PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6cp00732e

Prediction of diffusion coefficients in mixtures with tensor completion

Zeno Romero, Kerstin Münnemann, Hans Hasse and Fabian Jirasek *

Predicting diffusion coefficients in mixtures is crucial for many applications, as experimental data remain scarce, and machine learning (ML) offers promising alternatives to established semi-empirical models. Among ML models, matrix completion methods (MCMs) have proven effective in predicting thermophysical properties, including diffusion coefficients in binary mixtures. However, MCMs are restricted to single-temperature predictions, and their accuracy depends strongly on the availability of high-quality experimental data for each temperature of interest. In this work, we address this challenge by presenting a hybrid tensor completion method (TCM) for predicting temperature-dependent diffusion coefficients at infinite dilution in binary mixtures. The TCM employs a Tucker decomposition and is jointly trained on experimental data for diffusion coefficients at infinite dilution in binary systems at 298 K, 313 K, and 333 K. Predictions from the semi-empirical SEGWE model serve as prior knowledge within a Bayesian training framework. The TCM then extrapolates linearly to any temperature between 268 K and 378 K, achieving markedly improved prediction accuracy compared to established models across all studied temperatures. To further enhance predictive performance, the experimental database was expanded using active learning (AL) strategies for targeted acquisition of new diffusion data by pulsed-field gradient (PFG) NMR measurements. Diffusion coefficients at infinite dilution in 19 solute + solvent systems were measured at 298 K, 313 K, and 333 K. Incorporating these results yields a substantial improvement in the TCM's predictive accuracy. These findings highlight the potential of combining data-efficient ML methods with adaptive experimentation to advance predictive modeling of transport properties.

Received 27th February 2026,
Accepted 26th May 2026

DOI: 10.1039/d6cp00732e

rsc.li/pccp

Introduction

Diffusion is the fundamental process governing mass transport and determines the rate and selectivity of many processes in nature and technology. Despite their importance, experimental data on diffusion coefficients remain scarce, as there are simply too many mixtures of interest to study more than a tiny fraction of them through tedious diffusion measurements. Accordingly, the development of reliable prediction methods for diffusion coefficients is an important research topic,[1–8] which is currently evolving rapidly, driven by machine learning (ML).[9–11] We focus here on the demanding, practically important task of predicting diffusion coefficients in liquid mixtures.

There are two distinct classes of diffusion coefficients: *mutual* diffusion coefficients, which describe the collective motion of molecules driven by chemical-potential gradients, and *self*-diffusion coefficients, which characterize the Brownian motion of individual molecules.[12] Liquid-phase mutual diffusion is commonly modeled using either the Maxwell–Stefan or Fickian framework. Established measurement methods include diaphragm cells,[13] Taylor-dispersion experiments,[14] dynamic light scattering,[15] and concentration-profile monitoring in quiescent fluids.[16]

Pulsed-field gradient NMR allows a calibration-free and accurate self-diffusion measurement in both pure liquids and liquid mixtures.[7,17–19] It uses a short magnetic-field-gradient pulse to label nuclear spins with position-dependent phases, followed by a second gradient pulse, applied after a defined delay, which rephases the spins. Molecular diffusion during this delay leads to incomplete rephasing, resulting in a decrease in signal intensity that scales with the diffusion coefficient; specifically, the faster the diffusion, the greater the signal reduction.

The diffusion coefficient D_{ij}^{∞} of a solute i at infinite dilution in a solvent j is of particular interest for several reasons: at this limit, self- and mutual diffusion coefficients coincide, and the Maxwell–Stefan and Fickian descriptions become identical. Moreover, if both infinite-dilution coefficients in a binary mixture are known (i in j and j in i), an extrapolation to finite concentrations is possible, *e.g.*, *via* the empirical Vignes correlation,[20] with possible extension to multi-component systems.[12] The semi-empirical Stokes–Einstein–Gierer–Wirtz estimation (SEGWE) model[2] is currently the most accurate semi-empirical model for the prediction of D_{ij}^{∞}.

Laboratory of Engineering Thermodynamics, RPTU Kaiserslautern-Landau, Erwin-Schrödinger-Str. 44, 67663 Kaiserslautern, Germany. E-mail: fabian.jirasek@rptu.de

Recently, we have introduced matrix completion methods (MCMs) from ML, which are well established in recommender systems,[21] for predicting D_{ij}^{∞} at 298 K.[9] The key idea is to represent experimental data measured for different binary mixtures as a matrix whose rows and columns correspond to components i and j, with each entry containing the available data for mixture $i + j$.[22,23] Because this matrix is sparsely populated with experimental data, predicting the properties of unstudied mixtures reduces to a matrix completion problem. MCMs have since been developed for various thermodynamic properties, including activity coefficients,[22,24–28] Henry's law constants,[29,30] and diffusion coefficients,[9,31] as well as for pair-interaction parameters in thermodynamic models.[32–36]

For predicting D_{ij}^{∞}, hybrid approaches that incorporate prior physical knowledge from the SEGWE model[2] in the MCM training are especially promising, outperforming all available semi-empirical alternatives in prediction accuracy at 298 K.[9] However, because MCMs require a matrix structure in their training data, they are restricted to predicting a single property of binary mixtures under fixed conditions; for D_{ij}^{∞}, this means single-temperature predictions. Industrial practice, however, demands knowledge across a wide temperature range, not only at 298 K, where data are even sparser.

There are two general ways for extending MCMs to higher dimensions, *e.g.*, for predicting temperature-dependent thermodynamic properties. The first route is feasible if the temperature dependence of the property of interest is known. Then, the MCMs can be applied to predict the mixture-specific parameters of the equation describing the temperature dependence. This route was introduced by Damay *et al.*[25] for predicting temperature-dependent activity coefficients at infinite dilution using the Gibbs–Helmholtz relation. The second route is to extend the MCM to a tensor completion method (TCM), whereby a three-dimensional tensor is spanned by the two components that make up the mixtures and the temperature. The TCM concept was transferred to thermodynamics by Damay *et al.*,[37] who again considered the temperature-dependent prediction of activity coefficients at infinite dilution.

The TCM approach, unlike the first route, is also applicable when no general equation describing the temperature dependence is available. Liquid-phase diffusion coefficients are sometimes approximated as having a linear temperature dependence, consistent with the Stokes–Einstein theory[38] if the temperature dependence of the solvent viscosity is neglected. However, this is not generally applicable, making TCMs an interesting option in this field.

The predictive capabilities of an ML model can also be enhanced by purposefully incorporating new data into the training set through active learning (AL) methods. Such AL strategies iteratively select the presumed most informative data points for experimental measurement, without prior knowledge of their values, and thereby aim to maximize model improvement with minimal experimental effort. To this end, a query strategy is employed within an AL framework.[39] In a previous work, we found that uncertainty sampling, *i.e.*, selecting the data point to be measured based on the current model's largest uncertainty, is an effective query strategy for improving the performance of an MCM in predicting D_{ij}^{∞}.[31]

In this work, we present a novel hybrid TCM for predicting D_{ij}^{∞} across temperatures. Our method employs a Tucker decomposition,[40] analogous to that in the study by Damay *et al.*,[37] and integrates SEGWE priors, following our earlier MCM approach.[9] This TCM is trained on D_{ij}^{∞} data at 298 K, 313 K, and 333 K. While D_{ij}^{∞} is obviously of interest well beyond these three discrete temperatures, substantially less experimental information is available outside this range, preventing the development of MCMs for predicting D_{ij}^{∞} at these temperatures. However, the developed TCM can also predict temperatures absent from the training set, and we evaluate its predictions at temperatures between 268 K and 378 K, comparing them to experimental diffusion data and SEGWE[2] predictions within this extended range. The dependence of liquid-phase diffusion coefficients on the pressure is generally small, especially at low to moderate pressures. Since all experimental D_{ij}^{∞} values used in this work were reported at (or near) atmospheric pressure in the original literature, we neglect the influence of the pressure and note that the developed model should not be used to predict diffusion coefficients at very high pressures.

Furthermore, we extend the available experimental data on D_{ij}^{∞} at 298 K, 313 K, and 333 K by measuring D_{ij}^{∞} using pulsed-field gradient NMR spectroscopy[7,17,19] and selecting the measured systems using AL[31,39] and uncertainty sampling. We systematically evaluate the influence of the new training data on the prediction accuracy.

Methodology

Database

Experimental data for liquid-phase D_{ij}^{∞} at 298 K in binary mixtures were, on the one hand, taken from the database of Großmann *et al.*,[9] which is based on data from the Dortmund Data Bank (DDB) 2019[41] and several other sources. Most of the data for D_{ij}^{∞} reported by Großmann *et al.*[9] were obtained from an extrapolation of data at finite concentrations. After extending the database of Großmann *et al.*[9] with new D_{ij}^{∞} from the 2025 version of the DDB[41] in this work and data from our previous works,[19,31] the three temperatures for which by far the most data were available are 298 K, 313 K, and 333 K (*cf.* Fig. S1 in the SI). Thus, we carried out a comprehensive literature search for D_{ij}^{∞} at those temperatures, adding 89 additional data points at 313 K and 333 K from various sources.[19,31,41–60] In all cases, we used the extrapolation scheme of Großmann *et al.*[9] to obtain the D_{ij}^{∞} if the data were not directly reported at infinite dilution in the literature. To facilitate model evaluation *via* leave-one-out analysis, data were filtered to include only solutes i and solvents j that each occurred in at least two distinct solute–solvent pairs $i + j$ with available experimental D_{ij}^{∞} data, irrespective of temperature.

The resulting database of experimental values of D_{ij}^{∞} used in this work consists of 224 data points at 298± 1 K, 75 data points at 313 ± 1 K, and 56 data points at 333 ± 1 K. It covers 45

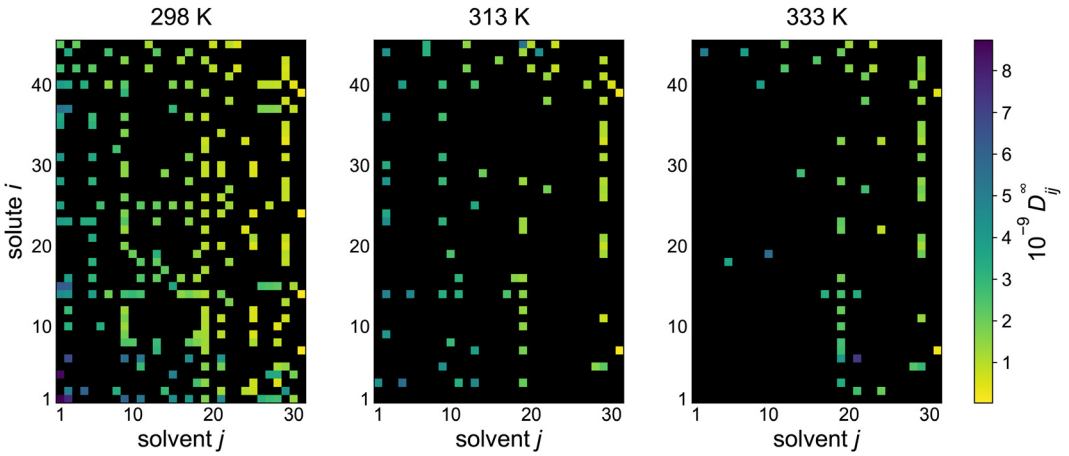

Fig. 1 Experimental data for liquid-phase diffusion coefficients D_{ij}^{∞} of solutes i at infinite dilution in solvents j at temperatures 298 K, 313 K, and 333 K in the database used as the starting point in the present work. Numbers identify solutes and solvents, *cf.* Tables S1 and S2 in the SI. Solutes are ordered with respect to their molar mass (bottom: low; top: high), and solvents are ordered with respect to their viscosity (left: low; right: high). The color code indicates the value of $D_{ij}^{\infty,\mathrm{exp}}$, and black cells denote missing data.

different solutes i infinitely diluted in 31 different solvents j. The data can be arranged in temperature-specific matrices, where the rows represent the solutes and the columns represent the solvents, *cf.* Fig. 1. The solutes and solvents included in the matrix are listed in Tables S1 and S2 in the SI. The included compounds consist mostly of water and organic molecules that are liquid under ambient conditions and have low reactivity. The solutes additionally contain 5 substances that are gaseous under ambient conditions. Values of D_{ij}^{∞} range from 10^{-11} to 10^{-8} m^2 s^{-1}.

The dataset can also be represented as a third-order tensor, with the three dimensions being the solutes, solvents, and temperatures. In total, this tensor has 4185 elements, of which, however, only 8.5% are occupied by experimental data for D_{ij}^{∞}. As shown in Fig. 1, this tensor is not only sparsely but also heterogeneously occupied. Most data are available at 298 K, and there are some solvents and solutes for which much more data are available than for the others; there are even several solvents and solutes for which no data are available at 313 K and 333 K at all. Details of the data availability per temperature are provided in Table 1.

To facilitate sample handling during the measurements planned by AL in this work, we further filtered the data from Fig. 1 to obtain a reduced dataset by excluding all gaseous compounds under ambient conditions. Due to this exclusion, we again had to filter data, so only solutes i and solvents j for which experimental data points for $D_{ij}^{\infty}(T)$ in at least two different mixtures $i + j$ were available were included. We chose to exclude these substances beforehand to follow the query strategy as closely as possible, rather than intervening during the AL workflow by skipping selected systems. The temperature-specific matrix arrangement of this reduced database is shown in Fig. 2 and is the underlying database for the experimental AL workflow. The solutes and solvents included in this matrix are listed in Tables S3 and S4 in the SI.

Table 1 Information on the availability of D_{ij}^{∞} in our database for each studied temperature, *cf.* Fig. 1

T	298 K	313 K	333 K
Number of data points	224	75	56
Matrix occupation rate	16.1%	5.4%	4.0%
Number of available solvents	31	24	18
Number of available solutes	45	35	33

While we focus on three temperatures here to compare TCM predictions with temperature-specific MCMs, a continuous-temperature approach, as explained in the following sections, enables generalization to arbitrary temperatures. To assess the performance of the TCM over the continuous temperature range, we use another dataset from the DDB 2025[41] spanning 268 K to 378 K (but excluding 298 K, 313 K, and 333 K) and containing 98 data points. No data for these temperatures were, however, used for training the TCM.

In addition to the experimental data, henceforth called $D_{ij}^{\infty,\mathrm{exp}}$, a synthetic database, $D_{ij}^{\infty,\mathrm{SEGWE}}$, was used for pre-training both temperature-specific MCMs and the TCM. This synthetic database consists of predictions of D_{ij}^{∞} at 298 K, 313 K, and 333 K using the SEGWE model[2] for the same solutes and solvents as in the experimental database. The solvent viscosities required for the SEGWE model[2] were obtained from the DDB 2025,[41] and the effective density (a parameter in the SEGWE model) was set to the recommended value ρ_{eff} = 627 kg m^{-3},[2] as it was done in our previous works.[9,31]

All experimental diffusion coefficient data were reported at (or near) ambient pressure in their original publications, as are the SEGWE[2] predictions. While pressure effects on liquid diffusion coefficients are generally much weaker than temperature effects, they may become significant at elevated pressures, which were not considered in this work.

Matrix completion method

In the present work, we have used a hybrid MCM that combines experimental data on D_{ij}^{∞} with synthetic data obtained using

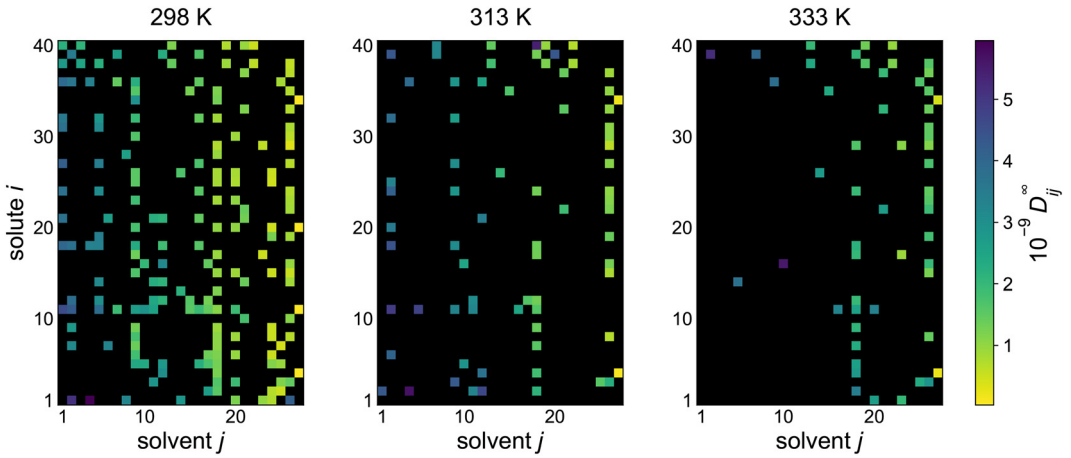

Fig. 2 Experimental data for liquid-phase diffusion coefficients $D_{ij}^{\infty,\mathrm{exp}}$ of solutes i at infinite dilution in solvents j at temperatures 298 K, 313 K, and 333 K in the database used as the starting point for the AL study. Numbers identify solutes and solvents, *cf.* Tables S3 and S4 in the SI. Solutes are ordered with respect to their molar mass (bottom: low; top: high), and solvents are ordered with respect to their viscosity (left: low; right: high). The color code indicates the value of D_{ij}^{∞}, and black cells denote missing data.

the SEGWE model during the training in a Bayesian framework and that was introduced and termed "MCM-Whisky" in our previous works[9,24,31] to predict isothermal D_{ij}^{∞} as a benchmark for comparison with the newly developed TCM, which predicts temperature-dependent D_{ij}^{∞}. This MCM is based on a low-rank matrix factorization of an isothermal D_{ij}^{∞} matrix, which is only sparsely populated with experimental data (*cf.* Fig. 1). The training of the MCM involves two steps with different (experimental or synthetic) training data; in both steps, the training data are modeled as

$$\ln D_{ij}^{\infty} = u_i \cdot v_j + \varepsilon_{ij} \tag{1}$$

where u_i and v_j are the component-specific feature vectors of solute i and solvent j, representing the fitting parameters of the model, and ε_{ij} are the deviations between model predictions and the training data. Both feature vectors have length $K = 2$, which is a hyperparameter of the model and was adapted from our previous work.[9,31]

In the first training step, an MCM is trained on the complete synthetic $\ln D_{ij}^{\infty,\mathrm{SEGWE}}$ data matrix according to eqn (1) using uninformed normal prior distributions with $\mu_0 = 0$ and $\sigma_0 = 1$ and a Cauchy likelihood with scale parameter $\lambda = 0.2$. The resulting preliminary features u_i^* and v_j^*, obtained by minimizing the residuals ε_{ij} and described by the posterior probability distributions from this first training step, were scaled and then used as informed normal prior distributions for the second MCM trained on the sparse $\ln D_{ij}^{\infty,\mathrm{exp}}$ matrix following eqn (1), again using a Cauchy likelihood with scale parameter $\lambda = 0.2$, minimizing the residuals ε_{ij} now referring to the experimental data, and resulting in the final solute and solvent features, u_i and v_j. For the scaling, the mean of the posterior distributions of u_i^* and v_j^* was adopted, whereas their standard deviation was scaled with a constant factor to obtain an average value (averaged over all solutes i and solvents j) of $\bar{\sigma} = 0.5$. The resulting distributions were finally multiplied by the uninformed normal prior ($\mu_0 = 0$ and $\sigma_0 = 1$) used in the first training step. This probabilistic hybrid approach allows prior physical information from the SEGWE model to be incorporated into the MCM, while maintaining the flexibility of the model to adapt to experimental data. More details on this hybrid approach can be found in our earlier work.[9,24,31]

Since we use a Bayesian approach for this second training step as well, we obtain posterior distributions over the model parameters after training, from which probability distributions for each predicted matrix entry can be calculated using eqn (1). The mean of these distributions was considered as the predicted diffusion coefficient $\ln D_{ij}^{\infty,\mathrm{pred}}$. Furthermore, from the obtained probability distributions, the standard deviation σ_{ij} was calculated as a measure for model uncertainty.

The MCM approach was used to predict isothermal diffusion coefficients. Hence, an individual MCM was trained on data for a single temperature, *i.e.*, a single matrix from Fig. 1, and generated predictions for the missing values at the same temperature, which was done here for 298 K, 313 K, and 333 K. It does not use information on the D_{ij}^{∞} at multiple temperatures and cannot extrapolate from one temperature to another. In the following, we will refer to the hybrid MCM approach simply as the MCM.

Tensor completion method

In this work, we introduce a novel hybrid tensor completion method for the temperature-dependent prediction of D_{ij}^{∞} as an extension of the hybrid temperature-specific MCM. For this purpose, we use a low-rank Tucker decomposition modeling the training data as

$$\ln D_{ij}^{\infty}(T) = \sum_{\alpha=1}^{r_u}\sum_{\beta=1}^{r_v}\sum_{\gamma=1}^{r_w} u_{i\alpha} \cdot v_{j\beta} \cdot w_{\gamma}(T) \cdot \kappa_{\alpha\beta\gamma} + \varepsilon_{ij}(T) \tag{2}$$

where u, v, and w are the latent features of the solute, the solvent, and the temperature, respectively; r_u, r_v, and r_w are their respective latent feature dimensions; and κ is the core tensor, which is used to combine the latent feature matrices

into a single tensor. We note that u and v are temperature-independent parameters, whereas w is temperature-dependent.

The TCM approach is *a priori* discrete with respect to temperatures and was trained and evaluated simultaneously at the three temperatures 298 K, 313 K, and 333 K. For each temperature, it learns an independent set of latent temperature features $w_\gamma(T)$, which contain no prior information about temperature and do not incorporate any physically motivated scaling. However, these learned features $w_\gamma(T)$ were subsequently correlated with the temperature T to enable their prediction at any temperature, not just the discrete ones. The correlation of the temperature features and the application of the TCM for predictions across a broad temperature range are discussed in the Results section.

Tucker decomposition, which becomes equivalent to canonical polyadic decomposition if κ is the unit tensor, was chosen because of its flexibility by introducing κ. Analogous to the MCM, we propose the hybrid TCM approach, which additionally incorporates SEGWE[2] predictions into its training. This approach consists of two steps, schematically shown in Fig. 3.

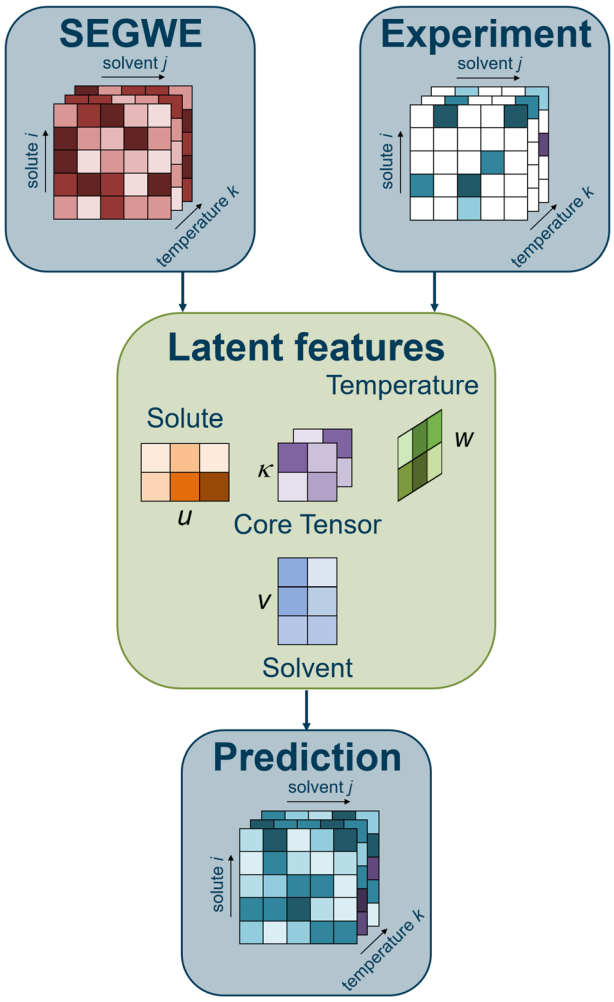

Fig. 3 Schematic representation of the hybrid TCM for predicting temperature-dependent D_{ij}^{∞} developed in this work. The TCM incorporates prior information from the SEGWE model[2] and uses the Tucker decomposition for tensor factorization.

Analogous to the MCM, the TCM is trained in two steps. First, the TCM is fitted to the fully completed synthetic tensor of $\ln D_{ij}^{\infty,\mathrm{SEGWE}}$ using uninformed normal prior distributions with $\mu_0 = 0$ and $\sigma_0 = 1$ and a Cauchy likelihood with scale parameter $\lambda = 0.2$. The posterior distributions of the latent features $u^\star$, $v^\star$, $w^\star$, and $\kappa^\star$ from this run then serve as priors for a second MCM on the sparse experimental tensor $\ln D_{ij}^{\infty,\mathrm{exp}}$ using a Cauchy likelihood with scale parameter $\lambda = 0.2$. For each feature, we keep the posterior mean and rescale its standard deviation by a constant so that the overall average becomes $\bar{\sigma} = 0.5$ (averaged over all i, j, and T). These informed priors are multiplied by the default uninformative prior ($\mu_0 = 0$ and $\sigma_0 = 1$). Following our earlier work,[9,31] this scheme injects prior physical knowledge from the SEGWE[2] model into the TCM while retaining flexibility to fit the experimental data.

$\ln D_{ij}^{\infty,\mathrm{pred}}$ are calculated analogously to MCM, from the posterior distributions of the model parameters, according to eqn (2). The mean of the resulting distribution for each tensor entry is taken as $\ln D_{ij}^{\infty,\mathrm{pred}}$, whereas their standard deviation $\sigma_{ij}(T)$ serves as a measure for model uncertainty.

The use of κ generally allows different latent feature dimensions r_u, r_v, and r_w, which are the hyperparameters of the model. We have carried out hyperparameter optimization in this work, using system-wise leave-one-out cross-validation, *cf.* below for details. The results of the hyperparameter study are given in Fig. S2 of the SI. We found that the best prediction accuracy was achieved using $r_u = r_v = r_w = 2$.

Because the training data are restricted to measurements performed at (or near) ambient pressure, the present TCM (and MCM) should only be applied at low to moderate pressures. Extrapolation to elevated pressures would require additional training data and incorporation of pressure as an explicit model dimension, which could be the subject of future work.

Active learning

The objective of AL is to enhance the predictive capabilities of an ML model by purposefully incorporating new data into the training set. Ideally, these data points are selected to maximize the model's performance gain without prior knowledge of their values. To this end, a query strategy is employed within an AL framework.[39]

In this work, the ML model to be improved is the TCM for predicting D_{ij}^{∞} in binary mixtures at 298 K, 313 K, and 333 K. For the AL, we thereby constrain the newly measured data and evaluations to the three discrete temperatures, *i.e.*, the possible solute–solvent–temperature tuples (i, j, T), thereby limiting the experimental space, which would otherwise be infinitely large due to continuous temperature. Consequently, we used a pool-based sampling approach, where all solute–solvent pairs (i, j) for which no $D_{ij}^{\infty,\mathrm{exp}}$ exist at any temperature T comprise the sampling pool $\mathcal{U}$, which contains the solute–solvent pairs from which the query strategy may choose new mixtures to be measured.

Fig. 4 shows the general AL framework used in this work, which was adopted from our previous work.[31]

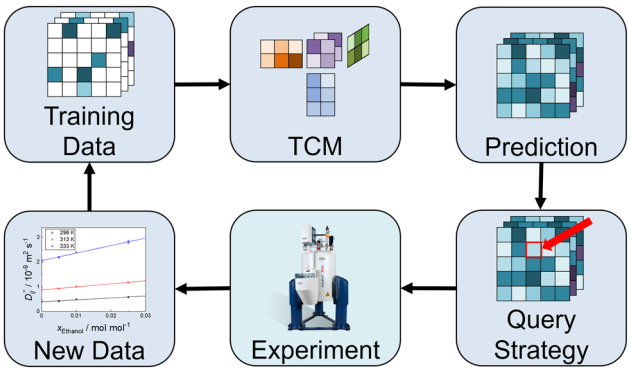

Fig. 4 Active learning workflow for the targeted improvement of the TCM developed in this work.

As illustrated in Fig. 4, the AL workflow is an iterative process. We begin with the initial training data set, *i.e.*, the initially available experimental data for D_{ij}^{∞}, *cf.* Fig. 1. The TCM is trained on this data set and can then be used to generate a complete tensor of predicted diffusion coefficients $D_{ij}^{\infty,\mathrm{pred}}$. Based on the obtained predictions, a query strategy is used to select a solute–solvent pair $(i,j)^* \in \mathcal{U}$ for which no experimental data are available at any temperature $T \in \Theta$, where $\Theta = \{298\ \mathrm{K}, 313\ \mathrm{K}, 333\ \mathrm{K}\}$. $D_{ij}^{\infty,\mathrm{exp}}\ \forall\ T \in \Theta$ are then measured for this selected system by PFG NMR spectroscopy at all temperatures T, and the new data are subsequently added to the training data set. This procedure is repeated several times, increasing the training data size at each iteration and thus (hopefully) improving the prediction accuracy of the model. The key to this improvement lies in choosing a suitable query strategy.

In our previous study, we found that uncertainty sampling was the most suitable query strategy for improving the prediction of diffusion coefficients with MCMs,[31] which is why we again use this strategy in this work for the TCM approach. For this purpose, we average the prediction uncertainty $\sigma_{ij}(T)$ resulting from our TCM over the three studied temperatures T. This results in a solute–solvent matrix of temperature-averaged uncertainties, from which the entry $(i, j)^*$ with the highest associated prediction uncertainty $\bar{\sigma}_{ij}$ was selected, *cf.* eqn (3), and the new diffusion coefficients $D_{ij}^{\infty,\mathrm{exp}}$ are measured using PFG NMR spectroscopy.

$$(i,j)^* = \underset{(i,j)}{\mathrm{argmax}}\ \frac{1}{|\Theta|}\sum_{T\in\Theta}\sigma_{ij}(T) \tag{3}$$

In practice, uncertainty sampling tends to sample outliers that are not representative of the underlying data distribution, which we also observed in our prior work.[31] While sampling some outliers can improve the model's prediction accuracy, continuously sampling them yields little new information and leads to redundancy.[61,62] Specifically, in the context of the Bayesian MCM (and TCM), after repeated sampling within a single row or column, *i.e.*, repeated sampling of the same solute i or solvent j, the information gain by inclusion of another data point in the same row or column is small. At the same time, the posterior for all other compounds can remain wide.[63] We thus introduce the simple rule of removing a compound from the sampling pool after it has been sampled in too many consecutive rounds. This approach encourages exploration of the chemical space and reduces redundancy in a simple way.

Computational details and evaluation

Bayesian inference was performed using automatic differentiation variational inference[64,65] implemented in the probabilistic programming language Stan[66] in its Python package CmdStanPy. The code is provided in the SI.

The predictive performance of the models was evaluated using leave-one-out analysis.[67] This procedure is stricter than a random train-test split of individual data points, since the model cannot use information for the excluded solute–solvent pair at any temperature. This evaluation therefore assesses prediction of unseen binary systems rather than interpolation of isolated missing entries in the tensor. Each model was trained on a subset of $D_{ij}^{\infty,\mathrm{exp}}$, which includes all available experimental data except for one binary system to be predicted. In the case of the MCM, this means that the training data included all experimental data for one specific temperature T, except for one solute–solvent pair (i, j), which was then predicted at the same temperature. In the case of the TCM, the training data included all experimental data for all three temperatures $T \in \Theta$, except for one solute–solvent pair (i, j), which was excluded at all temperatures and predicted at all temperatures. Thus, in all cases, the predictions were made on diffusion coefficients for truly unseen solute–solvent pairs (i,j).

To evaluate the prediction accuracy at each temperature $T \in \Theta$, we computed the absolute relative error ($\mathrm{ARE}_{ij}(T)$) for each data point. These per-point errors were calculated using

$$\mathrm{ARE}_{ij}(T) = \left|\frac{D_{ij}^{\infty,\mathrm{pred}}(T) - D_{ij}^{\infty,\mathrm{exp}}(T)}{D_{ij}^{\infty,\mathrm{exp}}(T)}\right| \tag{4}$$

These errors were aggregated over the set $\mathcal{L}(T)$, which contains all (i, j) pairs where experimental data are available at temperature T and reported as box plots. The temperature-specific relative mean absolute error (rMAE(T), *cf.* eqn (5)), and the relative mean squared error (rMSE(T), *cf.* eqn (6)), are also reported:

$$\mathrm{rMAE}(T) = \frac{1}{|\mathcal{L}(T)|}\sum_{(i,j)\in\mathcal{L}(T)}\left|\frac{D_{ij}^{\infty,\mathrm{pred}}(T) - D_{ij}^{\infty,\mathrm{exp}}(T)}{D_{ij}^{\infty,\mathrm{exp}}(T)}\right| \tag{5}$$

$$\mathrm{rMSE}(T) = \frac{1}{|\mathcal{L}(T)|}\sum_{(i,j)\in\mathcal{L}(T)}\left(\frac{D_{ij}^{\infty,\mathrm{pred}}(T) - D_{ij}^{\infty,\mathrm{exp}}(T)}{D_{ij}^{\infty,\mathrm{exp}}(T)}\right)^2 \tag{6}$$

Furthermore, to demonstrate the generalization of the TCM to continuous temperature values, we trained a TCM on all data for $\Theta = \{298\ \mathrm{K}, 313\ \mathrm{K}, 333\ \mathrm{K}\}$ and report the errors across the temperature range [268 K, 378 K] (excluding data at $T \in \Theta$) using a box plot with aggregated 10 K temperature bins. In all cases, we compare the TCM results to SEGWE predictions and, for $T \in \Theta$, also to isothermal MCMs, using the same error metrics.

Measurement of diffusion coefficients by PFG NMR spectroscopy

In this work, self-diffusion coefficients were measured using PFG NMR, following the method described in our previous works.[7,17,31] The experiments were conducted with a Bruker NMR spectrometer (magnet: Ascend 400, console: Avance III HD 400, probe: PABBO 5.0 mm) with a magnetic field strength of 9.4 T (proton resonance frequency: 400.13 MHz) and a maximum gradient of 0.45 T m^{-1}. The temperature control (uncertainty $\pm$0.1 K) was calibrated using a certified Pt-100 resistance thermometer (PTB, Braunschweig). The measurements were performed using 2.5 mm diffusion tubes (Deuteron GmbH) to minimize convection. The chemicals were used as received, at natural isotope abundance; details are provided in Table S5 of the SI.

The pulse sequence stebpgp1s,[68] a stimulated echo sequence with bipolar gradients, was used as implemented in TopSpin 3.6.5 (Bruker). The Stejskal–Tanner equation was used to calculate the self-diffusion coefficients D_i:[69]

$$\ln\left(\frac{I}{I_0}\right) = -D_i\gamma^2\delta^2\left(\Delta - \frac{\delta}{3} - \frac{\tau}{2}\right)g^2 \qquad (7)$$

Here, I is the signal intensity, I_0 is the intensity at the lowest gradient strength, γ is the gyromagnetic ratio, δ is the gradient duration, Δ is the diffusion time, τ is the correction for bipolar gradients, and g is the gradient strength. D_i was obtained by fitting the equation to the measured I/I_0 ratios using a least-squares approach with the Python package lmfit.[70] Peak integrals were evaluated manually using MNova (Mestrelab). The experimental uncertainty σ_i^{exp} was estimated from the root-mean-square error of the fit residuals, reported as a 95% confidence interval assuming a t-distribution. The uncertainty is indicated with the experimental results.

The pulse sequence parameters were Δ = 50 ms and τ = 0.2 ms. The gradient strengths g were varied from 0.023 to 0.431 T m^{-1} in eight increments with equal squared spacing. 32 scans were conducted at each increment. The gradient duration δ was adjusted (300–5000 μs) to ensure at least 80% signal attenuation from lowest to highest g.

Solutions with three different solute concentrations (0.005, 0.01, and 0.025 mol mol^{-1}) were gravimetrically prepared for each measured solute–solvent system and measured at three temperatures (298 K, 313 K, and 333 K) and ambient pressure. When multiple peaks were present for the same compound, their respective measured D_i values were averaged. The diffusion coefficients measured at the three concentrations were linearly extrapolated to infinite dilution of the solute to obtain $D_{ij}^{\infty,exp}(T)$. The overall uncertainty $\sigma_{ij}^{\infty,exp}(T)$ was calculated by combining propagated measurement errors and extrapolation uncertainty, reported as a 95% confidence interval assuming a t-distribution.

Results and discussion

Accuracy of diffusion coefficient prediction

In Fig. 5, the performance of the hybrid TCM developed in this work for predicting D_{ij}^{∞} is compared to that of the semiempirical SEGWE[2] model and that of the isothermal MCM models[9] in terms of the ARE_{ij} based on the discrete-temperature dataset compiled in this work from the literature. Note that the results of the MCMs at 313 K and 333 K shown in Fig. 5 include some predictions for systems containing components, for which no experimental data were part of the training set, which is a consequence of the leave-one-out analysis and the fact that the MCMs are trained only on data for a single temperature. Consequently, the MCMs could infer the latent features of some components only from the synthetic SEGWE data. In Fig. S3 in the SI, an analogous plot is shown, including only

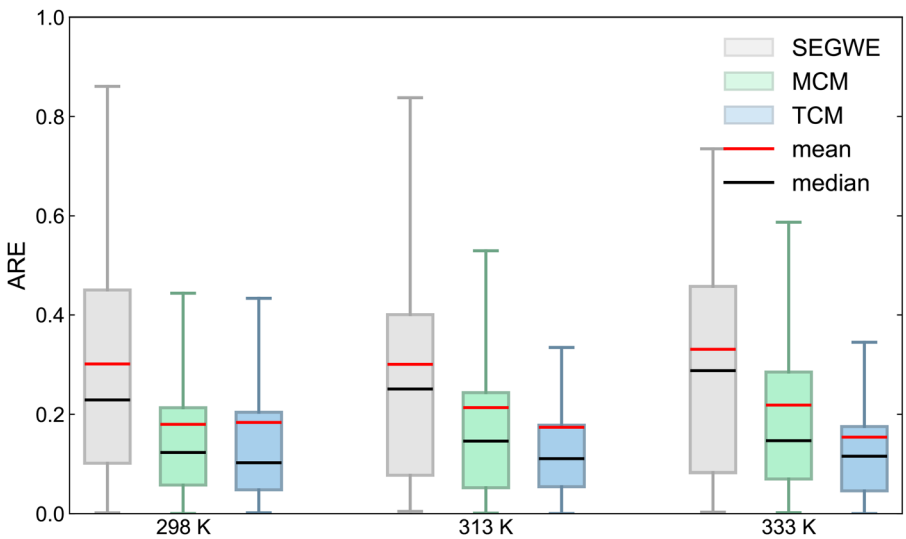

Fig. 5 Boxplot of the ARE_{ij} of the D_{ij}^{∞} predicted using the SEGWE model,[2] the MCM,[9] and the developed TCM. MCM and TCM results were obtained using leave-one-out analysis, and the SEGWE model was used as proposed by the original authors.[2] Boxes represent interquartile ranges (IQRs), and whiskers represent 1.5 IQR.

systems whose components appeared in all training data sets; the results are very similar.

Fig. 5 demonstrates that the MCMs (green) yield substantially lower prediction errors than SEGWE (red),[2] confirming previous results at 298 K reported by our group.[9] Notably, the MCM approach maintains superior performance over SEGWE, also at elevated temperatures (313 K and 333 K), despite the significantly lower availability of experimental data at these temperatures.

The TCM predictions (blue) exhibit lower error scores, including narrower interquartile ranges (IQRs) and 1.5 × IQR whiskers than both the SEGWE model and the MCM, indicating that the TCM predictions are more robust and have fewer outliers than the previously available methods. This robustness in predictive performance is further illustrated by the histograms of the relative prediction errors of D_{ij}^{∞} for each method and temperature shown in Fig. S4 in the SI.

The TCM developed in this work further improves the predictive accuracy over both the SEGWE model and the MCM across all three studied temperatures. This result is astonishing, as one could have expected a deterioration going from an individual fit for each temperature to a global fit over all temperatures. It is likely that the inclusion of additional training data across multiple temperatures allows the TCM to compensate for the more limited data sets at the higher temperatures, where substantially fewer measurements are available. This interpretation is supported by the larger performance gains observed at those temperatures. The results demonstrate that incorporating diffusion coefficient data across multiple temperatures into a model's training not only broadens the model's predictive scope but also enhances its accuracy at individual temperatures.

Analysis of the temperature features

Fig. 6 shows the latent temperature features w_γ (or rather their means) calculated using the TCM trained on the entire discrete data set as a function of T.

Fig. 6 shows a linear dependence of the discrete w_1 and w_2 (circles) on the temperature T. With the goal of predicting D_{ij}^{∞} across a continuous T range, we thus model the temperature dependence of w_γ using eqn (8):

$$w_\gamma(T) = A_\gamma + B_\gamma T \tag{8}$$

The linear regression statistics obtained from fitting eqn (8) to the discrete w_γ are detailed in Table 2, including the coefficient of determination (R^2) and mean squared error (MSE) of the fit.

Table 2 shows a very strong ($R^2 > 0.99$) linear correlation between the w_γ and T. Considering the temperature independence of the solute and solvent features u and v and the core tensor κ, this implies that within the considered temperature range $\ln D_{ij}^{\infty}$ is well-approximated as being linear in T. This result does not directly correlate with Stokes–Einstein theory[38] or SEGWE,[2] as they require the solvent viscosity, the temperature dependence of which is not easily described. Rather than implying a fundamentally linear temperature dependence, the more complex underlying dependence can be represented adequately by a linear approximation within the limited temperature range studied here.

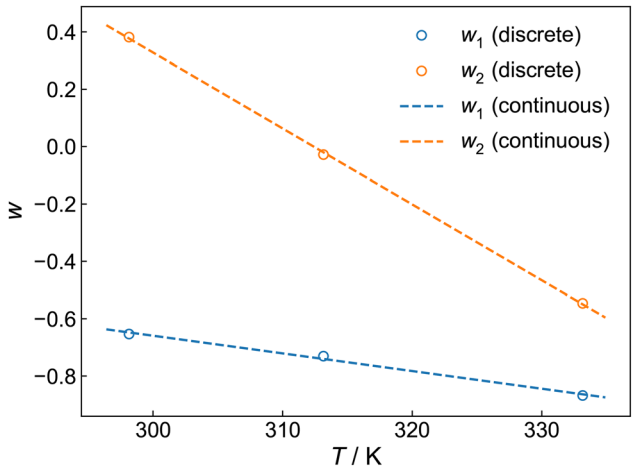

Fig. 6 Temperature features $w_1(T)$ and $w_2(T)$ of a TCM trained on the full discrete data set as a function of T and linear fits.

It is worth noting that the TCM learned this correlation only from the experimental values of D_{ij}^{∞} at 298 K, 313 K, and 333 K. No information on the actual physical temperature was provided, as was also the case for solutes and solvents. It is thus most astonishing that the TCM was able to learn a correlation between its parameters and the temperature purely from experimental data. Using this correlation, we predicted D_{ij}^{∞} for the same solute–solvent matrix of Fig. 1, at temperatures between 268 K and 378 K, the prediction error of which is shown as a function of T in Fig. 7, using 10 K temperature bins.

The TCM maintains high performance across the temperature range of 268 K to 378 K, with a total rMAE of 0.118, while the SEGWE model has a total rMAE of 0.263 for the same data set. Additionally, the TCM substantially outperforms SEGWE,[2] even at temperatures not present in the TCM's training set. It is most surprising that, despite the training data containing only a very small temperature range (35 K) and only three temperatures, the TCM can extrapolate easily to any unseen temperature in a much broader range (110 K), with barely any loss in accuracy. As expected with increasing temperatures, the prediction accuracy gradually worsens; thus, the linear scaling with T should be used only within the specified 268 K to 378 K range. To improve prediction accuracy at higher temperatures, alternative scaling and the inclusion of experimental data at higher temperatures could be used.

Improvement of the TCM by active learning

In Table 3, the experimental diffusion coefficients at infinite dilution $D_{ij}^{\infty,\mathrm{exp}}$ at 298 K, 313 K, and 333 K measured in this work by PFG NMR spectroscopy are reported with their

Table 2 Regression statistics obtained from fitting eqn (8) to the discrete w_γ, *cf.* Fig. 6

γ	A_γ	B_γ	R^2	MSE
1	1.195	-6.179×10^{-3}	0.9939	4.81×10^{-5}
2	8.271	-2.648×10^{-2}	0.9997	3.18×10^{-5}

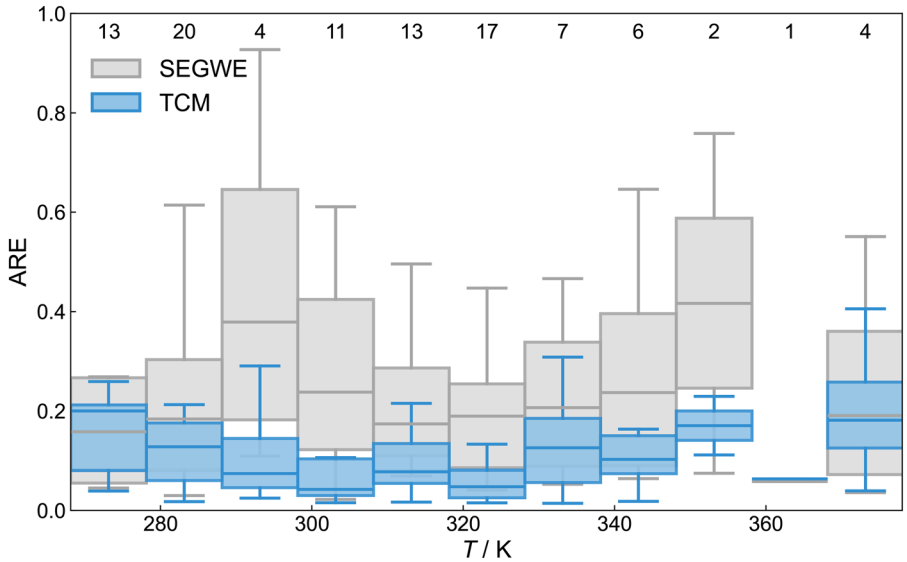

Fig. 7 Boxplot of the ARE_{ij} of the D_{ij}^{∞} predicted using SEGWE[2] and the TCM as a function of T. The numbers above each box indicate the number of data points per bin, horizontal lines represent the median, boxes represent the IQR, and whiskers represent 1.5 IQR.

respective uncertainties. The complete list of measured self-diffusion coefficients D_i at the studied finite concentrations, from which $D_{ij}^{\infty,\mathrm{exp}}$ was derived by extrapolation, is reported in the SI. In total, $D_{ij}^{\infty,\mathrm{exp}}$ in 19 different binary mixtures were measured.

As expected, $D_{ij}^{\infty,\mathrm{exp}}$ increases with increasing temperature for all systems studied. The experimental uncertainty shows significant variation. In some cases, the uncertainty is as high as 10% (partly even higher), mainly caused by the decreasing sensitivity of the NMR experiment at high temperatures in combination with low solute concentrations.

In the experimental implementation of the AL approach, we observed that some solvents were proposed particularly often by uncertainty sampling. In this study, as shown in Table 3, 1,2-propanediol was initially suggested most often, which we thus excluded from the sampling pool after the third measurement. The same effect was observed for dimethoxymethane towards the end of our measurements.

Using the entire set of new data for 19 previously unstudied mixtures for the training, the prediction rMSE decreased from 0.18 to 0.15 at 298 K, from 0.10 to 0.08 at 313 K, and from 0.07 to 0.06 at 333 K, while the occupation rate of the matrix increased only by 1.8%. The prediction rMAE and rMSE after each AL iteration are shown in the SI, Fig. S5. These results show that substantial improvements in the prediction of

Table 3 Liquid-phase diffusion coefficients at infinite dilution $D_{ij}^{\infty,\mathrm{exp}}$ measured by PFG NMR spectroscopy in this work, including experimental uncertainty $\sigma_{ij}^{\mathrm{exp}}$. All measurements were performed at ambient pressure. The temperature uncertainty is ±0.1 K across all temperatures. The systems are sorted according to the order in which they were selected by the AL strategy, measured, and subsequently included in the TCM training

			$D_{ij}^{\infty,\mathrm{exp}}/10^{-9}$ m^2 s^{-1}		
No.	Solute i	Solvent j	298 K	313 K	333 K
1	Methyl isopropyl ketone	1,2-Propanediol	0.052 ± 0.002	0.115 ± 0.009	0.248 ± 0.044
2	Butyl acetate	1,2-Propanediol	0.059 ± 0.002	0.122 ± 0.003	0.267 ± 0.006
3	Benzaldehyde	1,2-Propanediol	0.059 ± 0.009	0.127 ± 0.001	0.296 ± 0.022
4	Dimethoxymethane	Acetone	4.064 ± 0.025	4.894 ± 0.041	6.161 ± 0.102
5	Water	Dimethoxymethane	5.737 ± 0.132	7.085 ± 0.131	9.270 ± 0.177
6	2-Methyl-2,4-pentanediol	Acetone	5.019 ± 0.012	6.323 ± 0.024	8.356 ± 0.030
7	*m*-Cresol	Acetonitrile	2.675 ± 0.016	3.368 ± 0.072	4.357 ± 0.042
8	Water	2,4,6-Trioxaheptane	2.559 ± 0.087	3.249 ± 0.126	3.965 ± 0.049
9	Hexafluorobenzene	1-Butanol	0.896 ± 0.005	1.241 ± 0.008	1.828 ± 0.009
10	Chlorobenzene	Methyl isopropyl ketone	2.575 ± 0.005	3.140 ± 0.018	4.381 ± 0.200
11	Benzene	Butyl chloride	3.292 ± 0.021	3.957 ± 0.035	4.994 ± 0.058
12	Methyl isopropyl ketone	Acetone	3.736 ± 0.008	4.448 ± 0.020	5.495 ± 0.039
13	Glycerol	Acetonitrile	2.668 ± 0.039	3.219 ± 0.081	4.099 ± 0.042
14	Water	2,4,6,8,10-Pentaoxaundecane	0.742 ± 0.028	0.937 ± 0.046	1.276 ± 0.054
15	Water	2,4,6,8-Tetraoxanonane	1.327 ± 0.005	1.730 ± 0.034	2.351 ± 0.024
16	2,4,6-Trioxaheptane	Dimethoxymethane	3.464 ± 0.041	4.186 ± 0.026	4.869 ± 0.026
17	Butyric acid	Dimethoxymethane	2.381 ± 0.109	3.262 ± 0.015	4.175 ± 0.026
18	Di-*tert*-butyl sulfide	Dimethoxymethane	2.757 ± 0.008	3.347 ± 0.010	4.306 ± 0.025
19	Phenol	Dimethoxymethane	2.768 ± 0.012	3.418 ± 0.018	4.470 ± 0.045

diffusion coefficients can be achieved with only a few additional measurements, consistent with our previous findings.[31]

In the SI, we report the final TCM parameters obtained after training on the complete data set, comprising all literature data (*cf.* Fig. 1) and the new data measured in this work (*cf.* Table 3); these parameters should be used when applying the model to predict diffusion coefficients.

Conclusions

In this work, we have introduced a novel tensor completion method (TCM) for predicting the diffusion coefficients at infinite dilution, D_{ij}^{∞}, in binary systems at different temperatures. The method is trained on D_{ij}^{∞} data at 298 K, 313 K, and 333 K, but allows predictions at any temperature between 268 K and 378 K, thereby extending previously available matrix completion methods (MCMs), which were restricted to the isothermal case. The TCM achieves significantly higher prediction accuracies than the semi-empirical SEGWE model.[2] The global TCM model also provides better predictions than MCMs trained individually at different temperatures, indicating that the TCM benefits from joint training across different temperatures while simultaneously enabling generalization over a large temperature range.

Furthermore, the available data on D_{ij}^{∞} were extended through measurements using PFG NMR spectroscopy, in which the systems were selected using an active learning (AL) approach guided by the model's uncertainty. In total, 19 systems for which no prior data were available were measured at 298 K, 313 K, and 333 K. Even though this only increases the tensor's occupation rate by 1.8%, considerable improvements in prediction quality were observed. However, further improvements could be achieved by developing tailored query strategies in future work.

Author contributions

Zeno Romero: data curation, investigation, methodology, software, validation, visualization, and writing – original draft. Kerstin Münnemann: funding acquisition, resources, and writing – review & editing. Hans Hasse: funding acquisition, resources, supervision, and writing – review & editing. Fabian Jirasek: conceptualization, funding acquisition, resources, supervision, and writing – review & editing.

Conflicts of interest

There are no conflicts of interest to declare.

Data availability

Most of the experimental data on D_{ij}^{∞} used for training and testing in this work were used under license for this study; they are available directly from Dortmund Data Bank (DDB)[41] version 2025.

Additional experimental data for D_{ij}^{∞} found during our comprehensive literature study are available from their original sources.[19,31,41–60]

In the supplementary information (SI) of this work, we report the new diffusion data measured in this work, the complete Stan code used in processing the data sets in this work, as well as a set of parameters obtained using the TCM after training on all literature data (*cf.* Fig. 1) and the new data measured in this work. See DOI: https://doi.org/10.1039/d6cp00732e.

Acknowledgements

We gratefully acknowledge financial support by the Carl Zeiss Foundation in the frame of the project ''Process Engineering 4.0'' and by DFG in the frame of the Research Training Group GRK 2908 ''Valuable Wastewater (WERA)'' (grant number 503479768) and the Priority Program SPP 2363 ''Molecular Machine Learning'' (grant number 497201843). The authors acknowledge support from the Core Facility INST 248/370-1. Furthermore, FJ gratefully acknowledges financial support by DFG in the frame of the Emmy-Noether program (grant number 528649696).

References

1 J. R. Elliot, V. Diky, T. A. Knotts IV and W. V. Wilding, *The Properties of Gases and Liquids*, McGraw-Hill Professional, New York, NY, 17th edn, 2023.
2 R. Evans, G. D. Poggetto, M. Nilsson and G. A. Morris, Improving the interpretation of small molecule diffusion coefficients, *Anal. Chem.*, 2018, **90**, 3987–3994.
3 G. Guevara-Carrion, T. Janzen, Y. M. Muñoz Muñoz and J. Vrabec, Mutual diffusion of binary liquid mixtures containing methanol, ethanol, acetone, benzene, cyclohexane, toluene, and carbon tetrachloride, *J. Chem. Phys.*, 2016, **144**, 124501.
4 S. Parěz, G. Guevara-Carrion, H. Hasse and J. Vrabec, Mutual diffusion in the ternary mixture of water + methanol + ethanol and its binary subsystems, *Phys. Chem. Chem. Phys.*, 2013, **15**, 3985.
5 S. Schmitt, H. Hasse and S. Stephan, Entropy scaling for diffusion coefficients in fluid mixtures, *Nat. Commun.*, 2025, **16**, 2611.
6 C. J. Kankanamge, A. I. D. Zosel, T. Klein and A. P. Fröba, Prediction of Fick Diffusion Coefficients in Binary Electrolyte Mixtures, *Int. J. Thermophys.*, 2025, **46**, 142.
7 D. Bellaire, O. Großmann, K. Münnemann and H. Hasse, Diffusion coefficients at infinite dilution of carbon dioxide and methane in water, ethanol, cyclohexane, toluene, methanol, and acetone: A PFG-NMR and MD simulation study, *J. Chem. Thermodyn.*, 2022, **166**, 106691.
8 F. D. Lenahan, M. Piszko, T. Klein and A. P. Fröba, Prediction of Fick Diffusion Coefficients in Binary Mixtures of Liquids with Dissolved Gases at Infinite Dilution – A Review, *J. Chem. Eng. Data*, 2024, **69**, 692–702.

9 O. Großmann, D. Bellaire, N. Hayer, F. Jirasek and H. Hasse, 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.
10 F. Jirasek, N. Hayer, R. Abbas, B. Schmid and H. Hasse, Prediction of parameters of group contribution models of mixtures by matrix completion, *Phys. Chem. Chem. Phys.*, 2023, **25**, 1054–1062.
11 J. P. Aniceto, B. Zêzere and C. M. Silva, Prediction of diffusion coefficients in aqueous systems by machine learning models, *J. Mol. Liq.*, 2024, **405**, 125009.
12 R. Taylor and R. Krishna, Wiley Series in Chemical Engineering, *Multicomponent Mass Transfer*, John Wiley & Sons, Nashville, TN, 1993.
13 M. Tham, K. Bhatia and K. Gubbins, Steady-state method for studying diffusion of gases in liquids, *Chem. Eng. Sci.*, 1967, **22**, 309–311.
14 G. I. Taylor, Dispersion of soluble matter in solvent flowing slowly through a tube, *Proc. R. Soc. London, Ser. A*, 1953, **219**, 186–203.
15 R. D. Mountain and J. M. Deutch, Light scattering from binary solutions, *J. Chem. Phys.*, 1969, **50**, 1103–1108.
16 D. Bellaire, K. Münnemann and H. Hasse, Mutual diffusion coefficients from NMR imaging, *Chem. Eng. Sci.*, 2022, **255**, 117655.
17 D. Bellaire, H. Kiepfer, K. Münnemann and H. Hasse, PFG-NMR and MD simulation study of self-diffusion coefficients of binary and ternary mixtures containing cyclohexane, ethanol, acetone, and toluene, *J. Chem. Eng. Data*, 2020, **65**, 793–803.
18 T. Specht, K. Münnemann, H. Hasse and F. Jirasek, Rational method for defining and quantifying pseudo-components based on NMR spectroscopy, *Phys. Chem. Chem. Phys.*, 2023, **25**, 10288–10300.
19 S. Mross, S. Schmitt, S. Stephan, K. Münnemann and H. Hasse, Diffusion coefficients in mixtures of poly(oxymethylene) dimethyl ethers with alkanes, *Ind. Eng. Chem. Res.*, 2024, **63**, 1662–1669.
20 A. Vignes, Diffusion in binary solutions. Variation of diffusion coefficient with composition, *Ind. Eng. Chem. Fundam.*, 1966, **5**, 189–199.
21 Y. Koren, R. Bell and C. Volinsky, Matrix Factorization Techniques for Recommender Systems, *Computer*, 2009, **42**, 30–37.
22 F. Jirasek, R. A. S. Alves, J. Damay, R. A. Vandermeulen, R. Bamler, M. Bortz, S. Mandt, M. Kloft and H. Hasse, Machine learning in thermodynamics: Prediction of activity coefficients by matrix completion, *J. Phys. Chem. Lett.*, 2020, **11**, 981–985.
23 F. Jirasek and H. Hasse, Perspective: Machine learning of thermophysical properties, *Fluid Phase Equilib.*, 2021, **549**, 113206.
24 F. Jirasek, R. Bamler and S. Mandt, Hybridizing physical and data-driven prediction methods for physicochemical properties, *Chem. Commun.*, 2020, **56**, 12407–12410.
25 J. Damay, F. Jirasek, M. Kloft, M. Bortz and H. Hasse, Predicting activity coefficients at infinite dilution for varying temperatures by matrix completion, *Ind. Eng. Chem. Res.*, 2021, **60**, 14564–14578.
26 D. Gond, J.-T. Sohns, H. Leitte, H. Hasse and F. Jirasek, Hierarchical matrix completion for the prediction of properties of binary mixtures, *Comput. Chem. Eng.*, 2025, 109122.
27 N. Hayer, T. Specht, J. Arweiler, D. Gond, H. Hasse and F. Jirasek, Prediction of activity coefficients by similarity-based imputation using quantum-chemical descriptors, *Phys. Chem. Chem. Phys.*, 2025, **27**, 4307–4315.
28 J. Zenn, D. Gond, F. Jirasek and R. Bamler, Balancing molecular information and empirical data in the prediction of physicochemical properties, *Digital Discovery*, 2025, **4**, 683–693.
29 N. Hayer, F. Jirasek and H. Hasse, Prediction of Henry's law constants by matrix completion, *AIChE J.*, 2022, **68**, e17753.
30 N. Hayer, H. Hasse and F. Jirasek, Prediction of temperature-dependent Henry's law constants by matrix completion, *J. Phys. Chem. B*, 2024, **129**, 409–416.
31 Z. Romero, K. Münnemann, H. Hasse and F. Jirasek, Improvement of Diffusion Coefficient Prediction by Active Learning, *J. Phys. Chem. B*, 2025, **129**, 9219–9228.
32 F. Jirasek, R. Bamler, S. Fellenz, M. Bortz, M. Kloft, S. Mandt and H. Hasse, Making thermodynamic models of mixtures predictive by machine learning: matrix completion of pair interactions, *Chem. Sci.*, 2022, **13**, 4854–4862.
33 N. Hayer, T. Wendel, S. Mandt, H. Hasse and F. Jirasek, Advancing thermodynamic group-contribution methods by machine learning: UNIFAC 2.0, *Chem. Eng. J.*, 2025, **504**, 158667.
34 M. Hoffmann, N. Hayer, M. Kohns, F. Jirasek and H. Hasse, Prediction of pair interactions in mixtures by matrix completion, *Phys. Chem. Chem. Phys.*, 2024, **26**, 19390–19397.
35 F. Jirasek and H. Hasse, Combining Machine Learning with Physical Knowledge in Thermodynamic Modeling of Fluid Mixtures, *Annu. Rev. Chem. Biomol. Eng.*, 2023, **14**, 31–51.
36 F. Jirasek, N. Hayer, R. Abbas, B. Schmid and H. Hasse, Prediction of parameters of group contribution models of mixtures by matrix completion, *Phys. Chem. Chem. Phys.*, 2023, **25**, 1054–1062.
37 J. Damay, G. Ryzhakov, F. Jirasek, H. Hasse, I. Oseledets and M. Bortz, Predicting temperature-dependent activity coefficients at infinite dilution using tensor completion, *Chem. Ing. Tech.*, 2023, **95**, 1061–1069.
38 W. L. X. X. V. Sutherland, A dynamical theory of diffusion for non-electrolytes and the molecular mass of albumin, *London, Edinburgh Dublin Philos. Mag. J. Sci.*, 1905, **9**, 781–785.
39 B. Settles, *Active Learning Literature Survey; Computer Sciences Technical Report 1648*, 2009.
40 L. R. Tucker, Implications of factor analysis of three-way matrices for measurement of change, *Problems in measuring change*, 1963, vol. 15, p. 3.
41 Dortmund Data Bank. DDBST – Dortmund Data Bank Software And Separation Technology GmbH, Dortmund Data Bank, 2024. https://www.ddbst.com (accessed 2025-07-01).
42 T. Tominaga, S. Yamamoto and J.-I. Takanaka, Limiting interdiffusion coefficients of benzene, toluene, ethylbenzene and hexafluorobenzene in water from 298 to 368 K, *J. Chem. Soc., Faraday Trans. 1*, 1984, **80**, 941–947.

43 T. Tominaga and S. Matsumoto, Diffusion of polar and nonpolar molecules in water and ethanol, *Bull. Chem. Soc. Jpn.*, 1990, **63**, 533–537.
44 T. Tominaga and S. Matsumoto, Limiting interdiffusion coefficients of some hydroxylic compounds in water from 265 to 433 K, *J. Chem. Eng. Data*, 1990, **35**, 45–47.
45 Y.-P. Hsieh, R. B. Leron, A. N. Soriano, A. R. Caparanga and M.-H. Li, Diffusivity, density and viscosity of aqueous solutions of choline chloride/ethylene glycol and choline chloride/malonic acid, *J. Chem. Eng. Jpn.*, 2012, **45**, 939–947.
46 S. F. Li and H. M. Ong, Infinite dilution diffusion coefficients of several alcohols in water, *J. Chem. Eng. Data*, 1990, **35**, 136–137.
47 P. Schatzberg, Diffusion of water through hydrocarbon liquids, *J. Polym. Sci., Part C: Polym. Symp.*, 1965, 87–92.
48 I.-H. Lin and C.-S. Tan, Measurement of diffusion coefficients of p-chloronitrobenzene in CO2-expanded methanol, *J. Supercrit. Fluids*, 2008, **46**, 112–117.
49 R. L. Hurle and L. A. Woolf, Tracer diffusion in methanol and acetonitrile under pressure, *J. Chem. Soc., Faraday Trans. 1*, 1982, **78**, 2921–2928.
50 M. T. Tyn and W. F. Calus, Temperature and concentration dependence of mutual diffusion coefficients of some binary liquid systems, *J. Chem. Eng. Data*, 1975, **20**, 310–316.
51 S. A. Sanni, C. J. Fell and H. P. Hutchison, Diffusion coefficients and densities for binary organic liquid mixtures, *J. Chem. Eng. Data*, 1971, **16**, 424–427.
52 L. Bonoli and P. Witherspoon, Diffusion of aromatic and cycloparaffin hydrocarbons in water from 2 to 60, *J. Phys. Chem.*, 1968, **72**, 2532–2534.
53 M. J. te Riele, E. D. Snijder and W. P. van Swaaij, Diffusion coefficients at infinite dilution in water and in N-methylpyrrolidone, *J. Chem. Eng. Data*, 1995, **40**, 34–36.
54 A. Alizadeh and W. Wakeham, Mutual diffusion coefficients for binary mixtures of normal alkanes, *Int. J. Thermophys.*, 1982, **3**, 307–323.
55 D. Anderson, J. Hall and A. Babb, Mutual diffusion in nonideal binary liquid mixtures, *J. Phys. Chem.*, 1958, **62**, 404–408.
56 E. Hashim and H. Al-Shorachi, Diffusion coefficient of a binary liquid system, *Pet. Sci. Technol.*, 2007, **25**, 1519–1525.
57 E. Yumet, H.-C. Chen and S.-H. Chen, Tracer diffusion of carbon tetrachloride, S-trioxane, 12-crown-4, 15-crown-5, 18-crown-6 in acetonitrile, benzene, and chlorobenzene, *AIChE J.*, 1985, **31**, 76–81.
58 S.-H. Chen, D. Evans and H. Davis, Tracer diffusion in methanol, 1-butanol and 1-octanol from 298 to 433 K, *AIChE J.*, 1983, **29**, 640–645.
59 W. Clark and R. Rowley, The mutual diffusion coefficient of methanol–n-hexane near the consolute point, *AIChE J.*, 1986, **32**, 1125–1131.
60 J. Wagner, Z. Romero, K. Münnemann, T. Specht, F. Jirasek and H. Hasse, Thermodynamic modeling of poorly specified mixtures using NMR fingerprinting and group-contribution equations of state, *Fluid Phase Equilib.*, 2025, **596**, 114446.
61 D. D. Kim, *et al.*, Active learning in brain tumor segmentation with uncertainty sampling and annotation redundancy restriction, *J. Imaging Inf. Med.*, 2024, **37**, 2099–2107.
62 J. Zhu, H. Wang, B. K. Tsou and M. Ma, Active learning with sampling by uncertainty and density for data annotations, *IEEE Trans. Audio Speech Lang. Process.*, 2010, **18**, 1323–1331.
63 N. Houlsby, J. M. Hernandez-Lobato and Z. Ghahramani, Cold-start active learning with robust ordinal matrix factorization, Proceedings of the 31st International Conference on Machine Learning. Bejing, China, 2014, pp. 766–774.
64 A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman and D. M. Blei, Automatic differentiation variational inference, *J. Mach. Learn. Res.*, 2017, **18**, 1–45.
65 D. M. Blei, A. Kucukelbir and J. D. McAuliffe, Variational inference: A review for statisticians, *J. Am. Stat. Assoc.*, 2017, **112**, 859–877.
66 Stan Development Team, Stan Modeling Language Users Guide and Reference Manual, Version 2.35. https://mc-stan.org.
67 G. C. Cawley and N. L. Talbot, Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers, *Pattern Recogn.*, 2003, **36**, 2585–2592.
68 D. Wu, A. Chen and C. Johnson, Flow imaging by means of 1D pulsed-field-gradient NMR with application to electroosmotic flow, *J. Magn. Reson., Ser. A*, 1995, **115**, 123–126.
69 E. O. Stejskal and J. E. Tanner, Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient, *J. Chem. Phys.*, 1965, **42**, 288–292.
70 M. Newville, R. Otten, A. Nelson, T. Stensitzki, A. Ingargiola, D. Allan, A. Fox, F. Carter and M. Rawlik, *LMFIT: Non-Linear Least-Squares Minimization and Curve-Fitting for Python*. 2025, https://zenodo.org/doi/10.5281/zenodo.598352.