# Prediction of Pair Interactions in Mixtures by Matrix Completion[†]

Marco Hoffmann, Nicolas Hayer, Maximilian Kohns, Fabian Jirasek,* and Hans Hasse

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

E-mail: fabian.jirasek@rptu.de

**Abstract**

Molecular simulations enable the prediction of physicochemical properties of mixtures based on pair-interaction models of the pure components and combining rules to describe the unlike interactions. However, if no adjustment to experimental data is made, the existing combining rules often do not yield sufficiently accurate predictions of mixture data. To address this problem, adjustable binary parameters $\xi_{ij}$ describing the pair interactions in mixtures $(i + j)$ are used. In this work, we present the first method for predicting $\xi_{ij}$ for unstudied mixtures based on a matrix completion method (MCM) from machine learning (ML). Considering molecular simulations of Henry's law constants as an example, we demonstrate that $\xi_{ij}$ for unstudied mixtures can be predicted with high accuracy. Using the predicted $\xi_{ij}$ significantly increases the accuracy of the Henry's law constant predictions compared to using the default $\xi_{ij} = 1$. Our approach is generic and can be transferred to molecular simulations of other mixture properties and even to combining rules in equations of state, granting predictive access to the description of unlike intermolecular interactions.

[†]Electronic Supplementary Information (ESI) available.

# Introduction

Molecular simulations are widely used in science and engineering for elucidating phenomena and making quantitative predictions. The backbone of all molecular simulations are molecular models that describe the interactions in the molecular system, which is usually done based on the concept of pair-interaction energies. Once the molecular model, often consisting of a combination of Lennard-Jones (LJ) sites with superimposed electrostatic sites, is specified, the system properties of interest can be determined using appropriate sampling methods. In a binary mixture ($i + j$), there are three types of interactions: the like interactions $i - i$ and $j - j$, as well as the unlike interaction $i - j$. While the like pair interactions are specified by the pure-component models, this is not the case for the unlike interactions, which only occur in the mixture. The unlike electrostatic interactions are given by the laws of electrostatics. However, combining rules must be employed to obtain the unlike LJ interaction parameters. Many different combining rules have been proposed in the literature, but it turns out that a prediction of mixture properties from pure-component data alone is difficult,[1] so that in many cases, an adjustable binary interaction parameter $\xi_{ij}$ is introduced in the combining rule, which is determined from a fit to experimental data for the mixture ($i + j$). However, for many mixtures of interest, no suitable data for fitting $\xi_{ij}$ are available.

This problem is tackled in the present work using an approach from machine learning (ML), namely a matrix completion method (MCM).[2] MCMs are widely used in recommender systems,[3] and different mathematical routines for solving the completion problem are available. In the field of thermodynamics, MCMs have been successfully applied for predicting binary mixture properties, such as activity coefficients,[4–6] Henry's law constants,[7] and diffusion coefficients.[8] Besides purely data-driven MCMs,[4,6] hybrid MCMs[5,7,8] have been developed. Moreover, MCMs were also applied to predict missing pair-interaction parameters of the UNIQUAC[9] and UNIFAC method.[10]

In this work, we use an MCM to predict unknown unlike pair interactions in molecular models by predicting values for $\xi_{ij}$ in the combining rule. The thermodynamic property we

consider as an example is the Henry's law constant $H_{ij}$, where $i$ is the solute and $j$ is the solvent. MCMs are a natural choice for working with binary mixtures, as data for binary mixtures can be represented conveniently in a matrix. After choosing a particular set of $M$ solutes $i$ and $N$ solvents $j$, an $M$x$N$ matrix is obtained, in which we arrange the available data for $\xi_{ij}$. The problem is that in many cases this matrix is only sparsely occupied, as usually experimental data from which $\xi_{ij}$ can be determined are only available for a small fraction of all possible combinations of the $M$ solutes and $N$ solvents. The MCM is used to fill the gaps in this matrix. It solves this task by identifying similarities between the different rows and between the different columns in the matrix, i.e., between solutes as well as between solvents, based on the entries of the matrix. This knowledge is then used for predicting the unknown entries. Thus, the minimum requirement for applying our method to a binary system $(i + j)$ is that at least one experimental data point is available for the solute $i$ (with any solvent) and that at least one experimental data point is available for the solvent $j$ (with any solute).

This paper is organized as follows: first, the data basis of Henry's law constants $H_{ij}$ is presented. Then, the set of molecular models and the combining rule we use are introduced together with the simulation method for determining the Henry's law constant. We can only determine $\xi_{ij}$ from the experimental value of $H_{ij}$ if we have the data and the molecular models for the studied system $(i + j)$. This leads to a fairly small matrix. Finally, the MCM used to predict the missing entries in that matrix is explained, the results are discussed, and conclusions are drawn.

# Data Basis

For fitting $\xi_{ij}$ values, molecular models for the pure components $i$ and $j$ and at least one experimental data point for the mixture $(i + j)$ are required. Thus, the dimension and occupation of the matrix containing $\xi_{ij}$ depend on the availability of both molecular models

and experimental mixture data. Experimental data for Henry's law constants $H_{ij}$ were taken from the Dortmund Data Bank[11] (DDB). Since $\xi_{ij}$ is assumed to be temperature-independent,[12] one experimental data point per binary system is sufficient to adjust $\xi_{ij}$. The MolMod Database[13] was chosen as the source for the molecular models. The experimental data were pre-processed before use (see ESI for details), and the molecular models were tested for their applicability in the simulation methods used in this work. Finally, both databases were intersected to form the final data basis of this work. The process of the model and data selection, as well as the data basis assembly, are described in more detail in the ESI. The final data basis contains 213 experimental Henry's law constants $\ln\left(H_{ij}^{\mathrm{exp}}/\mathrm{kPa}\right)$ and covers 34 solutes and 15 solvents. The resulting 34x15 matrix has 510 elements, of which 213 are known, corresponding to an occupation rate of 40%. This matrix is considerably smaller than the matrices used in previous works on predicting thermodynamic properties of binary mixtures by MCMs,[4–6,8–10] but it has a higher occupation rate. The small size of the matrix is a challenge for the MCM and at the beginning of our work, it was unclear whether the method would work at all on such a small data set. The temperatures for which data are available range from 273.15 K to 313.15 K, with 167 data points at 298.15 K.

## Molecular Models and Simulation

The software package *ms*2[14] was used for all molecular simulations in this work. The molecular models for the solutes and solvents were taken from the MolMod database[13] and consist of LJ sites in combination with electrostatic sites (points charges, dipoles, and quadrupoles). For a complete list of the molecular models used in this work, the reader is referred to the ESI. The equations for the electrostatic contributions to the intermolecular potential are well documented in the literature[15,16] and, hence, not repeated here. The contributions of repulsive and dispersive interactions to the potential between two molecules $i$ and $j$ were

modeled by the LJ potential

$$u_{ij} = \sum_{a=1}^{S_i} \sum_{b=1}^{S_j} 4\varepsilon_{ij,ab} \left[ \left( \frac{\sigma_{ij,ab}}{r_{ij,ab}} \right)^{12} - \left( \frac{\sigma_{ij,ab}}{r_{ij,ab}} \right)^{6} \right], \tag{1}$$

employing the modified Lorentz[17] and Berthelot[18] combining rules for the unlike interaction parameters

$$\sigma_{ij,ab} = \frac{\sigma_{ii,aa} + \sigma_{jj,bb}}{2}, \tag{2}$$

$$\varepsilon_{ij,ab} = \xi_{ij} \sqrt{\varepsilon_{ii,aa} \varepsilon_{jj,bb}}. \tag{3}$$

In these equations, $S_i$ and $S_j$ denote the number of LJ sites on molecules $i$ and $j$, respectively. $\sigma_{ii,aa}$ and $\varepsilon_{ii,aa}$ are the size and energy parameters of the individual LJ sites of molecule $i$, respectively. While the unlike size parameter $\sigma_{ij,ab}$ is calculated directly from the arithmetic mean of the like size parameters, the geometric mean of the like energy parameters is additionally scaled with the binary interaction parameter $\xi_{ij}$ to yield the unlike energy parameter $\varepsilon_{ij,ab}$. The interaction between unlike LJ sites within the same molecules was always modeled with the standard Lorentz and Berthelot rules, i.e., $\xi_{ii} = 1$. Therefore, the binary interaction parameter $\xi_{ij}$ only scales all interactions between LJ sites of unlike molecules.

Henry's law constants of a solute $i$ in a solvent $j$ were obtained by measuring the residual chemical potential $\mu_i^\infty$ of $i$ at infinite dilution in $j$ with Widom's test particle method[19]

$$\mu_i^\infty = -k_\mathrm{B} T \ln \frac{\langle V \exp\left(-\Psi_i / k_\mathrm{B} T\right) \rangle}{\langle V \rangle}, \tag{4}$$

with the Boltzmann constant $k_\mathrm{B}$, the temperature $T$, the volume $V$, the increase in potential energy due to the presence of a test particle $\Psi_i$, and angle brackets representing the $NpT$ ensemble average. The Henry's law constant is then obtained from

$$H_{ij} = \rho_j k_\mathrm{B} T \exp\left( \frac{\mu_i^\infty}{k_\mathrm{B} T} \right), \tag{5}$$

with the solvent density $\rho_j$. Widom's method uses virtual test molecules that are not part of the simulated system and, therefore, do not impact the system's trajectory. Consequently, this allows for the simultaneous calculation of Henry's law constants for an arbitrary number of solutes in the same solvent (and at the same temperature and pressure), thereby vastly increasing the computational efficiency. However, Widom's test particle method is prone to errors when inserting large molecules, and if the solvent density is relatively high. In these cases, it becomes likely that the inserted solute molecule overlaps with solvent molecules, thus yielding very high potential energies and the failure to sample Eq. (4) accurately. Correlations for the density and the vapor pressure were used to initialize the simulations to speed up equilibration. The pressure was set to 105 % of the pressure obtained by the vapor pressure correlation to ensure a liquid phase at all times. The employed correlations and technical details on the molecular simulations are given in the ESI.

Molecular simulations were carried out for all binary systems for which molecular models were available in the MolMod data bank for both solute and solvent and for which an experimental data point was reported in the DDB (see ESI for details). Each Henry's law constant was simulated three times using the binary interaction parameters $\xi_{ij} \in \{0.95, 1.00, 1.05\}$. From the results, the value of $\xi_{ij}$ for the studied system was determined as explained below. For details on the molecular simulations, the reader is referred to the section Molecular Simulation in the ESI. All simulation results were checked for convergence and faulty or questionable results were omitted from the further analysis. Problems were always caused by a failure of Widom's test particle method. Only binary systems for which all three simulations at different $\xi_{ij}$ yielded valid results were investigated further. The statistical uncertainties of the Henry's law constants obtained by molecular simulation were estimated by evaluating the standard deviation of $\mu_i^\infty$ with the block averaging method.[20]

### Fitting of Binary Interaction Parameters

The three Henry's law constants $\ln\left(H_{ij}^{\mathrm{MolSim}}/\mathrm{kPa}\right)$ for each system $(i + j)$ obtained by molecular simulations with $\xi_{ij} \in \{0.95, 1.00, 1.05\}$ were used to fit the binary interaction parameter $\xi_{ij}^{\mathrm{exp}}$ to the experimental Henry's law constant. We empirically found a linear dependence of $\ln\left(H_{ij}^{\mathrm{MolSim}}/\mathrm{kPa}\right)$ on $\xi_{ij}$ for all studied systems, so that a linear fit was used:

$$\ln\left(H_{ij}/\mathrm{kPa}\right) = C_{1,ij} \cdot \xi_{ij} + C_{0,ij}. \tag{6}$$

The parameters $C_{1,ij}$ and $C_{0,ij}$ were determined from a fit to the three simulation data points. The experimental data point for $\ln\left(H_{ij}^{\mathrm{exp}}/\mathrm{kPa}\right)$ was then inserted into Eq. (6) to find the value for $\xi_{ij}$ for the studied system. In the following, these values are designated as $\xi_{ij}^{\mathrm{exp}}$. Eq. (6) was also used throughout this work to obtain values for $\ln\left(H_{ij}/\mathrm{kPa}\right)$ from $\xi_{ij}$ values (e.g., from those predicted by the MCM). The linear correlation was also found to be accurate in cases where the values of $\xi_{ij}$ are not in the range [0.95,1.05] used for determining the parameters. In the ESI, section 'Validity of the Linear Correlation', this is demonstrated for two solutes as examples by comparing the extrapolation up to $\xi_{ij} = 0.5$ and $\xi_{ij} = 1.5$ with dedicated molecular simulations using these values. It is important to note that the fit in Eq. (6) and thus the parameters $C_{1,ij}$ and $C_{0,ij}$ are only valid at the temperature of the corresponding experimental data point.

## Matrix Completion Method (MCM)

For predicting the missing entries in the sparse matrix of $\xi_{ij}^{\mathrm{exp}}$, an MCM was developed. Each binary interaction parameter $\xi_{ij}^{\mathrm{exp}}$ is thereby modeled by Eq. (7):

$$\xi_{ij}^{\mathrm{exp}} = \boldsymbol{u}_i^T \cdot \boldsymbol{v}_j + b_i + b_j + \epsilon_{ij}. \tag{7}$$

Here, $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are column vectors of length $K$ containing features of the molecules $i$ and $j$, respectively, and $b_i$ and $b_j$ are scalars that can be considered as 'biases' of molecule $i$ and $j$, respectively, describing a general tendency of these molecules for higher or lower values of $\xi_{ij}$. Moreover, $\epsilon_{ij} = \xi_{ij}^{\text{exp}} - \xi_{ij}^{\text{pred}}$ describes the differences between the training data $\xi_{ij}^{\text{exp}}$ and the model predictions $\xi_{ij}^{\text{pred}}$ and is minimized during the training by adjusting the model parameters $\boldsymbol{u}_i$, $\boldsymbol{v}_j$, $b_i$, and $b_j$.

A Bayesian approach was used to train the model, where all parameters and data points are modeled as random variables following a probability distribution. The training aims to find the so-called *posterior*, which describes the probability distribution over the model parameters conditioned on the training data; by sampling from the posterior, the most probable model parameters for describing the training data can be retrieved. Bayes' rule relates the posterior to the product of two other distributions, the so-called *prior* distribution over the model parameters, which describes expectations on the model parameters prior to the training, and the so-called *likelihood* distribution, which describes the probability of the observed data conditioned on the model parameters.

In this work, the priors over all parameters were modeled with zero-centered normal distributions with standard deviation of $\sigma = 0.5$ for $u_i$ and $v_j$, and $\sigma = 5$ for $b_i$ and $b_j$. These rather broad distributions reflect the lack of prior knowledge of the model parameters but act as weak regularizers of the model by penalizing very large parameter values. The likelihood was also modeled with normal distributions centered around $\boldsymbol{u}_i^T \cdot \boldsymbol{v}_j + b_i + b_j$; the respective standard deviation as well as the dimension $K$ of $u_i$ and $v_j$ were treated as hyperparameters of the model and optimized in a preliminary study. A dimension of $K = 2$ and a standard deviation of the likelihood normal distribution of $\sigma = 0.02$ yielded the best results.

After obtaining the probability distributions of the model parameters $u_i$, $v_j$, $b_i$, and $b_j$ from training, predictions for $\xi_{ij}$ were made with Eq. (7) using the posterior means for all parameters. This is done by sampling 1000 values from the posterior probability distributions of the parameters after the training and using these to calculate 1000 samples for

$\xi_{ij}$. From these 1000 samples, the mean and the standard deviation are calculated. The model training was performed with the probabilistic programming language *Stan*[21] using variational inference. For more details on variational inference the reader is referred to our previous work.[4]

The predictive performance of the developed model was tested using a leave-one-out (LOO) strategy: the model was trained and evaluated multiple times, with a single data point being omitted during the training and considered as the test data point in that run. This way, a true prediction was obtained for each data point, which could be compared to the experimental data. For the evaluation of the predictive performance of the developed model, the mean absolute error (MAE) and mean squared error (MSE) were used:

$$\mathrm{MAE} = \frac{1}{N}\sum_{k=1}^{N}\left|X_{ij}^{\mathrm{pred}} - X_{ij}^{\mathrm{exp}}\right|, \tag{8}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{k=1}^{N}\left(X_{ij}^{\mathrm{pred}} - X_{ij}^{\mathrm{exp}}\right)^2. \tag{9}$$

where $X_{ij}^{\mathrm{pred}}$ and $X_{ij}^{\mathrm{exp}}$ are the predicted and experimental data point, respectively. $k$ runs over all $N$ test data points. Eqs. (8) and (9) were used to evaluate both the predicted binary interaction parameters and the thereof predicted $\ln\left(H_{ij}^{\mathrm{MolSim}}/\mathrm{kPa}\right)$ from Eq. (6).

# Results and Discussion

## Fitted Binary Interaction Parameters

In the first step, molecular simulations to determine Henry's law constants for three different binary interaction parameters were performed for the 213 selected binary systems and the values of $\xi_{ij}^{\mathrm{exp}}$ were determined from the experimental data as explained above. The obtained values for $\xi_{ij}^{\mathrm{exp}}$ range from 0.447 to 1.679 and are depicted as a heatmap in Fig. 1 and provided as csv-file with the Supplementary Material. Considering only systems for a given solute or

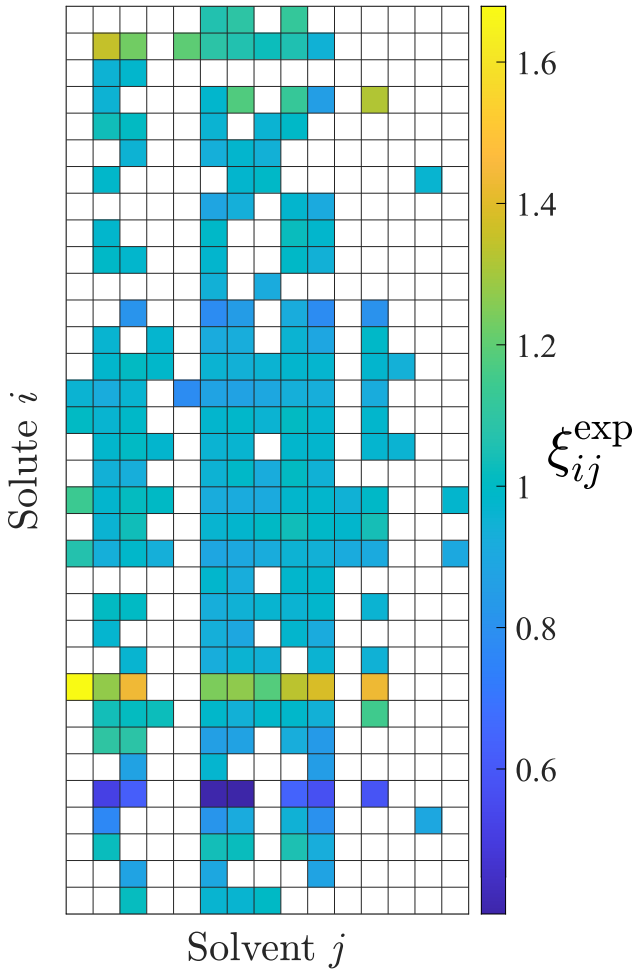


Figure 1: Heatmap of the binary interaction parameters $\xi_{ij}^{\mathrm{exp}}$ fitted to experimental Henry's law constants. Solute and solvent models are sorted by their ID, as defined in the ESI.

solvent, one finds that the number of available entries in the matrix varies strongly among the solvents and among the solutes. Furthermore, tendencies to higher or lower values of $\xi_{ij}$ are observed for some of the solutes and solvents. Fig. 2 (left) shows a plot of the experimental Henry's law constants over the corresponding fitted binary interaction parameters, and Fig. 2 (right) shows a histogram of $\xi_{ij}^{\mathrm{exp}}$. Fig. 2 (left) shows that the investigated binary systems cover a wide range of $H_{ij}$, where low values for $H_{ij}$ indicate a high solubility of the solute $i$ in the solvent $j$ and vice versa.

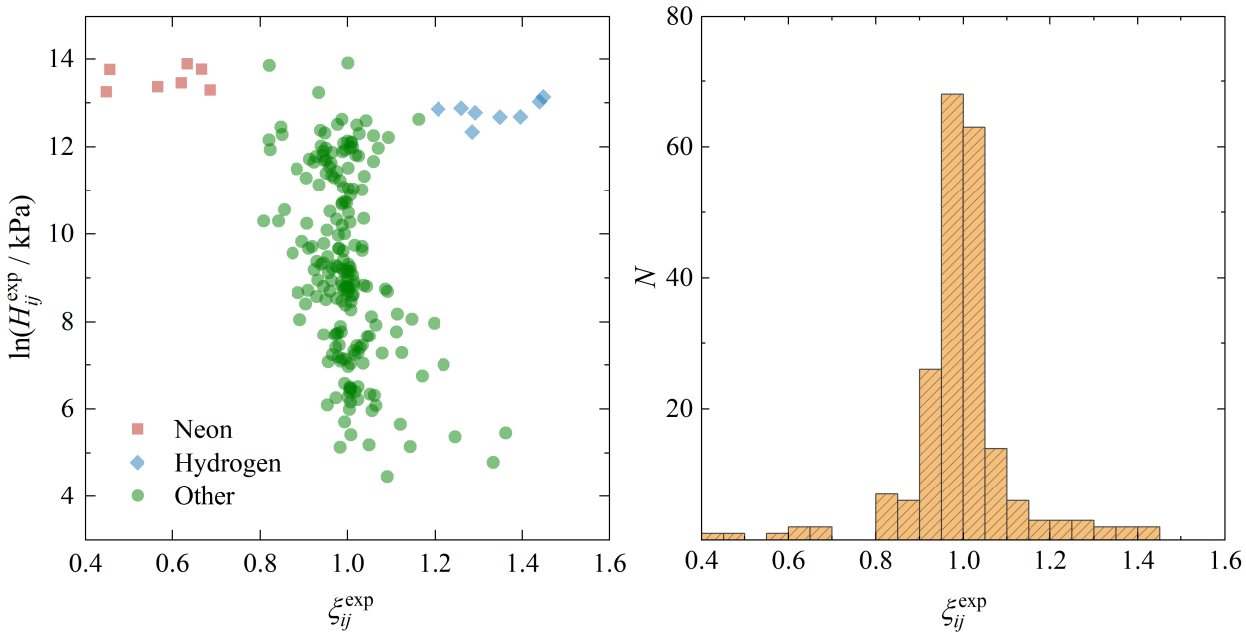


Figure 2: Left: experimental Henry's law constants plotted against the fitted binary interaction parameters $\xi_{ij}^{\mathrm{exp}}$. Right: histogram showing the number of systems as a function of $\xi_{ij}^{\mathrm{exp}}$.

For more than 90 % of the binary systems, the calculated binary interaction parameters deviate less than 0.2 from the default value of 1, showing that in most cases, moderate adjustments in $\xi_{ij}$ are sufficient to match experimental data points. It should be noted that the impact of $\xi_{ij}$ on the Henry's law constant varies strongly between systems. It is determined by the slope of the linear relation between $\ln(H_{ij}/\mathrm{kPa})$ and $\xi_{ij}$ given by Eq. (6), i.e., the parameter $C_{1,ij}$. The values of $C_{1,ij}$ for the studied systems vary by up to a factor

of 10.

For the solutes Neon and Hydrogen the $C_{1,ij}$ have particularly low values, indicating a weak influence of $\xi_{ij}$ on $\ln(H_{ij}/\mathrm{kPa})$, which explains the large deviations from $\xi_{ij} = 1$ that were found for systems with these two solutes. The small influence of $\xi_{ij}$ on $\ln(H_{ij}/\mathrm{kPa})$ for these compounds is a consequence of their low $\varepsilon_{ij}$ values (compare Eqs. (1)-(5)). Nevertheless, the extreme values of $\xi_{ij}$ for the systems containing Neon and Hydrogen were included in the MCM.

## Predicted Binary Interaction Parameters

Fig. 3 shows the $\xi_{ij}^{\mathrm{pred}}$ predicted by the MCM in comparison to the experimental binary interaction parameters $\xi_{ij}^{\mathrm{exp}}$. The numbers from the MCM were obtained by the LOO analysis as explained above and are, hence, true predictions. For comparison, also the corresponding results for the default value of $\xi_{ij} = 1$ are shown.

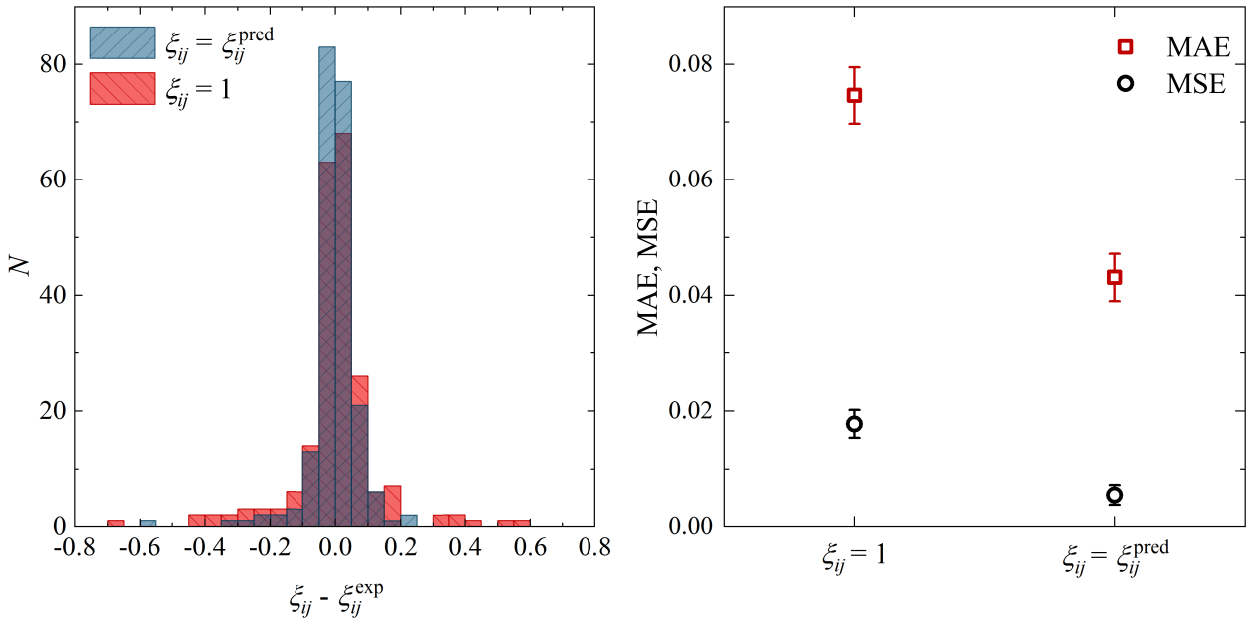


Figure 3: MCM predictions of the interaction parameter ($\xi_{ij}^{\mathrm{pred}}$) compared to the experimental values $\xi_{ij}^{\mathrm{exp}}$. The comparison is also carried out for the default value of $\xi_{ij} = 1$. Left: histogram representing the number of systems over their deviation from $\xi_{ij}^{\mathrm{exp}}$. Right: Mean Absolute Error (MAE) and Mean Squared Error (MSE) for both approaches. Error bars show the standard errors of the means.

The left panel of Fig. 3 shows a histogram indicating the number of systems that fall into certain bins of the deviations, while in the right panel, the results for MAE and MSE (see Eqs. (8), (9)) are shown. The deviations from $\xi_{ij}^{\mathrm{exp}}$ are much lower for the MCM compared to those obtained using $\xi_{ij} = 1$: the MAE is almost halved and the MSE is reduced to less than one-third, which indicates that the number of outliers is also significantly reduced using the MCM. Both effects are visible in the histogram shown in Fig. 3. In the section 'Model Uncertainty' in the ESI, a detailed analysis of the correlation between the predicted standard deviation of $\xi_{ij}^{\mathrm{pred}}$ and the number of available training data points in the same row (solute) and column (solvent) is given.

## Predicted Henry's Law Constants

Fig. 4 shows the results for the predicted Henry's law constants using $\xi_{ij}^{\mathrm{pred}}$ from the MCM in comparison to the experimental values. Again, results obtained with the default value of $\xi_{ij} = 1$ are also shown for comparison.

The representation is the same as in Fig. 3. The histogram in Fig. 4 (left) shows that using $\xi_{ij}^{\mathrm{pred}}$ significantly reduces the number of predictions with large deviations from the experimental values compared to using $\xi_{ij} = 1$. Using $\xi_{ij}^{\mathrm{pred}}$ instead of $\xi_{ij} = 1$ reduces both the MAE (from 0.461 to 0.342) and the MSE (from 0.454 to 0.299) significantly. Overall, the results show that the MCM approach works despite the very small database with only 34 solutes $i$ and 15 solvents $j$, which makes it extremely difficult to find similarities among the solutes, and especially among the solvents. With the Supplementary Material, a completed set of binary interaction parameters and respective standard deviations for the studied binary systems is reported. Both tables are available as csv-files. These binary interaction parameters were predicted with the developed method after training on all available fitted binary interaction parameters, i.e., without withholding data points in a LOO analysis. Based on the results of the LOO evaluation as discussed above, the accuracy of molecular simulations using these parameters is expected to be significantly better compared to using the default

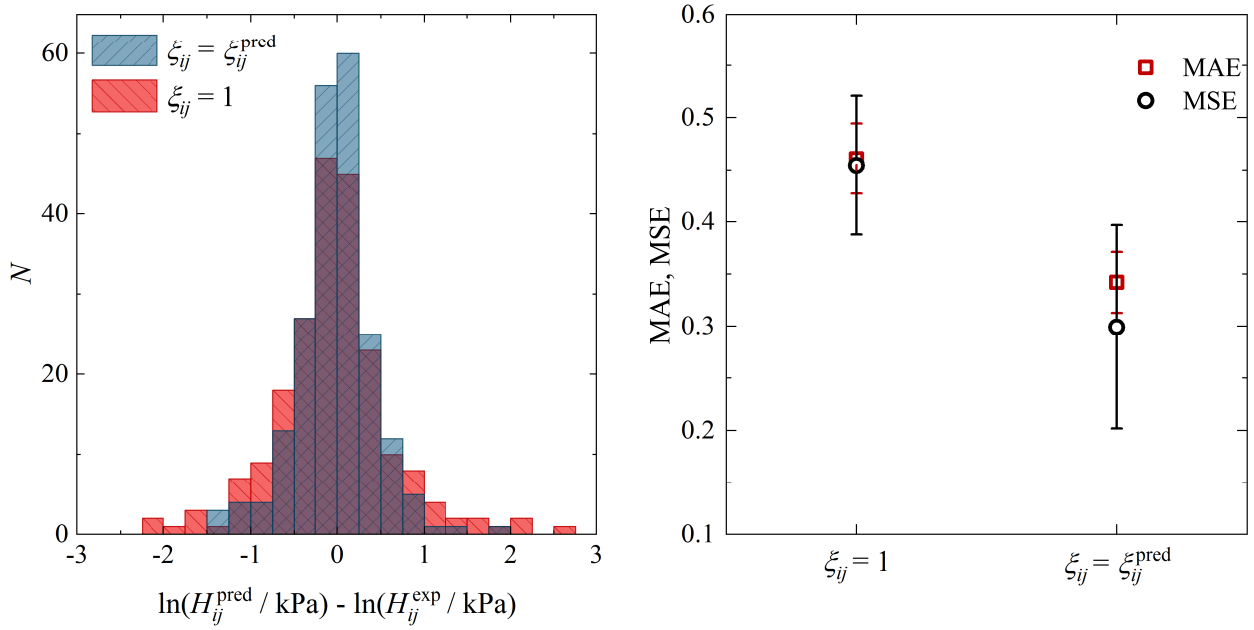


Figure 4: Predictions of Henry's law constants using Eq. (6) with $\xi_{ij}^{\text{pred}}$ and $\xi_{ij} = 1$ compared to the experimental data. Left: histogram representing the number of systems using $\xi_{ij}^{\text{pred}}$ and $\xi_{ij} = 1$ over the deviation of model results from the experimental data. Right: Mean Absolute Error (MAE) and Mean Squared Error (MSE) of $\ln\left(H_{ij}^{\text{exp}}/\text{kPa}\right)$ obtained for both approaches. Error bars show the standard errors of the means.

value of $\xi_{ij} = 1$.

## Application to Molecular Simulation

Additional molecular simulations of temperature-dependent Henry's law constants for four binary systems were performed to test the results obtained from single state points as described above. In each case, the molecular simulations were carried out using $\xi_{ij} = 1$, $\xi_{ij}^{\text{exp}}$, and $\xi_{ij}^{\text{pred}}$ (from the LOO analysis). The results are depicted in Fig. 5, which includes consolidated experimental data from the DDB for reference.

For the four investigated systems (Methane in Ethanol, Oxygen in Acetone, Hydrogen in Benzene, Acetylene in Acetone), the molecular simulations with the standard predictive approach $\xi_{ij} = 1$ predict the temperature dependence of the Henry's law constant *qualitatively* well, but the experimental results are not matched *quantitatively*. If at least one experimental data point for the system of interest is available to fit the respective binary interaction

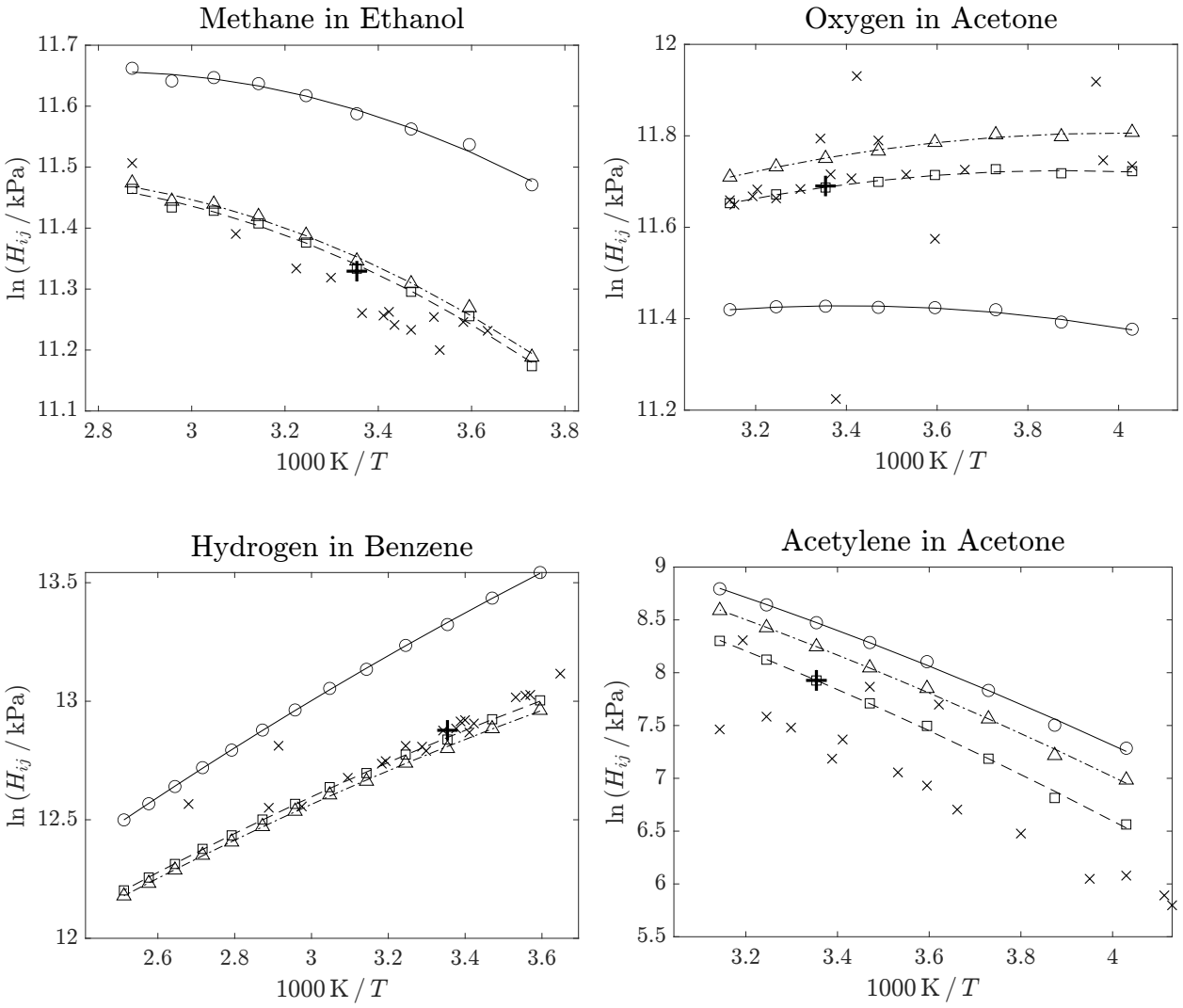


Figure 5: Results of molecular simulations of temperature-dependent Henry's law constants in four binary systems using $\xi_{ij} = 1$ (○, solid), $\xi_{ij}^{\mathrm{exp}}$ (□, dashed), and $\xi_{ij}^{\mathrm{pred}}$ (△, dash-dotted) predicted by the MCM with LOO. For comparison, consolidated experimental data (×) are also depicted (see ESI). The data points that were used for the fit of $\xi_{ij}^{\mathrm{exp}}$ are marked (+). Error bars of the statistical uncertainty of the molecular simulations are omitted for clarity. Lines are guides for the eye.

parameter, the results are much better. This demonstrates that molecular simulation is able to describe the experimental data also *quantitatively*, if $\xi_{ij}$ is suitably chosen (see results for $\xi_{ij}^{\mathrm{exp}}$ in Fig. 5). The simulation results using $\xi_{ij}^{\mathrm{pred}}$ as predicted by the MCM are very similar to those obtained using the fitted $\xi_{ij}^{\mathrm{exp}}$, although not a single experimental data point for the studied system was used the training the MCM (true predictions obtained using LOO analysis are shown in Fig. 5).

In Fig. 3 in the ESI, we show results of molecular simulations for the same systems using binary interaction parameters predicted by an MCM that was trained on all training data (without omitting test data in a LOO analysis). These results are very similar to those obtained using the truly predicted $\xi_{ij}^{\mathrm{pred}}$ shown in Fig. 5, which demonstrates the robustness of the developed MCM.

Overall, the results support the assumption that $\xi_{ij}$ is a temperature-independent parameter. Thus, predicting one $\xi_{ij}^{\mathrm{pred}}$ with the MCM suffices to obtain good simulation results for a wide temperature range. Inverting this argument, the results of molecular simulations with $\xi_{ij}^{\mathrm{pred}}$ from the MCM can be used for the critical assessment of experimental data and help to identify outliers that do not match the overall picture obtained in studies of the same property in other but similar binary systems.

# Conclusions

In this work, we have introduced a novel method for predicting unlike interaction parameters in molecular simulation of mixtures. Usually, the unlike interactions are calculated from the like interactions, which are, in turn, determined from pure-component data. To improve the representation of the mixture data, the combining rules that are used for this purpose often contain adjustable binary parameters $\xi_{ij}$. We suggest to predict these parameters using a matrix completion method (MCM) from machine learning and show that this can be done by studying a demanding example: the prediction of $\xi_{ij}$ for calculating Henry's law constants

$H_{ij}$ by molecular simulations. Even though only a small database was used for this feasibility study, good results were obtained. In predictions of systems that were not included in the training set, using values for $\xi_{ij}^{\mathrm{pred}}$ predicted by the MCM gave a clear improvement over working with $\xi_{ij} = 1$, which was so far the undisputed choice for systems for which no mixture data were available.

The new approach is generic and broadly applicable. It can be applied to any class of molecular models, any combination rule containing system-specific parameters, and can be trained on any class of mixture data. However, for its successful application, several requirements have to be fulfilled. (i) A sufficiently large set of compatible and consistent molecular models of pure components must be available. While individually adjusted models were used in the present work, the new method could also be applied to transferable force fields with their broad scope. (ii) A sufficiently large data set with experimental binary mixture data of a given thermodynamic property or, alternatively, a set of thermodynamic properties that can be represented simultaneously reasonably well with the chosen class of molecular models must be available. In the present pilot study, we have limited ourselves to data on Henry's law constants at a single temperature. A natural choice would be to extend this to vapor-liquid data in general, but other thermodynamic properties could also be included. (iii) The intersection of the two data sets (i) and (ii) must be sufficiently large to apply the MCM. A prerequisite for the MCM to work is not only that enough data are available, but also that they are sufficiently correlated. While for sufficiently large and accurate experimental data sets such correlations exist simply because similarities between different components exist (just think of homologous series), this is not necessarily the case for molecular models of different types plugged together by some combining rule. Hence, the consistency of the molecular models matters. Last but not least, our approach can also be transferred to the prediction of pair-interaction parameters in combining rules of equations of state.

## Data Availability Statement

The datasets for both the fitted and the predicted binary interaction parameters are made available as part of the supplementary material. The experimental training data were used under license for this study; they are available directly from Dortmund Data Bank (DDB) version 2022.[11] The *Stan* code for the matrix completion method is available with the ESI (*MCM_model.stan*). The input files for the molecular models used in this work can be downloaded free of charge from the MolMod Database.[13]

## Conflicts of Interest

There are no conflicts of interest to declare.

## Acknowledgement

The authors gratefully acknowledge financial support by the Carl-Zeiss-Stiftung (grant number: P2018-02-002) and Deutsche Forschungsgemeinschaft in the frame of the Emmy Noether grant of Fabian Jirasek and a grant of Fabian Jirasek and Hans Hasse in the Priority Program 2363. The present work was conducted under the auspices of the Boltzmann-Zuse Society of Computational Molecular Engineering (BZS) and the simulations were carried out on the Regional University Computing Center Kaiserslautern (RHRZ) under the grant RPTU-MTD.

# References


(1) Schnabel, T.; Vrabec, J.; Hasse, H. Henry's law constants of methane, nitrogen, oxygen and carbon dioxide in ethanol from 273 to 498 K: Prediction from molecular simulation. *Fluid Phase Equilibria* **2005**, *233*, 134–143.

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

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

(4) 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.

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

(6) 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.

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

(8) 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.

(9) 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.

(10) Jirasek, F.; Hayer, N.; Abbas, R.; Schmid, B.; Hasse, H. Prediction of parameters of group contribution models of mixtures by matrix completion. *Phys. Chem. Chem. Phys.* **2023**, *25*, 1054–1062.

(11) Dortmund Data Bank. www.ddbst.com, 2022.

(12) Schnabel, T.; Vrabec, J.; Hasse, H. Unlike Lennard–Jones parameters for vapor–liquid equilibria. *Journal of Molecular Liquids* **2007**, *135*, 170–178.

(13) Stephan, S.; Horsch, M. T.; Vrabec, J.; Hasse, H. MolMod – an open access database of force fields for molecular simulations of fluids. *Molecular Simulation* **2019**, *45*, 806–814.

(14) Deublein, S.; Eckl, B.; Stoll, J.; Lishchuk, S. V.; Guevara-Carrion, G.; Glass, C. W.; Merker, T.; Bernreuther, M.; Hasse, H.; Vrabec, J. ms2: A molecular simulation tool for thermodynamic properties. *Computer Physics Communications* **2011**, *182*, 2350–2367.

(15) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press, Oxford, 1987.

(16) Gray, C.; Gubbins, K. *Theory of Molecular Fluids, Volume 1: Fundamentals*; Clarendon Press, Oxford, 1985.

(17) Lorentz, H. A. Ueber die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gase. *Annalen der physik* **1881**, *248*, 127–136.

(18) Berthelot, D. Sure le mèlange de gaz. *Comptes rendus des séances de l'academie de sciences Paris* **1898**, *126*, 1703–1855.

(19) Widom, B. Some Topics in the Theory of Fluids. *The Journal of Chemical Physics* **1963**, *39*, 2808–2812.

(20) Flyvbjerg, H.; Petersen, H. G. Error estimates on averages of correlated data. *The Journal of Chemical Physics* **1989**, *91*, 461–466.

(21) Stan. www.mc-stan.org, 2022.