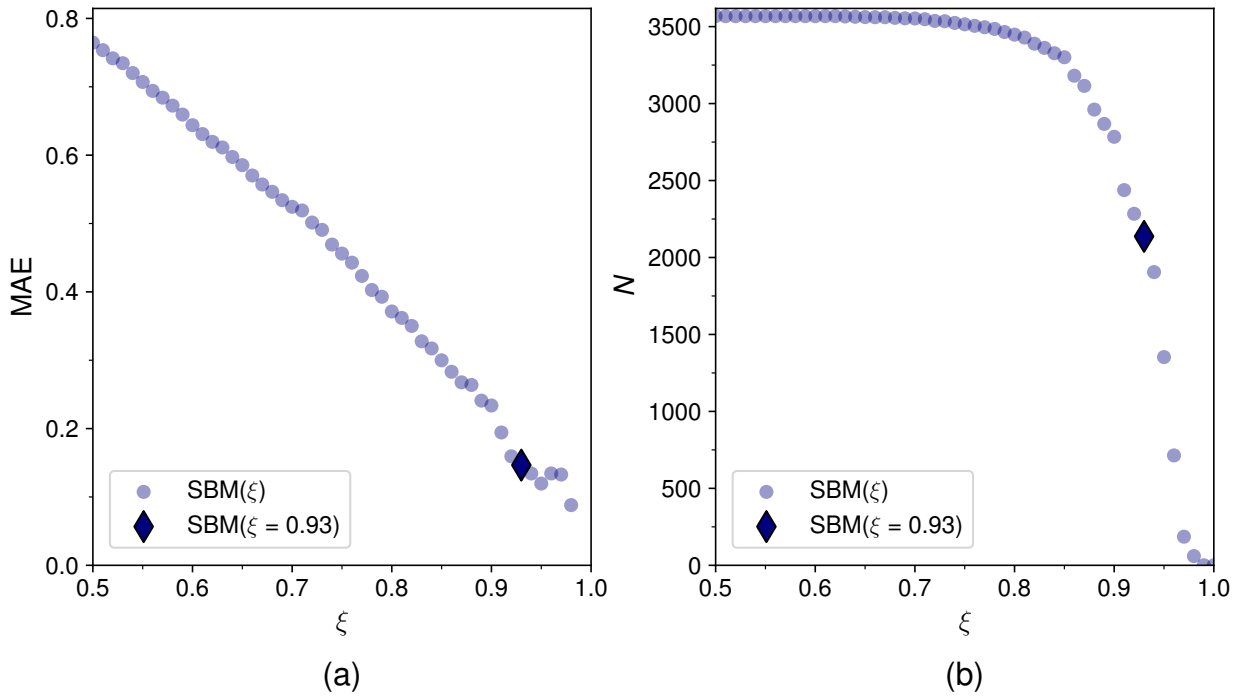

Figure 5: Mean absolute error (MAE) for the prediction of $\ln \gamma_{ij}^{\infty}$ (panel a) and number of predictable experimental data points N (panel b) of the best-performing SBM over the threshold ξ. The results for the SBM with $\xi = 0.93$ are highlighted.

Fig. 5a shows that increasing ξ results in a nearly linear decrease in MAE, indicating improving accuracy. In contrast, the relationship between N and ξ in Fig. 5b is more complex. For $\xi \leq 0.62$, the model achieves its maximum scope, i.e., predicting all experimental data points, while for $\xi > 0.98$, none of the mixtures considered in this work can be predicted. Between these two boundaries, N first decreases slowly with increasing ξ, followed by a steep decrease as ξ approaches 1. This sensitivity of N to ξ emphasizes the importance of selecting an optimal threshold. Overall, Fig. 5 supports the choice of $\xi = 0.93$, marked by the diamond, as a balance point that combines high predictive accuracy with substantial scope. While selecting a lower threshold would yield a broader scope, $\xi = 0.93$ is preferred here as it achieves an MAE in the range of typical experimental uncertainties.

Unlike ML models that rely on molecular descriptors (e.g., SMILES or molecular graphs) and can make, in principle, predictions for all mixtures that can be described by these descriptors, the SBM requires experimental data for similar mixtures to make predictions.

While this limits its applicability to cases where suitable data are available, the prediction accuracy of the SBM correlates directly with the similarity score S_{mn}, as shown in Fig. 5a, providing a clear and interpretable measure of reliability. This stands in contrast to many ML methods, which have broad applicability but often lack an intrinsic indicator of trustworthiness. Additionally, the SBM is basically immune to overfitting since it does not involve any adjustable parameters or training steps; instead, it relies entirely on a database search driven by the defined similarity score.

A detailed analysis of the results for the similarity S_{mn} of all pairs of solutes and all pairs of solvents is presented in Fig. 6. The results are plotted in matrices, which are symmetric as $S_{mn} = S_{nm}$. In these matrices, the solutes (solvents) were arranged so that similar solutes (solvents) were positioned nearby, which was done using a clustering algorithm adopted from a previous work.[37] The chosen arrangement of the solutes (solvents) leads to high values of S_{mn} along the diagonal, cf. Fig. 6.

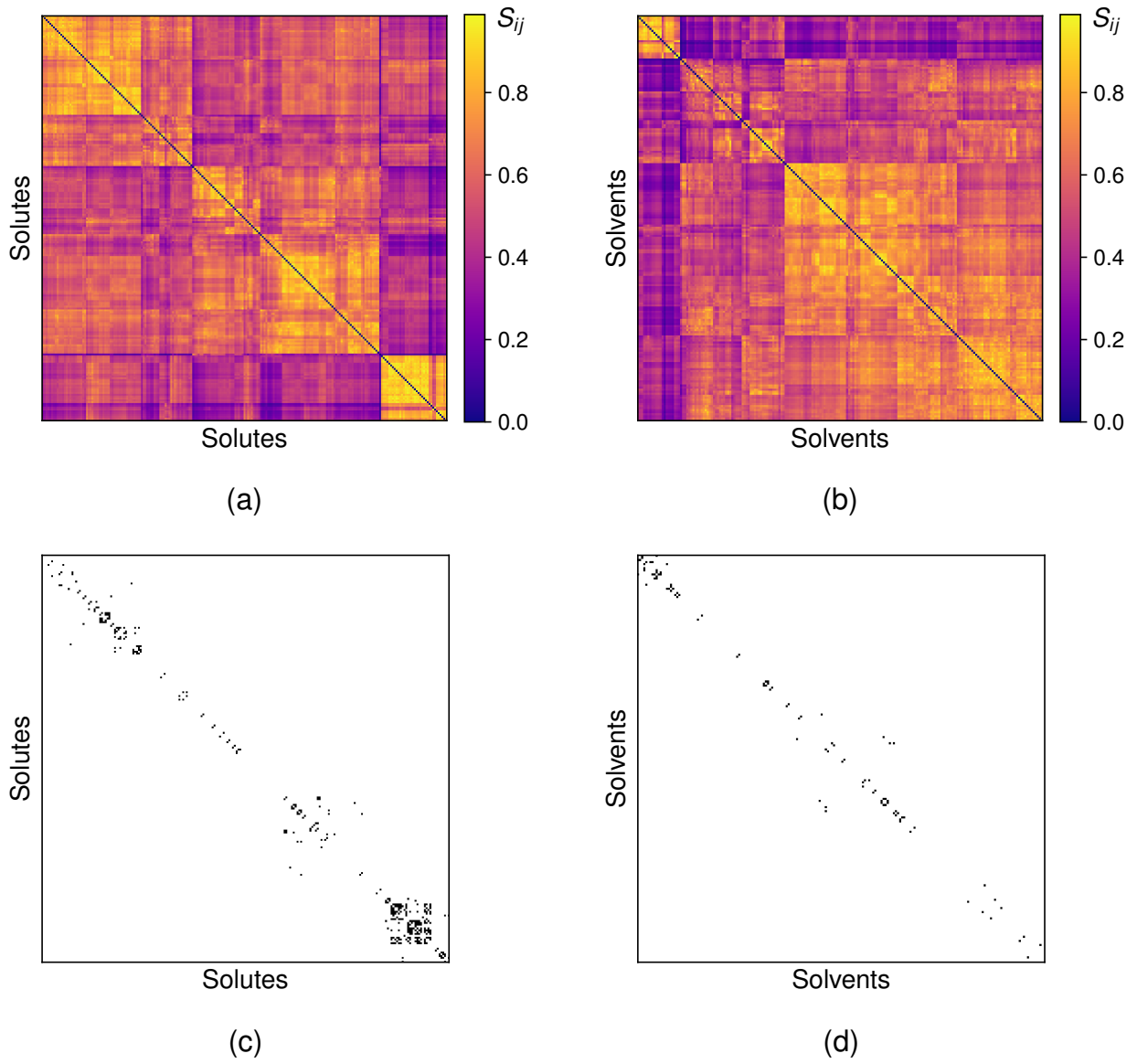

Figure 6: Heatmaps showing results for the pairwise similarity scores S_{mn} of the considered solutes (panel a) and solvents (panel b). For illustration, pairs with $S_{mn} > 0.93$ are highlighted in panels c and d.

The heatmaps in Fig. 6a and b reveal only a few strong similarities among the solutes and solvents in our database, as indicated by the few bright yellow areas. A notable exception is observed in the lower right corner of the solute matrix, cf. Fig. 6a, where a yellow square primarily represents alkanes classified as very similar according to our metric.

This observation becomes even more apparent when highlighting the solute-solute and solvent-solvent combinations with S_{mn} higher than $\xi = 0.93$, the threshold chosen for the detailed analysis discussed above, cf. Fig. 6c and d. Interestingly, only very few, or even just one, similar solutes or solvents for the mixture of interest are needed for the SBM to achieve the excellent predictive accuracy discussed earlier. Thus, for a set of similar mixtures, i.e., those with at least one similar solute or solvent according to our similarity metric, it is sufficient to measure $\ln \gamma_{ij}^{\infty}$ for just one of them. The other mixtures can then be predicted with high accuracy using the SBM. This finding is exciting for the planning of experiments in several ways. For example, it opens up ways to replace substances that are difficult to handle experimentally by suitable proxies, and it can also be used to devise strategies for an efficient design of experiments (DOE) to improve the accuracy and scope of the SBM with a minimum amount of additional experimental data.

Conclusions

This work has two primary outcomes. First, we introduce a novel similarity score, S_{mn}, for comparing two components based on their σ-profiles and surface areas, which are readily obtainable from quantum-chemical calculations or databases. The definition of this score, which ranges between 0 and 1, contains hyperparameters (weights) that can be adapted to different tasks. While this study focuses on predicting limiting activity coefficients γ_{ij}^{∞}, the definition of the similarity score should also be helpful for many other tasks related to predicting or assessing the thermodynamic properties of binary liquid systems.

Second, we developed a new similarity-based method (SBM) for predicting γ_{ij}^{∞}. The SBM uses a database of γ_{ij}^{∞} values and identifies sufficiently similar mixtures using S_{mn} to make predictions for systems without experimental data. The method leverages a user-defined similarity threshold, ξ, to balance accuracy and applicability. Higher ξ values improve prediction accuracy but reduce the number of systems for which predictions are possible, offering flexibility for various use cases.

The SBM we have developed here for predicting γ_{ij}^{∞} shows a remarkable accuracy, even though the database is not large and typically contains only very few (if any) highly similar systems for any given combination of solute i and solvent j. The new SBM outperforms the established physical benchmark methods UNIFAC (Dortmund), COSMO-SAC, and COSMO-SAC-dsp. In particular, restricting the prediction process to mixtures with high similarity scores allows the SBM to achieve accuracies comparable to experimental uncertainties, albeit with reduced applicability. Future integration of machine learning approaches could overcome this remaining limitation and extend the capabilities of the model.

The approach for designing SBMs based on our new similarity score S_{mn} is generic and holds promise for predicting other physical properties of binary liquid mixtures in future work. For thermodynamic applications, the hyperparameters of the calculation of S_{mn} determined in the present work should be a good starting point but could be adapted for other applications.

The observation that data for only a few similar mixtures are sufficient to achieve accurate predictions suggests that a comparatively low number of targeted experiments can considerably improve SBMs. More generally, this finding could form the basis for new guiding principles for the design of experiments in binary systems.

Data Availability Statement

The experimental data on limiting activity coefficients were used under license for this study; they are available directly from Dortmund Data Bank (DDB) version 2023.[11] The σ-profiles as well as the implementations of COSMO-SAC and COSMO-SAC-dsp were taken from the open-source database provided by Bell et al.[35]

Conflicts of Interest

There are no conflicts of interest to declare.

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 SPP2363 '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) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity – a Review. *QSAR & Combinatorial Science* **2003**, *22*, 1006–1026.

(2) Stumpfe, D.; Bajorath, J. Similarity searching. *WIREs Computational Molecular Science* **2011**, *1*, 260–282.

(3) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 379–386.

(4) Fligner, M. A.; Verducci, J. S.; Blower, P. E. A Modification of the Jaccard–Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *44*, 110–119.

(5) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **2015**, *7*, 20.

(6) Raymond, J. W.; Willett, P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *Journal of Computer-Aided Molecular Design* **2002**, *16*, 59–71.

(7) 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.

(8) Thormann, M.; Klamt, A.; Wichmann, K. COSMOsim3D: 3D-similarity and alignment based on COSMO polarization charge densities. *Journal of chemical information and modeling* **2012**, *52*, 2149–2156.

(9) Thormann, M.; Traube, N.; Yehia, N.; Koestler, R.; Galabova, G.; MacAulay, N.; Toft-Bertelsen, T. L. Toward New AQP4 Inhibitors: ORI-TRN-002. *International Journal of Molecular Sciences* **2024**, *25*, 924.

(10) Brouwer, T.; Schuur, B. Model Performances Evaluated for Infinite Dilution Activity Coefficients Prediction at 298.15 K. *Industrial & Engineering Chemistry Research* **2019**, *58*, 8903–8914.

(11) Dortmund Data Bank. 2023; `www.ddbst.com`.

(12) Orbey, H.; Sandler, S. I. Relative measurements of activity coefficients at infinite dilution by gas chromatography. *Industrial & Engineering Chemistry Research* **1991**, *30*, 2006–2011.

(13) Kojima, K.; Zhang, S.; Hiaki, T. Measuring methods of infinite dilution activity coefficients and a database for systems including water. *Fluid Phase Equilibria* **1997**, *131*, 145–179.

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

(15) 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.

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

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

(18) 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.

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

(20) Lin, S.-T.; Sandler, S. I. A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model. *Industrial & Engineering Chemistry Research* **2002**, *41*, 899–913.

(21) 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.

(22) 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.

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

(24) 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.

(25) 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.

(26) 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.

(27) 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.

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

(29) 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.

(30) 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.

(31) 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.

(32) 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.

(33) 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**,

(34) Hayer, N.; Wendel, T.; Mandt, S.; Hasse, H.; Jirasek, F. Advancing Thermodynamic Group-Contribution Methods by Machine Learning: UNIFAC 2.0. `http://arxiv.org/pdf/2408.05220`.

(35) 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.

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

(37) Gond, D.; Sohns, J.-T.; Leitte, H.; Hasse, H.; Jirasek, F. Hierarchical Matrix Completion for the Prediction of Properties of Binary Mixtures. http://arxiv.org/pdf/2410.06060.