Making thermodynamic models predictive by machine learning: matrix completion of pair interactions*

Fabian Jirasek[1], Sophie Fellenz[1], Robert Bamler[2], Michael Bortz[3], Marius Kloft[1], Stephan Mandt[4], and Hans Hasse[1]

[1] RPTU Kaiserslautern
[2] University of Tübingen
[3] Fraunhofer Institute for Industrial Mathematics, Kaiserslautern
[4] University of California, Irvine

Abstract. We propose a generic hybrid approach that combines classical thermodynamic models with matrix completion methods (MCMs) from machine learning. As an example, we embed an MCM into a widely-used physical model to predict pair-interaction energies in liquid mixtures. Using a Bayesian machine-learning framework we predict activity coefficients for any binary mixture of these components in a thermodynamically consistent way, thereby surpassing the accuracy of the established benchmark model.

1 Introduction

Information on thermodynamic properties of mixtures is of crucial importance in chemical engineering and chemistry. However, providing this information is hampered by a combinatorial problem: there are way too many components, let alone possible mixtures of components, thermodynamic properties, and state points to study all relevant combinations in experiments. Consequently, experimental data on thermodynamic properties are available only for a small fraction of the relevant mixtures. Therefore, methods for the prediction of thermodynamic properties of mixtures are essential in practice. While physical methods were the gold standard in the last decades, data-driven methods from machine learning (ML) [5–7, 10], and, in particular, hybrid models combining both worlds [8, 9, 11, 12] are recently offering more and more promising alternatives. Of outstanding significance are activity coefficients, which describe the deviation from ideal mixtures and are the key properties for modeling, e.g., phase equilibria.

Our contributions are as follows:

- We propose a hybrid method for predicting activity coefficients of binary mixtures by combining ML with a physical model.
- We accurately predict activity coefficients in unseen mixtures.
- We compare to and outperform the current state-of-the-art physical prediction model UNIFAC [3, 4].

* This paper is a shortened version of a previously published journal paper by the same authors [8].

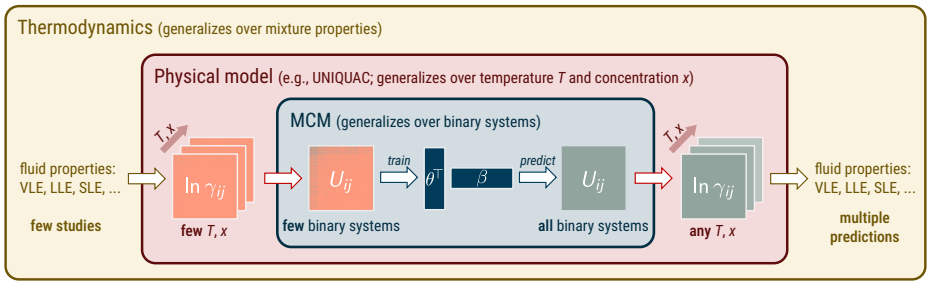

Fig. 1. Illustration of embedding an MCM into a physical model of mixtures (here: the lattice model UNIQUAC). Blue part: application of an MCM to pair-interaction energies U_{ij}. Red part: the physical model relates U_{ij} to temperature- and concentration-dependent activity coefficients γ_{ij}. Yellow part: γ_{ij} are directly related to observable mixture properties (e.g., vapor–liquid equilibria, liquid–liquid equilibria, and solid–liquid equilibria by thermodynamic laws.

2 Method: MCM-UNIQUAC

The idea behind our approach is shown in Fig. 1. Our goal is to predict activity coefficients γ_{ij} of pure solutes i in pure solvents j. We have a physical model [1, 14] that is able to predict γ_{ij} dependent on temperature and concentration, if the pair-interaction energies of the pure components U_{ii} and U_{jj} as well as of the mixture U_{ij} are known. However, for most binary mixtures, U is unknown. Thus, we use a matrix completion method (MCM) to predict the pair-interaction energies for unseen binary mixtures. For describing activity coefficients γ_{ij} of pure solutes i in pure solvents j, the physical model can be written as:

$$\ln \gamma_{ij}(T, x_i) = f_{\text{UNIQUAC}}(T, x_i, P_i, P_j, U_{ii}, U_{jj}, U_{ij}), \tag{1}$$

where the function f_{UNIQUAC} contains the UNIQUAC equation (see full paper [8]). Here, T is the temperature, x_i is the mole fraction (concentration) of component i in the mixture, and P_i, P_j are pure-component parameters of i and j, respectively, which are typically known.

We now use the MCM to generalize this model to binary mixtures where no experimental data is available. We factorize this matrix $U \in \mathbb{R}^{M \times M}$, where M is the number of components, as follows

$$U = \theta^T \beta + \beta^T \theta, \tag{2}$$

where $\theta \in \mathbb{R}^{K \times M}$ and $\beta \in \mathbb{R}^{K \times M}$ and K is the feature vector dimension (we use $K = 3$). The right-hand side of Eq. 2 is constructed in such a way that the physical constraint of symmetry in the pair-interaction energies, $U_{ij} = U_{ji}\ \forall i, j$, is enforced, resulting in a symmetric matrix.

We use a probabilistic model in which the $\ln \gamma_{ij}$ variable is modeled with a Cauchy likelihood

$$p(\ln \gamma | T, x_i, P_i, P_j, U_{ii}, U_{jj}, U_{ij}) = Cauchy(\ln \gamma | f_{UNIQUAC}(\cdot))$$

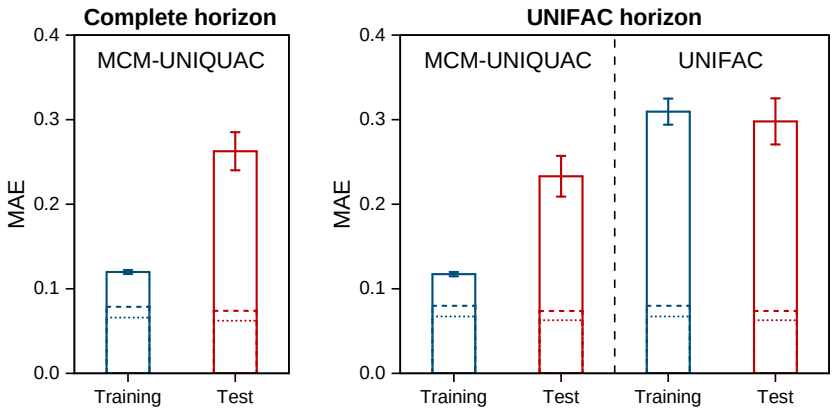

Fig. 2. Mean absolute error (MAE) of MCM-UNIQUAC on the training and test set (left) and comparison to UNIFAC [3] based only on those systems that can also be modeled with UNIFAC (right). Bars indicate the results of MCM-UNIQUAC and UNIFAC, and lines denote the baselines obtained by directly fitting UNIQUAC pair-interaction parameters (ΔU_{ij}, dotted) or pair-interaction energies (U_{ij}, dashed) to all available data points. Error bars denote standard errors of the means.

and Gaussian priors on the latent variables θ and β. We trained our model using the probabilistic programming language Stan [2] and resorted to mean-field Gaussian Variational Inference [13] for obtaining an approximate posterior distribution for the activity coefficients. The final prediction was made by averaging 1,000 samples from the posterior. [5]

MCM-UNIQUAC was trained end-to-end on a set of measured logarithmic temperature- and concentration-dependent activity coefficients in binary mixtures $\ln \gamma_{ij}$ from the Dortmund Data Bank (DDB) [15]. The considered $M = 1,146$ components result in $M(M-1)/2 = 656,085$ possible different binary systems, but experimental data are only available for 12,199 of these systems. Data for 80% of the systems were used for training, 10% of the systems were used for validation, and 10% for testing.

3 Results and discussion

The results obtained with MCM-UNIQUAC are shown in Fig. 2 (left), where the mean absolute error (MAE) is reported. In Fig. 2 (right), we compare it with the best available physical method for the prediction of activity coefficients, the group-contribution model modified UNIFAC (Dortmund) [3, 16]. In contrast to MCM-UNIQUAC, UNIFAC cannot be applied to all systems for which data are available (denoted as 'complete horizon' in Fig. 2 (left)) because multiple group-interaction parameters are missing. Hence, only those subsets of the training set

[5] More details in the full paper [8].

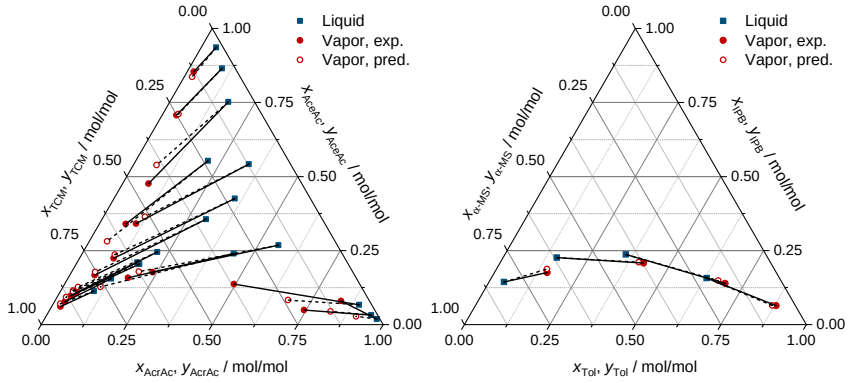

Fig. 3. Prediction of the vapor-liquid equilibrium in ternary systems at constant pressure with MCM-UNIQUAC and comparison to experimental data (exp.) from the DDB [15]. The pressure and the composition of the liquid phase were specified, the composition of the corresponding vapor phase was predicted (pred.). Left: acrylic acid (AcrAc) + acetic acid (AceAc) + tetrachloromethane (TCM) at 100 kPa. Right: toluene (Tol) + isopropylbenzene (IPB) + α-methyl styrene (α-MS) at 101 kPa.

and of the test set for which UNIFAC could be applied were used; this 'UNIFAC horizon' covers 7,578 of 9,759 systems from the training set and 961 of 1,220 systems from the test set. As baselines, the scores obtained for the different sets by directly fitting UNIQUAC parameters to all available data points are marked as lines in Fig. 2. For an in depth explanation of the lines, please refer to the full paper [8]. MCM-UNIQUAC not only allows modeling binary mixtures, but can also be applied to multi-component mixtures. As an example, isobaric VLE phase diagrams for two ternary systems are shown in Fig. 3. The constituent binary subsystems of these systems were part of neither training nor validation set. Predicted and experimental values agree very well.

4 Conclusion

In the present work, we describe a novel hybrid approach for predicting thermodynamic properties of mixtures, which combines matrix completion methods (MCMs) with physical modeling. The basic idea is to predict the pair-interaction energies, as used in basically all physical models of mixtures, between components in mixtures using MCMs. We thereby generalize the physical model to yield predictions for all binary systems. We demonstrate the predictive capacity of our model as compared to the existing physical benchmark model UNIFAC, which we can clearly outperform. Our hybridization approach is generic, it can be applied to any mixture property, and any physical model based on pair interactions can be used. In the future, among others, we plan to extend our approach to more sophisticated ML methods.

5 Acknowledgements

The authors gratefully acknowledge financial support by Carl Zeiss Foundation in the frame of the project 'Process Engineering 4.0' and by Bundesministerium für Wirtschaft und Energie (BMWi) in the frame of the project 'KEEN'. This material is in part based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0021. Stephan Mandt furthermore acknowledges support by the National Science Foundation under the NSF CAREER award 2047418 and Grants 1928718, 2003237, and 2007719, the Department of Energy under grant DE-SC0022331, as well as unrestricted gifts from Intel, Disney, and Qualcomm. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or NSF.

References

1. Abrams, D.S., Prausnitz, J.M.: Statistical thermodynamics of liquid mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems. AIChE Journal **21**, 116–128 (1975). https://doi.org/https://doi.org/10.1002/aic.690210115
2. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of Statistical Software **76**, 1–32 (2017). https://doi.org/10.18637/jss.v076.i01
3. Constantinescu, D., Gmehling, J.: Further development of modified UNIFAC (Dortmund): Revision and extension 6. Journal of Chemical & Engineering Data **61**, 2738–2748 (2016). https://doi.org/10.1021/acs.jced.6b00136
4. Fredenslund, A., Jones, R.L., Prausnitz, J.M.: Group-contribution estimation of activity coefficients in nonideal liquid mixtures. AIChE Journal **21**, 1086–1099 (1975). https://doi.org/https://doi.org/10.1002/aic.690210607
5. Grossmann, O., Bellaire, D., Hayer, N., Jirasek, F., Hasse, H.: Data base for diffusion coefficients at infinite dilution at 298 K and matrix completion methods for their prediction. Digital Discovery **1**, 886–897 (2022). https://doi.org/10.1039/D2DD00073C
6. Hayer, N., Jirasek, F., Hasse, H.: Prediction of Henry's law constants by matrix completion. AIChE Journal **68**, e17753 (2022). https://doi.org/10.1002/aic.17753
7. 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 **11**, 981–5 (2020). https://doi.org/10.1021/acs.jpclett.9b03657
8. 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 **13**, 4854–4862 (2022). https://doi.org/10.1039/D1SC07210B
9. Jirasek, F., Bamler, R., Mandt, S.: Hybridizing physical and data-driven prediction methods for physicochemical properties. Chemical Communications **56**, 12407–12410 (2020). https://doi.org/10.1039/D0CC05258B

10. Jirasek, F., Hasse, H.: Perspective: Machine learning of thermophysical properties. Fluid Phase Equilibria **549**, 113206 (2021). https://doi.org/https://doi.org/10.1016/j.fluid.2021.113206
11. Jirasek, F., Hasse, H.: Combining machine learning with physical knowledge in thermodynamic modeling of fluid mixtures. Annual Review of Chemical and Biomolecular Engineering **14**, 31–51 (2023). https://doi.org/10.1146/annurev-chembioeng-092220-025342
12. 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 **25**, 1054–1062 (2023). https://doi.org/10.1039/D2CP04478A
13. Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., Blei, D.M.: Automatic differentiation variational inference. Journal of Machine Learning Research **18**, 1–45 (2017), `http://jmlr.org/papers/v18/16-107.html`
14. Maurer, G., Prausnitz, J.: On the derivation and extension of the UNIQUAC equation. Fluid Phase Equilibria **2**, 91–99 (1978). https://doi.org/https://doi.org/10.1016/0378-3812(78)85002-X
15. Onken, U., Rarey-Nies, J., Gmehling, J.: The Dortmund Data Bank: A computerized system for retrieval, correlation, and prediction of thermodynamic properties of mixtures. International Journal of Thermophysics **10**, 739–747 (1989). https://doi.org/10.1007/BF00507993
16. Weidlich, U., Gmehling, J.: A modified UNIFAC model. 1. prediction of VLE, h^{E}, and γ^{∞}. Industrial & Engineering Chemistry Research **26**, 1372–1381 (1987). https://doi.org/10.1021/ie00067a018