Prediction of Diffusion Coefficients in Mixtures with Tensor Completion

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

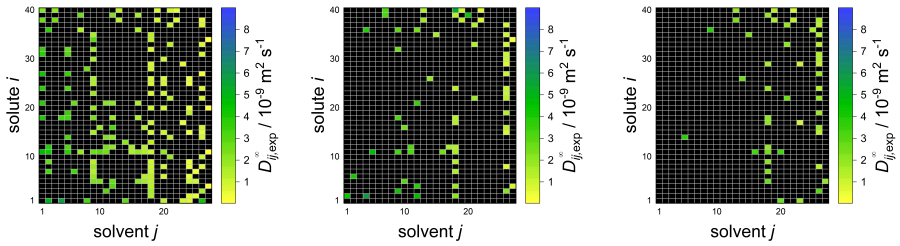
Laboratory of Engineering Thermodynamics (LTD), RPTU Kaiserslautern, Germany

Abstract. We propose a hybrid tensor completion method as a novel approach for predicting temperature-dependent diffusion coefficients, improving the scope and accuracy of existing models. Furthermore, we study active learning (AL) strategies for efficiently improving the predictions by purposefully planning and measuring diffusion coefficients with NMR spectroscopy. The results show that while AL enables significant improvements in a synthetic example, only a minor effect is observed in the experimental scenario.

1 Introduction

Information on the diffusion coefficients is essential in chemical engineering for modeling transport phenomena and simulating thermal separation processes. Unfortunately, experimental data are scarce, especially for mixtures; thus, models for predicting diffusion coefficients in mixtures are paramount in practice. Several prediction methods for this purpose, mainly semi-empirical, have been proposed in the literature [1, 2], but they all have significant limitations, including small scope and poor accuracy. For predicting diffusion coefficients in binary mixtures, matrix completion methods (MCMs) from machine learning (ML) are a fascinating option [2–4], exploiting that the properties of binary mixtures can be represented in matrices where the rows and columns denote the mixture components with many unobserved entries in most cases, cf. Fig. 1; the MCM is then trained to predict missing entries, i.e., properties of unstudied mixtures [2].

However, while MCMs have been shown to outperform all current benchmark models for predicting diffusion coefficients [2], they suffer from two limitations: they are restricted to a single temperature, i.e., they cannot describe the temperature dependence of diffusion coefficients, and they heavily rely on reliable experimental data for training, which is often not available. In this work, we address both limitations by developing novel tensor completion methods (TCMs) capturing the temperature dependence of diffusion coefficients and studying active learning (AL) strategies [5] for purposefully planning diffusion measurements by pulsed-field gradient (PFG) nuclear magnetic resonance (NMR) spectroscopy [6, 7].



(a) 298 K (b) 313 K (c) 333 K

Fig. 1: Available experimental data for binary liquid-phase diffusion coefficients D^{∞}_{ijk} at infinite dilution and three temperatures [2, 8].

2 Methods

2.1 Tensor Completion

We propose a TCM for predicting temperature-dependent diffusion coefficients of solutes i at infinite dilution D^{∞}_{ijk} in solvents j at temperatures k. For this purpose, we arrange the available experimental data from the Dortmund Data Bank [8] for 298, 313, and 333 K (cf. Fig. 1) in a three-dimensional tensor. This tensor contains D^{∞}_{ijk} for 27 solvents and 40 solutes, representing the first two dimensions of the tensor, while the third dimension represents the temperature. We use Tucker decomposition to factorize the tensor:

$$\ln D^{\infty}_{ijk} = \sum_{\alpha=1}^{r_i} \sum_{\beta=1}^{r_j} \sum_{\gamma=1}^{r_k} A(i,\alpha) \cdot B(j,\beta) \cdot C(k,\gamma) \cdot \kappa(\alpha,\beta,\gamma) \tag{1}$$

A, B, and C are feature matrices of solute, solvent, and temperature, respectively, learned during training, while r_i, r_j, and r_k are their respective feature dimensions. κ is the core tensor used in Tucker decomposition, which introduces additional flexibility as it allows different feature dimensions. Based on the results of preliminary tests, the hyperparameters were set to $r_i = r_j = r_k = 2$.

We use a probabilistic approach based on a Cauchy likelihood to train the model, which was implemented using the probabilistic programming language Stan [9]. Besides experimental data, we also use synthetic data from the SEGWE model [1] for training the TCM, which we incorporate as prior knowledge. The proposed model has a computation time of a few seconds and is thus very resource-friendly compared to contemporary ML models.

2.2 Active Learning

The experimental database for training the TCM was substantially extended by measuring diffusion coefficients with PFG NMR spectroscopy. AL strategies

were employed to select the most promising experiments and use the experimental resources as efficiently as possible [5]. In our AL workflow, we sequentially train the model on the training data, use a query strategy to select a new data point, measure it via NMR spectroscopy, and add the new data point to the training data. We thereby iteratively extend the training database and use it for retraining the model.

The query strategy at the core of the AL approach is essential as it defines on which basis the subsequent measurement is planned. In this work, we first systematically investigated established query strategies on a synthetic data set consisting of predictions from the SEGWE model at 298 K [1]. The tested strategies were random sampling, uncertainty sampling, maximum entropy sampling, and query-by-committee [5]. For this purpose, 15 % of the synthetic data were defined as the initial training set, 70 % were considered as the pool from which the AL approach can sample, and the remaining 15 % were used as the test set.

3 Results and Discussion

3.1 Testing Query Strategies on Synthetic Data

Fig. 2 shows the results for applying the AL approach based on different query strategies to the synthetic data set. Specifically, it shows the relative mean absolute error (rMAE) of the MCM on the test set after training on training sets of different sizes, denoted by the share of observed entries in the matrix.

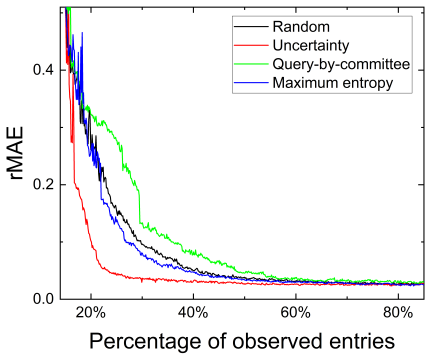


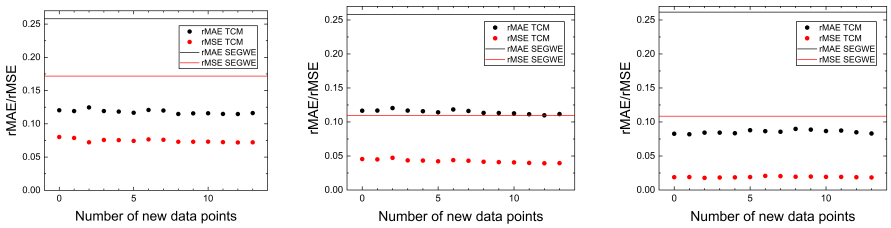
Fig. 2: rMAE of an MCM for reconstructing a synthetic data set for D_{ijk}^{∞} at 298 K over the growing training set using different query strategies.

In all cases, the prediction error decreases with increasing percentage of observed entries. Uncertainty sampling proved the most efficient of the studied query strategies, while the other query strategies did not perform significantly

better than random sampling. For this reason, uncertainty sampling was subsequently used to plan actual experiments.

3.2 Diffusion Coefficient Prediction with TCM

Fig. 3 shows the results of the TCM for predicting diffusion coefficients and applying AL based on using uncertainty sampling. Specifically, it shows the rMAE and relative mean squared error (rMSE) of the TCM for each temperature calculated using leave-one-out analysis after training on training sets of growing size, denoted by the number of new data points at each temperature.



(a) 298 K (b) 313 K (c) 333 K

Fig. 3: rMAE and rMSE of the temperature-dependent prediction of D_{ijk}^{∞} by TCM as a function of the additional training data points chosen by uncertainty sampling and measured in this work. Scores calculated by leave-one-out analysis.

The TCM (symbols in Fig. 3) shows significantly higher prediction accuracy than the established benchmark model SEGWE [1] (lines in Fig. 3) in both error scores at all temperatures. However, the additional data measured in this work did not significantly reduce the TCM prediction errors. Since only approx. 18% of the binary systems are studied, an improved accuracy could be expected, cf. Fig. 2. Hence, a direct transfer from synthetic to experimental data may not be feasible. One possible reason for the results is that the size of the experimental data set is too small to show the model's generalization ability and power of AL.

4 Conclusions

In the present work, we introduce a novel approach for predicting temperature dependent diffusion coefficients, combining the established semi-empirical SEGWE model with a TCM to a hybrid model. The hybrid TCM gives significantly more accurate predictions than the established benchmark. Regarding AL, we found in a study using synthetic data that the best performance gain could be realized using uncertainty sampling as the query strategy. Subsequently, we used uncertainty sampling to purposefully extend the existing experimental for diffusion coefficients, which, however, only marginally improved the performance of the TCM trained on that data further.

Acknowledgments. We gratefully acknowledge financial support by Carl Zeiss Foundation in the project "Process Engineering 4.0", as well as by Deutsche Forschungsgemeinschaft in the Priority Program 2363, and in the Emmy Noether Project of FJ.

References

1. R. Evans, G. Dal Poggetto, M. Nilsson, G. A. Morris, Anal. Chem. 90 (2018), 3987–3994.
2. O. Großmann, D. Bellaire, N. Hayer, F. Jirasek, H. Hasse, Digital Discovery 6 (2022) 886-897.
3. F. Jirasek, R. A. S. Alves, J. Damay, R. A. Vandermeulen, R. Bamler, M. Bortz, S. Mandt, M. Kloft, H. Hasse, J. Phys. Chem. Lett. 11 (2020), 981–985.
4. N. Hayer, F. Jirasek, H. Hasse, AIChE Journal 68 (2022), DOI 10.1002/aic.17753.
5. B. Settles, Active Learning, 1st ed., Springer Cham, Basel, Switzerland (2012).
6. D. Bellaire, H. Kiepfer, K. Münnemann, H. Hasse, J. Chem. Eng. Data 65 (2020) 793-803.
7. D. Bellaire, O. Großmann, K. Münnemann, H. Hasse, J. Chem. Thermodyn. 166 (2022) 106691.
8. Dortmund Data Bank, 2024, www.ddbst.com.
9. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., J. Stat. Softw. 76 (2017), 1–32.