Improvement of Diffusion Coefficient Prediction by Active Learning

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

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

E-mail: fabian.jirasek@rptu.de

Abstract

Methods for predicting diffusion coefficients in mixtures are essential in many applications, as experimental data are scarce. Machine learning (ML) methods offer promising alternatives to established semi-empirical models for predicting diffusion coefficients, but their performance strongly depends on the available training data. Increasing the size of data sets is a straightforward strategy for improving ML methods, but measuring diffusion coefficients is costly, limiting the number of experiments that can be carried out. We have therefore studied active learning (AL) strategies for planning diffusion coefficient measurements and the targeted improvement of ML methods for their prediction, specifically matrix completion methods (MCMs) for predicting diffusion coefficients at infinite dilution D_{ij}^{∞} in binary mixtures at 298 K. In the first step, different AL strategies were systematically tested on a synthetic data set for D_{ij}^{∞}, and uncertainty sampling was found to be a simple but effective choice. This strategy was therefore used for planning D_{ij}^{∞} measurements using pulsed-field gradient (PFG) nuclear magnetic resonance (NMR) spectroscopy. In total, D_{ij}^{∞} in 19 mixtures were measured for which previously no data were available, and the data were used for retraining two hybrid MCMs. The results show that significant improvement in the

prediction of D_{ij}^{∞} can be achieved with only a few suitably planned experiments, but also that the impact strongly depends on the used prediction model: while no clear influence on the performance of an MCM that was trained on the residuals of the semi-empirical SEGWE model was found, the accuracy of a hybrid MCM that incorporates SEGWE predictions as soft prior information could be substantially increased, almost halving the relative mean squared error on the test set.

Introduction

Diffusion is important in many fields of science and engineering. However, despite their significance, experimental data on diffusion coefficients are extremely scarce, especially in mixtures. Therefore, models for the prediction of diffusion coefficients in mixtures are indispensable and have been the subject of numerous investigations, e.g.,[1–5] including the development and application of machine learning (ML) methods.[6]

Two classes of diffusion coefficients have to be distinguished: mutual and self-diffusion coefficients. While mutual diffusion coefficients describe the motion of the collective of molecules of the same component, driven by gradients of the chemical potential, self-diffusion coefficients describe the movement of individual molecules due to Brownian motion.[7] Depending on the model used to describe mutual diffusion, Maxwell-Stefan and Fickian diffusion coefficients are distinguished.

Established methods for measuring mutual diffusion coefficients in liquid mixtures include diaphragm cells,[8] Taylor dispersion,[9] and dynamic light scattering.[10] Mutual diffusion coefficients can also be determined by measuring the time dependence of concentration fields in a quiescent fluid.[11] However, mutual diffusion coefficients must generally be measured in transient systems and require precise concentration measurements with sufficient temporal resolution. Therefore, accurate mutual diffusion coefficient measurements are challenging.[12]

In contrast, self-diffusion coefficients can be measured reliably and accurately in pure components and mixtures using pulsed-field gradient (PFG) nuclear magnetic resonance

(NMR) spectroscopy.[3,13–15] PFG NMR is non-invasive and a primary, calibration-free measuring method for diffusion coefficients. It uses a short magnetic field gradient pulse to label nuclear spins in the sample with position-dependent phases. After a defined delay, a second gradient pulse is used to rephase these spins. If the molecules have diffused during this delay, the rephasing will be incomplete, leading to a measurable decrease in signal intensity, directly dependent on the diffusion coefficient; specifically, the higher the diffusion coefficient, the stronger the decrease in signal intensity.

Among the different diffusion coefficients, that of a solute i infinitely diluted in a solvent j is of particular importance, as in this state, the self-diffusion coefficient of component i is identical with the mutual diffusion coefficient in the mixture $i+j$, and the differences between the Maxwell-Stefan and the Fickian diffusion coefficient vanish. Furthermore, suppose both diffusion coefficients at infinite dilution in a binary mixture are known (i in j and j in i). In that case, one can often extrapolate to the mutual diffusion coefficients at finite concentrations, e.g., using the Vignes correlation,[16] and the approach can also be extended to multicomponent mixtures.[12]

Several semi-empirical models for predicting diffusion coefficients D_{ij}^{∞} of pure solutes i at infinite dilution in pure solvents j have been proposed in the literature. Poling *et al.*[1] gives an overview of the older methods. More recently, the Stokes-Einstein Gierer-Wirtz Estimation (SEGWE) model[2] has been proposed. As an alternative, we have recently developed a matrix completion method (MCM) from ML for predicting D_{ij}^{∞} at a 298 K,[6] which has been shown to outperform all available semi-empirical methods in terms of prediction accuracy. The MCM approach is based on the idea that experimental data for a given property that were measured in different binary mixtures can be conveniently represented in the form of a matrix, where the rows and columns represent the components i and j, and the entries contain the available data for the resulting binary mixtures $i + j$.[17,18] Since these matrices are only sparsely occupied by experimental data in basically all cases, the prediction of the properties of the unstudied mixtures becomes a matrix completion problem.

MCMs have been developed for the prediction of different thermodynamic properties, including activity coefficients,[17,19–23] Henry's law constants,[24,25] and diffusion coefficients,[6] as well as for pair interactions in thermodynamic models.[26–30] However, the performance of MCMs, like that of ML methods in general, heavily depends on the available training data, and measuring thermodynamic properties is generally time-consuming and expensive, so only a tiny fraction of all relevant mixtures can be studied in experiments. For these reasons, active learning (AL) strategies[31] are highly interesting to a priori select those experiments that promise to yield the most informative data points for improving a given model. Active learning has been successfully applied for efficient experimental design in different fields, such as catalyst development and materials discovery.[32,33] However, it has not yet been used to systematically improve ML models for predicting thermodynamic properties.

In this work, we have studied and employed AL strategies for improving the prediction of binary diffusion coefficients at infinite dilution D_{ij}^{∞} at 298 K with MCMs. In the first step, we systematically evaluated and compared different AL strategies on a synthetic data set. Subsequently, we have employed the best-performing strategy for the targeted measurement of D_{ij}^{∞} via PFG NMR spectroscopy. As the starting point, we have used the consolidated data set compiled by Großmann *et al.*[6] extended with new experimental data from Mross *et al.*[15] and the Dortmund Data Bank 2024.[34] Thus, our initial data set comprises 161 data points for D_{ij}^{∞} for 40 solutes i and 27 solvents j. Hence, we have data for 14.9 % of the 1080 possible elements of the matrix. We have then performed measurements of 19 previously unstudied mixtures and used the new data to refine MCMs for predicting the D_{ij}^{∞} at 298 K for the missing mixtures. While our goal was to improve the prediction of D_{ij}^{∞} at 298 K by MCMs, we also report results for D_{ij}^{∞} of the 19 studied mixtures for 313 K and 333 K, which were, however, not used in training the models. These additional data were measured and provided to have a more complete dataset for future modeling approaches. The results of the case study demonstrate the benefits of AL in thermodynamic model development for predicting diffusion coefficients. The approach can easily be transferred to ML methods to

predict other thermodynamic properties.

Methods

Database

The experimental data for D_{ij}^{∞} used in this work were mainly taken from the database compiled by Großmann *et al.*,[6] who have consolidated liquid-phase diffusion coefficient data in binary systems at 298 K from the Dortmund Data Bank[34] and a vast number of other sources. As only a few diffusion coefficients at infinite dilution are reported in the literature, data points for D_{ij}^{∞} had to be obtained from an extrapolation of data for D_{ij} at finite concentrations.[6] We have extended this database with recent experimental data at 298 K from Mross et al.[15] and new data from the Dortmund Data Bank 2024.[34]

During preprocessing, this database was reduced as follows: First, all solutes and solvents that are not liquid under ambient conditions were removed to facilitate sample handling during the experiments. Furthermore, the dataset was reduced to include only solutes i and solvents j for which experimental data points for D_{ij}^{∞} in at least two different mixtures $i+j$ were available. This was done to ensure that each solute and solvent could be included in the test set while keeping at least one data point per solute and solvent in the training set to train the MCM.

The resulting database was used as the starting point of this work and consists of 161 experimental liquid-phase diffusion coefficients $D_{ij,\mathrm{exp}}^{\infty}$ at 298 ± 1 K of 40 different solutes i infinitely diluted in 27 different solvents j. The data for $D_{ij,\mathrm{exp}}^{\infty}$ were arranged as depicted in Figure 1, where the rows represent the solutes and the columns represent the solvents. The solutes and solvents included in the matrix are described in Table S1 in Supporting Information. The matrix has 1080 elements, of which only 14.9 % are occupied by experimental data. It can be seen from Figure 1 that the matrix is not only sparsely occupied, but the entries are also very unevenly distributed: there are a few solvents and solutes for which

much more data are available than for the others. The value of $D^{\infty}_{ij,\mathrm{exp}}$ is plotted over the solvent i and solute j for a few exemplary systems in the Supporting Information.

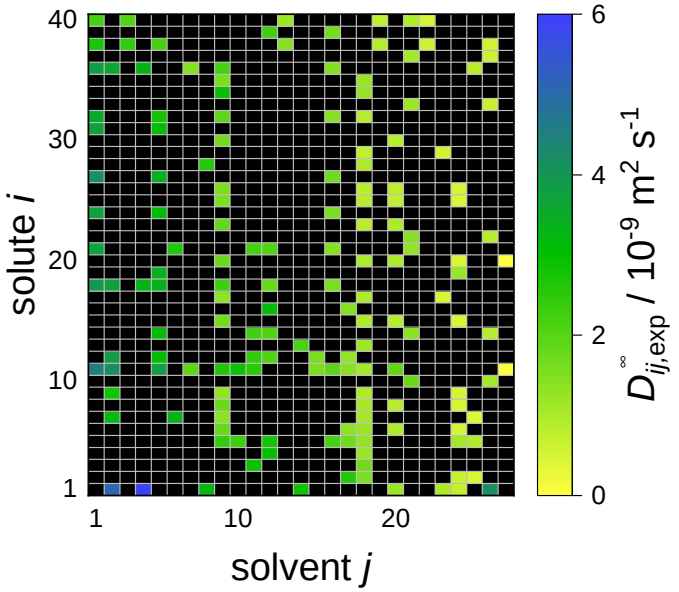


Figure 1: Experimental data for liquid-phase diffusion coefficients $D^{\infty}_{ij,\mathrm{exp}}$ of solutes i in solvents j at infinite dilution at 298 ± 1 K in the database used as starting point in the present work. Numbers identify solutes and solvents, cf. Supporting Information. Solutes are ordered 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 $D^{\infty}_{ij,\mathrm{exp}}$, and black cells denote missing data.

In addition to this sparsely occupied experimental database, a synthetic database for the same solutes and solvents was generated and used to study different AL strategies. This synthetic database consists of predictions for $D^{\infty}_{ij,\mathrm{exp}}$ at 298K with the SEGWE model, to which random noise was added by multiplying the SEGWE predictions with factors randomly drawn from a uniform distribution in the interval $[0.975, 1.025]$, which was done to simulate measurement errors in the order of ±2.5 %. The solvent viscosities required for the SEGWE model[2] were calculated using the correlations and parameters from the DIPPR database,[35] and the effective density (a parameter in the SEGWE model) was set to the recommended value $\rho_{\mathrm{eff}} = 627$ kg m^{-3}.[2] Figure 2 shows the resulting synthetic database.

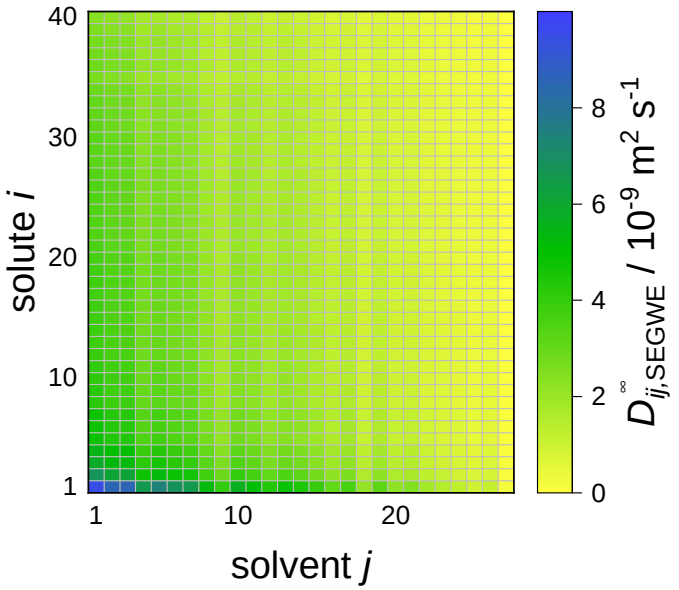


Figure 2: Synthetic database of $D^{\infty}_{ij,\mathrm{SEGWE}}$ at 298 K based on predictions from the SEGWE model[2] with added noise as described in the text. Solutes and solvents are sorted in the same way as in Figure 1. The color code indicates the value of $D^{\infty}_{ij,\mathrm{SEGWE}}$.

Matrix Completion Methods

In the present work, we have used three different MCMs for predicting D^{∞}_{ij}: one purely data-driven MCM and the two hybrid methods called MCM-Whisky and MCM-Boosting using a notation introduced in previous work on activity coefficients at infinite dilution.[17] All MCMs aim to learn the underlying structure within a matrix sparsely populated with available entries M_{ij}. For this purpose, a low-rank matrix factorization was used in all cases following (1):

$$M_{ij} = u_i \cdot v_j + \varepsilon_{ij} \tag{1}$$

where u_i and v_j are feature vectors of length K of the solute i and the solvent j, respectively, which contain the parameters of the MCM that are fitted to reproduce the training data M_{ij} by minimizing the errors ε_{ij}. These parameters are inferred using a Bayesian approach, where all data and parameters are modeled as random variables following a prob-

ability distribution. Bayesian inference was performed using automatic differentiation variational inference[36,37] implemented in the probabilistic programming language Stan[38] in its Python package CmdStanPy. The code is provided in the Supporting Information.

As a result of the training process, so-called posterior distributions over the model parameters are obtained, from which, in turn, probability distributions for each matrix entry to be predicted can be calculated using equation 1. The mean of these distributions was considered as the predicted matrix entry $M_{ij,\mathrm{pred}}$. Furthermore, from the obtained probability distributions, the standard deviation σ_{ij} was calculated as a measure for model uncertainty. In the present work, the same hyperparameters as in the original work[6] were used: As likelihood, a Cauchy distribution with scale $\lambda = 0.2$ centered around $u_i \cdot v_j$ and a feature vector length of $K = 2$ was used. As the size of the dataset increases, in particular, if additional solvents and/or solutes are added, this small feature vector length may, at some point, become insufficient to capture all relevant interactions. However, in prior work on applying MCMs to predict other thermophysical properties, where the data sets were significantly larger, we also found that small, though slightly larger feature vectors ($K = 3$ or $K = 4$) are sufficient.[17,24] Hence, we would not expect the immediate necessity of using significantly larger feature vectors if the data set on diffusion coefficients increases. For more details on the MCM and its training process, we refer to our original publication.[6] The three MCMs used in this work are described in more detail below.

Data-driven MCM: In the purely data-driven approach, the MCM is trained directly on the logarithmic $D^{\infty}_{ij,\mathrm{exp}}$, as described in equation (2):

$$\ln\left(D^{\infty}_{ij,\mathrm{exp}} \,/\, 10^{-9}\ \mathrm{m^2\ s^{-1}}\right) = u_i \cdot v_j + \varepsilon_{ij} \tag{2}$$

whereby for all MCM parameters, a normal distribution with mean $\mu_0 = 0$ and standard deviation $\sigma_0 = 1$ was used as prior.

MCM-Boosting: In the Boosting approach, an MCM is trained on the deviations

between SEGWE predictions and experimental data. The result of the MCM is then used to correct the SEGWE prediction. In our Boosting method, the MCM is trained on the residues between the logarithmic experimental diffusion coefficients $D^{\infty}_{ij,\mathrm{exp}}$ and the respective predictions with the SEGWE model $D^{\infty}_{ij,\mathrm{SEGWE}}$, see equation (3):

$$\ln D^{\infty}_{ij,\mathrm{SEGWE}} - \ln D^{\infty}_{ij,\mathrm{exp}} = u_i \cdot v_j + \varepsilon_{ij} \tag{3}$$

whereby, again, a normal distribution with mean $\mu_0 = 0$ and standard deviation $\sigma_0 = 1$ was used as prior for all MCM parameters.

MCM-Whisky: The MCM-Whisky approach consists of two steps. In the first step, an MCM is trained analogously to the data-driven MCM described above, but on the complete synthetic $\ln D^{\infty}_{ij,\mathrm{SEGWE}}$ data instead of the experimental data. The resulting preliminary features u_i^* and v_j^*, described by the posterior probability distributions from this first training step, are scaled and then used as informed prior distributions for a second MCM trained on the sparse $\ln D^{\infty}_{ij,\mathrm{exp}}$ matrix. 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 with the uninformed prior ($\mu_0 = 0$, $\sigma_0 = 1$) used in the data-driven MCM. This approach, which followed our previous work,[6,24] enables incorporating information from the SEGWE model into an MCM while preserving the model's flexibility to adjust to the experimental data.

Evaluation of Model Performance

The predictive performance of the MCMs was evaluated using predefined test sets, which were withheld during the training of the models. The test data were selected randomly, but ensuring that at least one data point for each solute i and each solvent j remained in the training set so that the MCM could learn the features of all components. The prediction

error on the test set is evaluated in the form of the relative mean absolute error (rMAE) and the relative mean squared error (rMSE), which were calculated following equations (4) and (5):

$$\text{rMAE} = \frac{1}{|\mathfrak{T}|} \sum_{(i,j)\in\mathfrak{T}} \left| \frac{D^{\infty}_{ij,\text{pred}} - D^{\infty}_{ij,\text{exp}}}{D^{\infty}_{ij,\text{exp}}} \right| \tag{4}$$

$$\text{rMSE} = \frac{1}{|\mathfrak{T}|} \sum_{(i,j)\in\mathfrak{T}} \left(\frac{D^{\infty}_{ij,\text{pred}} - D^{\infty}_{ij,\text{exp}}}{D^{\infty}_{ij,\text{exp}}} \right)^2 \tag{5}$$

Here, $\mathfrak{T}$ is the set of "test indices", i.e., the set of matrix indices (i, j) in the test set, for which experimental diffusion coefficient data $D^{\infty}_{ij,\text{exp}}$ exist but were not included in the training of the model, $|\mathfrak{T}|$ is the number of elements in the test set, and $D^{\infty}_{ij,\text{pred}}$ is the predicted diffusion coefficient.

Active Learning

General Framework

The goal of active learning (AL) is to improve the predictive performance of an ML model by purposefully adding new data to the training set. Ideally, these data points are chosen to lead to the highest possible performance gain of the model without knowing their values in advance. For this purpose, so-called query strategies are used within an AL framework.

In our case, the ML model that is to be improved is an MCM for predicting D^{∞}_{ij} in binary mixtures at 298 K. Therefore, the predictions are constrained to the matrix elements, i.e., the possible solute-solvent pairs. Hence, we consider a pool-based sampling approach, where all matrix entries for which no $D^{\infty}_{ij,\text{exp}}$ exists comprise the sampling pool $\mathfrak{U}$, containing the solute-solvent pairs from which the query strategy may sample, i.e., those mixtures that can be chosen to be measured. The general AL framework used in this work is depicted in Figure 3.

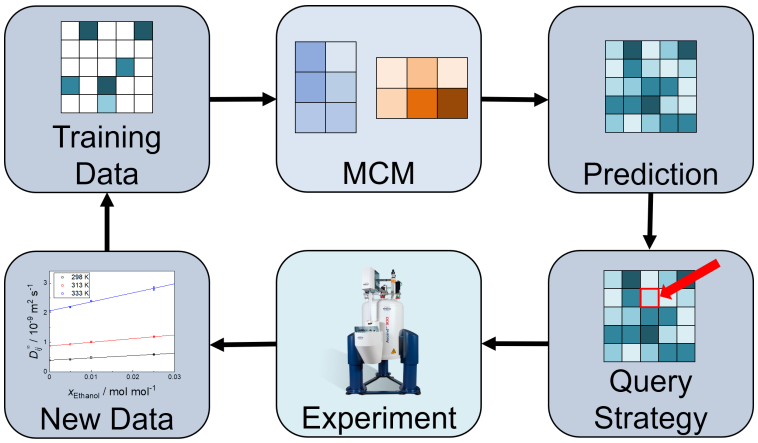


Figure 3: Active learning workflow used in this work. For an explanation of the different steps, see the text.

As shown in Figure 3, the AL workflow is an iterative process. We begin with an initial training data set (top left). The MCM is trained on this data set (top middle) to generate a complete matrix of predicted diffusion coefficients $D^{\infty}_{ij,\mathrm{pred}}$ (top right). Based on the obtained predictions, a query strategy is used to select a solute-solvent pair for which no experimental data are available (bottom right). The selected system is then studied by PFG NMR spectroscopy (bottom middle) to determine a new $D^{\infty}_{ij,\mathrm{exp}}$ (bottom left), which is subsequently added to the training data set. This procedure is repeated, continuously increasing the training data size at each iteration and thus (ideally) improving the prediction accuracy of the model. Key to this improvement is the choice of a suitable query strategy. AL has already been applied to MCMs in several fields;[33,39] however, to our knowledge, its application to thermophysical mixture property prediction with MCMs is new, and systematic comparisons of different query strategies for this scenario were not available. Thus, we first investigated different query strategies using a synthetic database, as described below.

Comparison of Query Strategies on Synthetic Data

Different query strategies were evaluated based on the completed noisy synthetic matrix of D^{∞}_{ij} at 298 K shown in Figure 2. In the first step, the data in the matrix (1080 data points) were randomly split into three subsets: an initial training set containing 15 % of

the synthetic data, a sampling pool containing 70 % of the data, and a test set containing the remaining 15 % of the synthetic data. We used the noisy data for all three data sets for better comparability with the experimental scenario, where training and test data are naturally noisy due to experimental uncertainties. The training data set was used to train the data-driven MCM. Since we used the SEGWE model predictions as the synthetic data set, the two hybrid models, MCM-Boosting and MCM-Whisky, could not be used here as they include the SEGWE predictions in their training and, hence, are already informed about the complete synthetic data set. Consequently, only the data-driven MCM was used for the synthetic study.

Based on the predictions obtained in each iteration and using one of the studied query strategies (see below), one data point from the sampling pool was selected and added to the training set, and the MCM was retrained. After each iteration, the MCM predictions were evaluated on the test set in terms of rMAE and rMSE. The whole process was carried out until all data points from the sampling pool had been selected. This process was repeated six times (whereby the randomly selected data in the three sets differed each time) for each query strategy using different data splits to increase the robustness of the results. In the data splits, it was ensured that each data point was not more than once in the initial training set and in the test set. The rMAE and rMSE obtained on the test data in each of the six runs were finally averaged to yield the final scores for each query strategy.

We have tested five query strategies, which are explained in the following. The data point selected from the sampling pool in each iteration is thereby indicated by $(i, j)^*$, where i and j represent the solutes and solvents, respectively, of the considered matrix containing the synthetic D_{ij}^{∞}. Some query strategies do not select a single data point in each step. In such a case, one of the selections was chosen randomly.

- **Random Sampling**: A new data point is chosen randomly from the sampling pool. As this is not a targeted strategy, it is considered an edge case here that all other query strategies should outperform.

- **Uncertainty Sampling**: Following this query strategy, the unstudied solute-solvent pair $(i, j)^*$ whose D_{ij}^{∞} is predicted with the highest uncertainty with the current model is selected. In this work, the standard deviation σ_{ij} of the MCM result for D_{ij}^{∞} trained in each round was used as the measure of the model uncertainty, cf. equation (6).[31]

$$(i, j)^* = \underset{(i,j)}{\operatorname{argmax}}\, \sigma_{ij} \tag{6}$$

- **Maximum Entropy Sampling**: This query strategy aims to maximize the homogeneity of the matrix in terms of available entries. Technically, in each step, the number of unavailable data points for each solute (M) and each solvent (N) in the training set is counted, and the unavailable data point with the largest sum, $M + N$, is selected. This procedure is described in equation (7). While this approach does not directly compute the informational entropy defined by Shannon,[40] the idea of maximizing the homogeneity of the matrix maximizes the Shannon entropy of the solute- and solvent-coverage distributions, i.e., the distribution of available data points across each solute and each solvent. This is further detailed in the Supporting Information.

$$(i, j)^* = \underset{(i,j)}{\operatorname{argmax}}\, |\{(x, y) \in \mathfrak{U} \mid x = i \vee y = j\}| \tag{7}$$

- **Query-by-Committee**: The core of this strategy is to identify unavailable data points whose predictions from a "committee" of models disagree most.[31] Our committee consisted of MCMs trained on three different subsets of the currently available training data, containing 85 % of the total training data each while ensuring that every available data point is contained in at least one of those subsets. The data point that is predicted with the highest absolute deviation from the mean of the three models'

predictions $M_{ij}^{(c)}$ is then selected from the sampling pool, as given in equation (8).

$$(i,j)^* = \underset{(i,j)}{\operatorname{argmax}} \sum_{c=1}^{3} \left(\hat{M}_{ij}^{(c)} - \frac{1}{3} \sum_{c=1}^{3} \hat{M}_{ij}^{(c)} \right)^2 \tag{8}$$

- **Expected Uncertainty Reduction**: The goal of expected uncertainty reduction is to sample the data point that leads to the largest decrease in model uncertainty. For this purpose, an estimation for each data point from the sampling pool is required, which is added to the current training data set $\mathfrak{L}$, used for training the model and assessing the overall model uncertainty (for all data points). This procedure is repeated for all data points from the sampling pool, and the one that led to the most significant overall reduction of the model uncertainty is selected.

  There are two general ways to estimate the data points from the sampling pool. Ideally, an alternative, independent prediction is available. If this is not the case, the standard approach is to use the predictions of the model of interest trained on the current training data set $\mathfrak{L}$.[31] In this work, we have used SEGWE predictions multiplied with a noise factor randomly chosen from the interval $[0.8, 1.2]$ as an estimate for the data points from the sampling pool. The larger noise factor compared to the one used for generating the completed synthetic data set was chosen to make our scenario more realistic: in our case, the ground truth (complete synthetic data set of SEGWE predictions +2.5% random noise) is known, which is, however, not the case in real applications where, at best, a suitable prediction method is available.

  Based on the SEGWE prediction with noise, we have chosen whichever data point led to the lowest total prediction uncertainty of a model trained on $\mathfrak{L} \cup \{(x,y)\}$, as described in equation (9). Here, $\sigma_{ij}^{(x,y)}$ is the prediction uncertainty at index (i,j) for

a model trained on data points $\mathfrak{L} \cup \{(x, y)\}$.[31]

$$(i, j)^* = \underset{(x,y)}{\operatorname{argmin}} \sum_{i=1}^{I} \sum_{j=1}^{J} \sigma_{ij}^{(x,y)} \tag{9}$$

For comparison, we have also determined the *optimal* sampling strategy by testing, in each iteration, all possible data points from the sampling pool and choosing the one that leads to the largest reduction of the rMAE. Obviously, the optimal sampling strategy can only be applied to the synthetic data set, where the ground truth is completely known a priori. In contrast, in the experimental setting, we do not know the data in the sampling pool and test set in advance.

Active Learning for Design of Experiments

The initial training set for the experimental scenario contained only 90 % (145 data points) of the available experimental data for D_{ij}^{∞}. 10 % (16 data points) of the data were withheld for testing. The test data points were chosen randomly and are listed in the Supporting Information. In each step of the AL procedure, the prediction accuracy of MCM-Boosting and MCM-Whisky was evaluated in terms of rMAE and rMSE on this test set.

Only one MCM and one query strategy can be used to plan the measurements of D_{ij}^{∞} in the experimental study. Based on the results for the synthetic data set, cf. Results and Discussion below, the uncertainty sampling strategy was used. As MCM, we have chosen to use MCM-Boosting, since it performed slightly better than MCM-Whisky for the prediction of D_{ij}^{∞} in the original publication.[6] We, however, also investigated and reported the evolution of the performance of MCM-Whisky using the newly measured data points.

Measurement of Diffusion Coefficients by PFG NMR Spectroscopy

Self-diffusion coefficients in binary liquid mixtures were measured by PFG NMR spectroscopy, as described in.[3,13] We have used 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 corresponding to a proton resonance frequency of 400.13 MHz and a maximum gradient of 0.45 T m^{-1}. Its temperature control unit was calibrated to 0.1 K uncertainty using a platinum resistance thermometer (Pt-100), which had been calibrated with a certified standard (PTB, Braunschweig). All measurements were carried out at 298 K, 313 K, and 333 K. All samples were measured in diffusion tubes with a reduced outer diameter of 2.5 mm (Deuteron GmbH) to minimize convection. All chemicals studied were used as commercially available at natural isotope abundance. The chemicals, their suppliers, and purities are listed in Table S4 in the Supporting Information.

The self-diffusion coefficients D_i of solute i in the solvent j were determined using the pulse sequence "stebpgp1s",[41] a stimulated echo pulse sequence with bipolar gradients, implemented in TopSpin 3.6.5 (Bruker). The self-diffusion coefficients of the solvent were not evaluated since they are not at infinite dilution. If possible, the observed nucleus was ^{1}H. One of the studied solutes, Hexafluorobenzene, does not contain hydrogen; in this case, the ^{19}F NMR spectra were used instead. The NMR results were evaluated using the Stejskal-Tanner equation.[42]

$$\ln\left(\frac{I}{I_0}\right) = -D_i\gamma^2\delta^2\left(\Delta - \frac{\delta}{3} - \frac{\tau}{2}\right)g^2 \tag{10}$$

where I is the peak integral of the diffusing component, I_0 is its peak integral at the lowest gradient strength, γ is the gyromagnetic ratio of the observed nucleus, δ is the duration of the gradient pulse, Δ is the diffusion time, τ is a correction constant for the application of bipolar gradients, and g is the gradient strength. Peak integrals were evaluated manually using MNova (Mestrelab). In practice, D_i was determined by fitting eq. (10) to the experimental I and I_0. The experimental uncertainty $\sigma_{i,\mathrm{exp}}$ of D_i was calculated from the root mean squared error of the residuals of this fit in the form of a 95 % confidence interval, assuming a t-distribution.

The parameters of the pulse sequence were set to $\Delta = 50$ ms and $\tau = 0.2$ ms. g was varied between 2.3 G cm^{-1} and 43.1 G cm^{-1} in eight steps, with their squares being equidistant and

32 scans per step for all spectra. The value for δ was chosen individually for each solute-solvent pair and temperature to ensure the decrease in I during the experiment was at least 80 % of the initial value. The δ were within a range of 300 to 5000 µs.

To obtain the diffusion coefficient of the solute i at infinite dilution in the solvent j, three mixtures for each system were prepared gravimetrically with low solute concentrations (0.005, 0.01, and 0.025 mol mol^{-1}). When a single solute produced multiple peaks in the NMR spectrum, a D_i could be obtained for each of those peaks, and their mean value was used in the following. The resulting concentration-dependent D_i were extrapolated linearly to find $D^{\infty}_{ij,\mathrm{exp}}$.[3] The uncertainty $\sigma_{ij,\mathrm{exp}}$ of this extrapolated $D^{\infty}_{ij,\mathrm{exp}}$ has two contributions that are summed up and reported in the form of a 95 % confidence interval assuming a t-distribution: the uncertainty σ_i of the measured D_i propagated through the linear extrapolation and the uncertainty of this linear extrapolation itself.

Results and Discussion

Comparison of Query Strategies on Synthetic Data

In Figure 4, the different query strategies are compared by showing results of predictions from the data-driven MCM as a function of the learning steps. The basis is the synthetic data set. Specifically, the rMAE of the predictions on the test set is plotted as a function of the share of the available entries of the matrix, which is 15 % at the beginning of the AL (as 15 % of the data were withheld and formed the test set for which the rMAE was calculated) and 85 % at the end, which is the maximum.

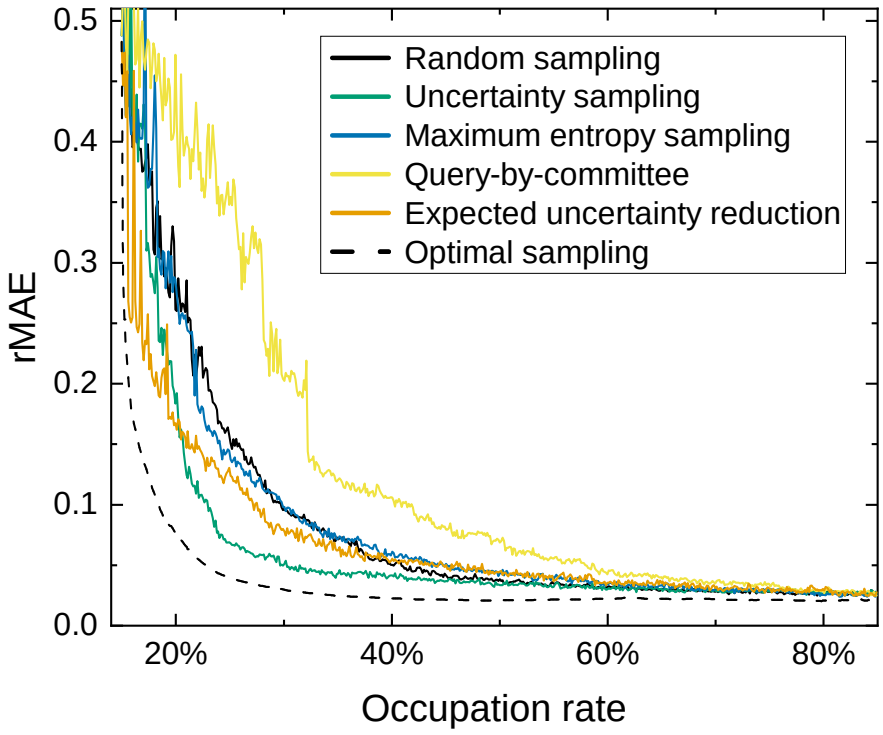


Figure 4: rMAE of the data-driven MCM for reconstructing the synthetic data set of D_{ij}^{∞} at 298 K from the SEGWE model as a function of the share of the available entries of the studied matrix. The scores were obtained by averaging the rMAE on the test sets of six random data splits.

Figure 4 shows that the prediction error decreases as the percentage of observed entries increases for all query strategies investigated, including random sampling. As expected, increasing the amount of available training data improves the model performance.

Applying the optimal sampling (which is not feasible in practical applications), the rMAE drops quickly as more entries become available. The limiting value of the rMAE of about 0.05 is already reached when the occupation rate is only 30 %; beyond this, additional data yield hardly any improvement. This is different for the five studied query strategies, even though all of them converge in the end to a value close to the limiting value of the rMAE. The best performance is found for the uncertainty sampling, for which, at an occupation rate of only about 30 %, the rMAE is already close to that limiting value. The query by committee

strategy shows the worst performance: an occupation rate of about 70 % is needed to come close to the limiting value. The results of the other three query strategies lie in between. Only in the first phase of the AL, for very low occupation rates and still high values of the rMAE, the computationally demanding expected uncertainty strategy outperforms the uncertainty sampling. As a result, we consider uncertainty sampling the best choice for the case studied, which is close to the AL task we want to solve. It was, therefore, used in the experimental AL study.

Active Learning for Planning Experiments

Based on the results presented in the synthetic AL study, uncertainty sampling was chosen as the query strategy for the experimental study of the present work. This strategy was combined with MCM-Boosting, which showed the best performance in our previous study,[6] for planning the D_{ij}^{∞} measurements. In the experimental study, beginning with the initial training data shown in Figure 1, the AL procedure described in the previous section was iteratively applied in 19 steps, whereby in each step, a new data point for D_{ij}^{∞} was measured.

Table 1 lists the measured $D_{ij,\mathrm{exp}}^{\infty}$ data for the selected systems and their measurement uncertainties. The complete list of measured self-diffusion coefficients at the studied finite concentrations is reported in the Supporting Information. Not only data for 298 K are reported, but also data for 313 K and 333 K, which were measured to get a more comprehensive database for the previously unstudied systems. This also contributes to future extensions of the MCMs to predict temperature-dependent diffusion coefficients.

In the experimental implementation of the AL approach, it was not always possible to strictly follow the suggestions made by uncertainty sampling: In step 17, uncertainty sampling proposed the binary mixture glycerol in hexadecane, which could not be dissolved in sufficiently high concentration at 298 K to observe a peak in the NMR spectrum. For this reason, 2-methyl-2,4-pentanediol in hexadecane, the second most uncertain prediction, was chosen instead.

Table 1: Liquid-phase diffusion coefficients at infinite dilution $D_{ij,\mathrm{exp}}^{\infty}$ measured by PFG NMR spectroscopy in this work, including experimental uncertainty $\sigma_{ij,\mathrm{exp}}$. The systems are sorted according to the order in which they were selected by the AL strategy, measured, and subsequently included in the MCM training.

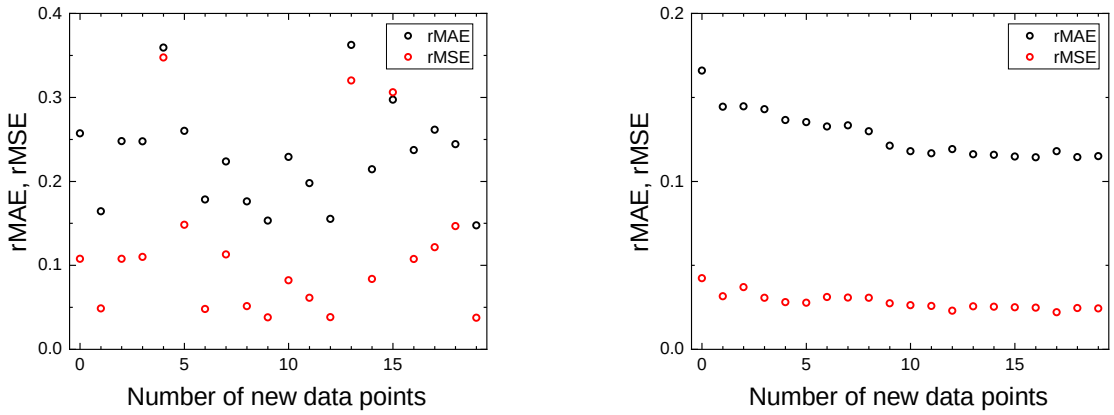
			$D_{ij,\mathrm{exp}}^{\infty}$ / $10^{-9}\mathrm{m}^2\mathrm{s}^{-1}$		
No.	Solute i	Solvent j	298 K	313 K	333 K
1	Ethanol	1,2-Propanediol	0.039 ± 0.004	0.088 ± 0.003	0.206 ± 0.006
2	Chlorobenzene	Hexadecane	0.673 ± 0.017	1.090 ± 0.041	1.734 ± 0.220
3	Ethylbenzene	Hexadecane	0.713 ± 0.016	1.035 ± 0.038	1.529 ± 0.142
4	Acetonitrile	Hexadecane	1.454 ± 0.069	2.219 ± 0.138	2.898 ± 0.142
5	Chlorotoluene	Hexadecane	0.757 ± 0.007	1.055 ± 0.011	1.499 ± 0.023
6	Iodomethane	1,2-Propanediol	0.079 ± 0.001	0.195 ± 0.005	0.408 ± 0.001
7	Ethyl acetate	Hexadecane	0.955 ± 0.024	1.240 ± 0.106	1.958 ± 0.091
8	Ethyl acetate	Chloroform	1.996 ± 0.011	2.460 ± 0.207	5.541 ± 0.255
9	Butyric acid	Hexadecane	0.541 ± 0.029	0.794 ± 0.122	1.228 ± 0.037
10	Hexafluorobenzene	Ethyl acetate	2.594 ± 0.018	3.220 ± 0.187	3.971 ± 0.058
11	Benzaldehyde	Hexadecane	0.807 ± 0.030	1.103 ± 0.076	1.600 ± 0.028
12	Di-*tert*-butyl sulfide	Hexadecane	0.456 ± 0.037	0.520 ± 0.040	0.708 ± 0.030
13	Benzyl alcohol	Butyl acetate	1.653 ± 0.074	1.979 ± 0.051	2.714 ± 0.113
14	Di-*tert*-butyl sulfide	Hexafluorobenzene	1.219 ± 0.008	1.585 ± 0.002	2.254 ± 0.107
15	Hexafluorbenzene	Acetone	3.463 ± 0.067	4.336 ± 0.065	5.180 ± 0.083
16	Acetophenone	Hexadecane	0.634 ± 0.020	1.062 ± 0.026	1.507 ± 0.046
17	2-Methyl-2,4-pentanediol	Hexadecane	0.317 ± 0.040	0.633 ± 0.065	1.141 ± 0.034
18	1-Chlorobutane	Hexadecane	0.860 ± 0.070	1.379 ± 0.126	1.797 ± 0.039
19	Propionic acid	Hexadecane	0.670 ± 0.033	0.700 ± 0.084	1.425 ± 0.016

As expected, $D_{ij,\mathrm{exp}}^{\infty}$ increases with increasing temperature for all studied systems. The experimental uncertainty varies greatly between $< 1\%$ and close to 10%, where the high values are mainly caused by the limited sensitivity of the NMR technique at the low solute concentrations studied. This effect is amplified at increasing temperature, as the signal-to-noise ratio of the NMR measurement decreases with increasing temperature.[43]

As seen in Table 1, many measured systems contain the solvent hexadecane. A possible explanation is that we have used MCM-Boosting to plan the experiments, which directly depends on the SEGWE predictions by operating on its residuals. However, the SEGWE model assumes spherical molecules and, therefore, only poorly approximates long-chained alkanes like hexadecane, the longest alkane chain in our data set. Hence, the SEGWE residuals for systems with hexadecane are unusually large compared to those of the other systems and, hence, are preferred by uncertainty sampling with MCM-Boosting. Uncertainty

sampling, in general, is known to suggest outliers particularly often.[31,44] Using a different ML model, such as MCM-Whisky, might have avoided this issue and would be a preferable choice in future studies.

In Figure 5a (left), the prediction error for the test set is shown as a function of the increasing size of the training set. The new data points were selected using uncertainty sampling based on MCM-Boosting, as described above.



(a) MCM-Boosting (b) MCM-Whisky

Figure 5: Performance of MCM-Boosting (a) and MCM-Whisky (b) for predicting D_{ij}^{∞} at 298 K in terms of relative mean absolute error (rMAE) and relative mean squared error (rMSE) as a function of the additional training data points selected using uncertainty sampling and measured in this work. Errors in (a) and (b) are plotted over different vertical scales.

At first glance, the results in Figure 5a (left) are disappointing. There is no clear trend of the influence of the additional training data on the prediction error of MCM-Boosting; quite often, adding a new training data point even leads to a significant increase in rMAE and rMSE. Although it was found to be the best-performing model for predicting D_{ij}^{∞} in our prior work,[6] the results indicate that MCM-Boosting lacks robustness with regards to its training data set and is thus not a good model choice when combined with AL. This does not mean that a Boosting model should never used for AL, but simply that for the improvement of D_{ij}^{∞} prediction by AL our MCM-Boosting model was not well-suited. Boosting models have been successfully improved with an AL approach by other authors, but not within the framework

of MCM-Boosting.[45,46] The effect of the additional training data on the performance of the data-driven MCM behaves similarly and is shown in Figure S2 in the Supporting Information.

In contrast, the prediction accuracy of MCM-Whisky significantly increases with the additional training data points, as shown in Figure 5b (right), although the measured data points were selected based on MCM-Boosting as described above. Specifically, the prediction rMAE of MCM-Whisky was reduced from 0.16 to 0.11, and the rMSE was almost halved from 0.042 to 0.024, which was achieved by increasing the matrix occupation rate by only 1.8 %.

Since hexadecane dominated the 19 AL acquisitions, the solvent-column corresponding to that species went from very sparse to suddenly over-represented. Thus, we can assume that the latent solvent feature for this column was weakly constrained at first, and each new hexadecane data point sharply updates its posterior which propagates through every D_{ij}^{∞} entry that involves that solvent; the global rMAE and rMSE can therefore oscillate whenever a single point lands far from the previous posterior mean.[44] MCM-Whisky, whose parameters were pretrained on a more uniformly covered data set, is far less sensitive to this single-column drift and thus stays numerically stable despite the same sampling bias.

Regarding the scalability of the AL approach, it should be noted that as the size of the existing training data set increases, the number of new data points required for substantial improvement will generally increase as well.[47] This can be observed in Figures 4 and 5b, and by comparison with previous works on AL with MCMs.[33,39] However, the extent to which training set size affects the informativeness of a new data point also depends on the model and data structure.[31] We can thus conclude that the viability of an AL approach depends not only on the size of the existing data set, but also on its structure, the model to be improved, as well as the cost of obtaining new data points.

In retrospect, using MCM-Whisky instead of MCM-Boosting together with uncertainty sampling from the start would have been preferable. However, knowing this in advance was impossible, especially since MCM-Boosting showed better performance in our previ-

ous work.[6] Additionally, we can see from the comparison of Figures 5a and 5b that in our case we obtain lower prediction errors for MCM-Whisky than for MCM-Boosting, seemingly contradicting the better performance of MCM-Boosting in the previous study.[6] This discrepancy, however, can be explained by the use of a fixed test set in this study, as opposed to the leave-one-out analysis used in the previous work,[6] and the use of a slightly different data base. The optimal query strategy in AL is always a joint function of the model and the structure of the observed and unobserved data. Thus, a query strategy tuned on the data-driven setting with synthetic data may become sub-optimal in the experimental setting. Nevertheless, the synthetic benchmark allows to explore the application of various different AL approached to the MCM setting, and provided a baseline for the experimental scenario. All in all, these results indicate that MCM-Whisky is a significantly more robust model regarding its training data than MCM-Boosting and that significant improvements in diffusion coefficient prediction with only a few measurements are possible if they are targeted using AL strategies.

In the Supporting Information, we report the predicted $D^{\infty}_{ij,\mathrm{pred}}$ obtained with MCM-Whisky trained on all literature data and the new data measured in this work.

Conclusions

We have investigated the application of active learning (AL) strategies to improve ML models for predicting physico-chemical properties of liquid mixtures, taking the diffusion coefficients at infinite dilution in binary mixtures D^{∞}_{ij} at 298 K as an example. A special feature of this study is that AL was combined with actual laboratory experiments by which an existing database D^{∞}_{ij} was extended. Hence, this work can be considered a field study on applying AL in connection with experiments in ML modeling.

The central issue in AL is the query strategy by which, in our case, the binary systems that were studied experimentally were selected. For selecting a suitable query strategy, we

conducted a preliminary study with a synthetic data set created with a semi-empirical model for D_{ij}^{∞}, so that it was closely related to our application. Among the studied query strategies, uncertainty sampling, which suggests the next system to be measured based on the largest model uncertainty, was found to give the best results and was therefore employed in the main study.

Three matrix completion methods (MCMs) were used for predicting D_{ij}^{∞}. A data-driven MCM was used in the preliminary study, while in the main study, two hybrid MCMs were used, in which the MCM is coupled with the semi-empirical SEGWE model for predicting D_{ij}^{∞}. In the Boosting method, the MCM learns to predict the errors of the SEGWE model, which can then be corrected. In the Whisky method, the MCM is pre-trained using the SEGWE predictions. Only one of these methods could be used to implement the AL strategy to guide the experimental program. Unfortunately, we selected MCM-Boosting for this, as it had yielded the best results in previous work, and not MCM-Whisky, which would have been the better choice, as we will explain below.

Hence, in the main study, the available data on D_{ij}^{∞} were extended stepwise by new experimental data for systems suggested by uncertainty sampling in connection with MCM-Boosting. A total of 19 steps were carried out, i.e., D_{ij}^{∞} was measured for 19 systems for which previously no data were available. The measurements were not only carried out for 298 K, as it would have been sufficient for the present work, but also for 313 K and 333 K, which extends the available database for the future development of ML methods for predicting temperature-dependent D_{ij}^{∞}.

The results of the AL were assessed using results for a test data set that was withheld from the initial data on D_{ij}^{∞} and not used in the training of the MCMs. The prediction of the test data was monitored in each step of the AL. To our surprise, adding new data yielded no significant improvement for MCM-Boosting (which was used to guide the choice of the newly studied systems). In many cases, the errors even increased upon adding the new information. This shows that MCM-Boosting, which basically corrects predictions from a physical model,

is less robust than we had assumed based on previous results. It also indicates that boosting methods may be ill-suited for guiding AL processes. This is plausible as boosting focuses on systems with which the physical model has the most problems, which may be difficult to heal empirically. In contrast, we found that the results from MCM-Whisky continuously improved upon adding new data (even though the new systems were selected using MCM-Boosting). Significant improvements of the predictions of D_{ij}^{∞} by MCM-Whisky were achieved even though only comparatively few data points were added.

Overall, the results from the present study demonstrate that AL is a promising technique for improving ML methods for predicting physico-chemical properties by the targeted planning of measurements, which enables efficient use of the available resources. The way the experimental program was designed in the present work can be seen as an example of a new paradigm in selecting systems for which physico-chemical data are measured, which was previously generally driven by specific applications. In contrast, here, it was driven by the desire to improve an ML model that applies to many different systems.

Acknowledgement

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 Resarch Training Group GRK2908 "Valuable Wastewater (WERA)" (grant number 503479768), the Priority Program SPP2363 "Molecular Machine Learning" (grant number 497201843) and 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).

Supporting Information Available

Supporting Information is available:

- Compound identifiers; diffusion coefficient values; maximum entropy sampling; test set; list of chemicals; AL with data-driven MCM. (PDF)
- Python code for testing the different query strategies. (PY)
- Experimental self-diffusion coefficients. (CSV)
- Diffusion coefficients $D_{ij,\mathrm{pred}}$ predicted for $T = 298\,\mathrm{K}$ by MCM-Whisky trained on the final dataset. (CSV)

References

(1) Poling, B. E.; Prausnitz, J. M.; O'Connell, J. *The Properties of Gases and Liquids*, 5th ed.; McGraw-Hill Professional: New York, NY, 2000.

(2) Evans, R.; Poggetto, G. D.; Nilsson, M.; Morris, G. A. Improving the interpretation of small molecule diffusion coefficients. *Analytical Chemistry* **2018**, *90*, 3987–3994.

(3) Bellaire, D.; Großmann, O.; Münnemann, K.; Hasse, H. 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. *The Journal of Chemical Thermodynamics* **2022**, *166*, 106691.

(4) Schmitt, S.; Hasse, H.; Stephan, S. Entropy scaling framework for transport properties using molecular-based equations of state. *Journal of Molecular Liquids* **2024**, *395*, 123811.

(5) Schmitt, S.; Hasse, H.; Stephan, S. Entropy scaling for diffusion coefficients in fluid mixtures. *Nature Communications* **2025**, *16*, 2611.

(6) 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.

(7) Cussler, E. L. *Diffusion: Mass Transfer in Fluid Systems*; Cambridge University Press, 2009.

(8) Tham, M.; Bhatia, K.; Gubbins, K. Steady-state method for studying diffusion of gases in liquids. *Chemical Engineering Science* **1967**, *22*, 309–311.

(9) Taylor, G. I. Dispersion of soluble matter in solvent flowing slowly through a tube. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **1953**, *219*, 186–203.

(10) Mountain, R. D.; Deutch, J. M. Light scattering from binary solutions. *The Journal of Chemical Physics* **1969**, *50*, 1103–1108.

(11) Bellaire, D.; Münnemann, K.; Hasse, H. Mutual diffusion coefficients from NMR imaging. *Chemical Engineering Science* **2022**, *255*, 117655.

(12) Taylor, R.; Krishna, R. *Multicomponent Mass Transfer*; Wiley Series in Chemical Engineering; John Wiley & Sons: Nashville, TN, 1993.

(13) Bellaire, D.; Kiepfer, H.; Münnemann, K.; Hasse, H. PFG-NMR and MD simulation study of self-diffusion coefficients of binary and ternary mixtures containing cyclohexane, ethanol, acetone, and toluene. *Journal of Chemical and Engineering Data* **2020**, *65*, 793–803.

(14) Specht, T.; Münnemann, K.; Hasse, H.; Jirasek, F. Rational method for defining and quantifying pseudo-components based on NMR spectroscopy. *Physical Chemistry Chemical Physics* **2023**, *25*, 10288–10300.

(15) Mross, S.; Schmitt, S.; Stephan, S.; Münnemann, K.; Hasse, H. Diffusion coefficients in mixtures of poly(oxymethylene) dimethyl ethers with alkanes. *Industrial and Engineering Chemistry Research* **2024**, *63*, 1662–1669.

(16) Vignes, A. Diffusion in binary solutions. Variation of diffusion coefficient with composition. *Industrial and Engineering Chemistry Fundamentals* **1966**, *5*, 189–199.

(17) 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.

(18) Jirasek, F.; Hasse, H. Perspective: Machine learning of thermophysical properties. *Fluid Phase Equilibria* **2021**, *549*, 113206.

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

(20) Damay, J.; Jirasek, F.; Kloft, M.; Bortz, M.; Hasse, H. Predicting activity coefficients at infinite dilution for varying temperatures by matrix completion. *Industrial and Engineering Chemistry Research* **2021**, *60*, 14564–14578.

(21) Gond, D.; Sohns, J.-T.; Leitte, H.; Hasse, H.; Jirasek, F. Hierarchical matrix completion for the prediction of properties of binary mixtures. *Computers & Chemical Engineering* **2025**, *199*, 109122.

(22) Hayer, N.; Specht, T.; Arweiler, J.; Gond, D.; Hasse, H.; Jirasek, F. Prediction of activity coefficients by similarity-based imputation using quantum-chemical descriptors. *Physical Chemistry Chemical Physics* **2025**, *27*, 4307–4315.

(23) Zenn, J.; Gond, D.; Jirasek, F.; Bamler, R. Balancing molecular information and empirical data in the prediction of physico-chemical properties. *Digital Discovery* **2025**, *4*, 683–693.

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

(25) Hayer, N.; Hasse, H.; Jirasek, F. Prediction of temperature-dependent Henry's law constants by matrix completion. *The Journal of Physical Chemistry B* **2024**, *129*, 409–416.

(26) 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.

(27) Hayer, N.; Wendel, T.; Mandt, S.; Hasse, H.; Jirasek, F. Advancing thermodynamic group-contribution methods by machine learning: UNIFAC 2.0. *Chemical Engineering Journal* **2025**, *504*, 158667.

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

(29) Jirasek, F.; Hasse, H. Combining machine learning with physical knowledge in thermodynamic modeling of fluid mixtures. *Annual Review of Chemical and Biomolecular Engineering* **2023**, *14*, 31–51.

(30) Jirasek, F.; Hayer, N.; Abbas, R.; Schmid, B.; Hasse, H. Prediction of parameters of group contribution models of mixtures by matrix completion. *Physical Chemistry Chemical Physics* **2023**, *25*, 1054–1062.

(31) Settles, B. *Active Learning Literature Survey*; Computer Sciences Technical Report 1648, 2009.

(32) Tran, K.; Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nature Catalysis* **2018**, *1*, 696–703.

(33) Donval, G.; Hand, C.; Hook, J.; Dupont, E.; Sabate Landman, M.; Freitag, M.; Lennox, M.; Düren, T. Ver. 1. ChemRxiv, May 10, 2021, DOI: 10.26434/chemrxiv.14555706.v1 (accessed 2025-07-09).

(34) Dortmund Data Bank. DDBST - Dortmund Data Bank Software And Separation Technology GmbH, Dortmund Data Bank, 2024. https://www.ddbst.com (accessed 2024-05-29).

(35) DIPPR Data Compilation of Pure Chemical Properties. Design Institute for Physical Properties (DIPPR), 2024; American Institute of Chemical Engineers (AIChE).

(36) Kucukelbir, A.; Tran, D.; Ranganath, R.; Gelman, A.; Blei, D. M. Automatic differentiation variational inference. *Journal of Machine Learning Research* **2017**, *18*, 1–45.

(37) Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association* **2017**, *112*, 859–877.

(38) Stan Development Team, Stan Modeling Language Users Guide and Reference Manual, Version 2.35. `https://mc-stan.org`.

(39) Chakraborty, S.; Zhou, J.; Balasubramanian, V.; Panchanathan, S.; Davidson, I.; Ye, J. Active matrix completion. 2013 IEEE 13th International Conference on Data Mining. 2013; pp 81–90.

(40) Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **1948**, *27*, 379–423.

(41) Wu, D.; Chen, A.; Johnson, C. Flow imaging by means of 1D pulsed-field-gradient NMR with application to electroosmotic flow. *Journal of Magnetic Resonance, Series A* **1995**, *115*, 123–126.

(42) Stejskal, E. O.; Tanner, J. E. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. *The Journal of Chemical Physics* **1965**, *42*, 288–292.

(43) Hoult, D.; Richards, R. The signal-to-noise ratio of the nuclear magnetic resonance experiment. *Journal of Magnetic Resonance (1969)* **1976**, *24*, 71–85.

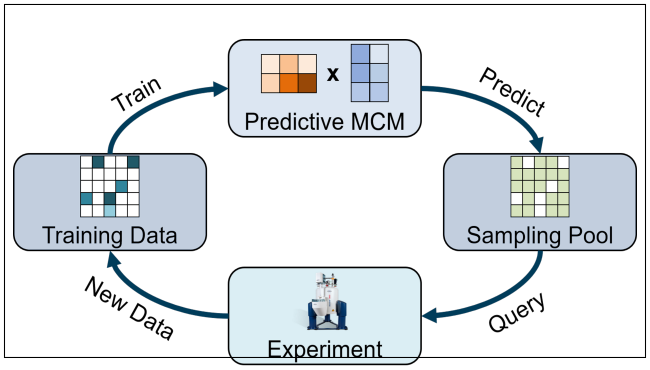
(44) Silva, J.; Carin, L. Active learning for online bayesian matrix factorization. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012; p 325–333.

(45) Sui, Q.; Ghosh, S. K. Active learning for stacking and AdaBoost-related models. *Stats* **2024**, *7*, 110–137.

(46) M Shankaranarayana, S. Model uncertainty based active learning on tabular data using boosted trees. The Third International Conference on Artificial Intelligence and Machine Learning Systems. 2023; p 1–9.

(47) Castro, R. M.; Nowak, R. D. *Learning Theory*; Springer Berlin Heidelberg, 2007; p 5–19.

TOC Graphic


Train
x
Predictive MCM
Predict
Training Data
Sampling Pool
New Data
Experiment
Query