Embedding Machine Learning Methods in Physical Thermodynamic Models

Nicolas Hayer[1], Stephan Mandt[2], Hans Hasse[1], and Fabian Jirasek[1]

[1] RPTU Kaiserslautern, Germany
[2] University of California, Irvine, USA

Abstract. Predicting the thermophysical properties is crucial in chemical engineering. Physical group-contribution methods (GCM) are widely used for this purpose but suffer from incomplete and inconsistent parameterizations, severely limiting their applicability and accuracy. In this work, we solve both issues by combining the most successful GCM, UNIFAC, with matrix completion methods (MCM) from machine learning, whereby the MCM is used to predict the pair-interaction parameters for the GCM. The resulting hybrid model, UNIFAC 2.0, has significantly higher prediction accuracy and scope than the original model.

1 Introduction

Knowledge of the thermodynamic properties of mixtures is crucial for chemical engineering. However, the sheer combinatorial diversity of mixtures makes it impossible to study each relevant mixture experimentally, making reliable prediction methods indispensable. Group-contribution methods (GCM) are widely used for this purpose. The best-established GCM is UNIFAC for predicting activity coefficients in liquid mixtures. Since its introduction in 1975 [3] it has been constantly revised and improved [15, 4, 12, 16, 5, 17] and is implemented in basically all process simulators, underlining its enduring relevance and success.

We use the latest published version of UNIFAC [17], labeled as UNIFAC 1.0 here, as a reference. UNIFAC 1.0 decomposes components into structural groups, and applying it to a given mixture requires pair-interaction parameters (a_{mn}) for each binary combination of the occurring main groups m and n. However, interaction parameters are missing for 56% of all pairs of groups, in some cases due to the challenging fitting process and in other cases due to the lack of experimental data for direct fitting, which severely hampers the applicability of UNIFAC 1.0 (a single missing relevant parameter prevents using the model). Unknown a_{mn} can be estimated using artificial training data from COSMO-based prediction methods or atomic interaction parameters. However, both approaches yield unreliable results and cannot match the accuracy of fitting to experimental vapor-liquid equilibrium (VLE) data [13].

In this work, we introduce a new way of predicting the interaction parameters of GCM based on machine learning. The approach is based on the idea that the pair-interaction parameters can be treated as elements of a square matrix and

a matrix completion method (MCM) [14, 9, 7, 6, 8, 10] is used to calculate the entries. Thereby, numbers for all entries are found, and the problem of missing parameters is solved. The proposed MCM, UNIFAC 2.0, enhances our previous work [10] by incorporating end-to-end training on extensive experimental data.

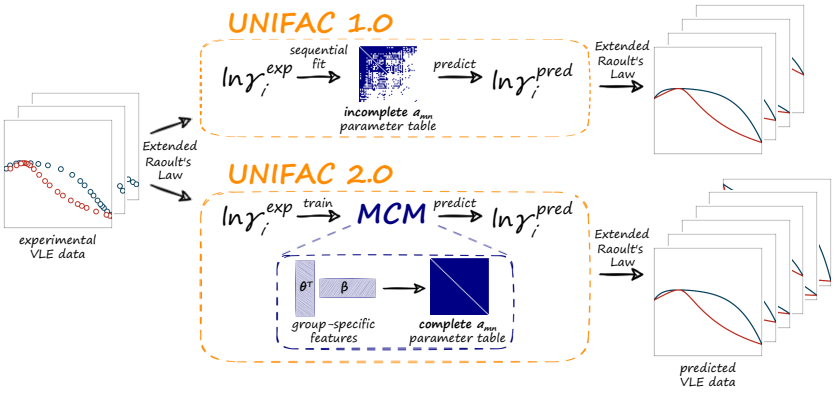

Fig. 1. Comparison of UNIFAC 1.0 and UNIFAC 2.0. UNIFAC 1.0 relies on sequential parameter fitting, whereas UNIFAC 2.0 integrates an MCM for predicting pair-interaction parameters and is trained end-to-end on experimental logarithmic activity coefficients ($\ln \gamma_i$) derived from VLE data.

2 Method: UNIFAC 2.0

Fig. 1 compares the original UNIFAC 1.0, based on sequential and sometimes inconsistent individual parameter fitting, with the proposed UNIFAC 2.0, based on end-to-end training of MCM features on 224,562 experimental data points, giving a consistent, complete parameterization; both UNIFAC variants rely on the same structural groups and physical model equations.

The MCM is based on decomposing the matrix of the pair-interaction parameters a_{mn} into the product of two feature matrices, thereby enabling the prediction of missing matrix entries through learned interactions. Each a_{mn} is thereby modeled as follows:

$$a_{mn} = \boldsymbol{\theta}_n^{\mathrm{T}} \cdot \boldsymbol{\beta}_m. \tag{1}$$

Here, $\boldsymbol{\theta}_n$ and $\boldsymbol{\beta}_m$ are column vectors of length K, with K representing the latent dimension, a hyperparameter that was determined to be $K = 8$ in preliminary studies.

Our proposed probabilistic model integrates observations ($\ln \gamma_i$) and the latent variables (LVs) that characterize UNIFAC main groups ($\boldsymbol{\theta}_n$, $\boldsymbol{\beta}_m$) within a Bayesian framework. Specifically, all $\ln \gamma_i$ and LVs are modeled as independent random variables. A standard normal distribution is used as prior for each

LV. The likelihood of observing $\ln \gamma_i$, given the LVs, follows a Cauchy distribution centered around the predicted activity coefficients $\ln \gamma_i^{\text{UNIFAC 2.0}}$ with scale parameter $\lambda = 0.4$:

$$p(\ln \gamma_i \,|\, \boldsymbol{\theta}_n, \boldsymbol{\beta}_m) = \text{Cauchy}(\ln \gamma_i^{\text{UNIFAC 2.0}},\ \lambda), \tag{2}$$

where $\ln \gamma_i^{\text{UNIFAC 2.0}}$ is determined via the standard UNIFAC equations [17] using the predicted interaction parameters a_{mn}.

Written in Pyro [1], our probabilistic model adopts stochastic variational inference (VI) [2] for posterior approximation. This approach leverages the Adam optimizer [11], with a learning rate of 0.15. A normal distribution is employed as the variational distribution, with all LVs being treated independently.

3 Results and Discussion

Fig. 2 compares the performance of UNIFAC 2.0 to that of the original UNIFAC 1.0 in terms of mean absolute error (MAE) and the mean squared error (MSE) for a test set containing 27,287 data points and covering 2,603 different binary mixtures, cf. Fig. 2. Since UNIFAC 2.0 has a larger scope than UNIFAC 1.0,

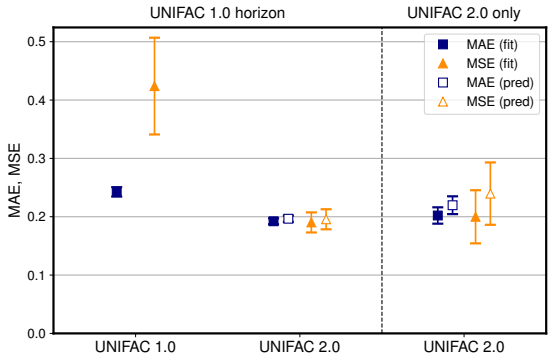

Fig. 2. MAE and MSE of the predicted $\ln \gamma_i$ of the test set (pred). For comparison, the results of UNIFAC 2.0 trained on all experimental data and UNIFAC 1.0 are also shown (fit). UNIFAC 1.0 can only predict 25,998 data points for 2,202 binary mixtures (UNIFAC 1.0 horizon), while an additional 1,289 test data points for 401 binary mixtures can only be predicted by UNIFAC 2.0 (UNIFAC 2.0 only). Error bars denote standard errors of the means.

a distinction is made: all data points that both methods can predict are summarized in the *UNIFAC 1.0 horizon*, all data points where only UNIFAC 2.0 is applicable are summarized as *UNIFAC 2.0 only*. Fig. 2 clearly shows the superior predictive accuracy of UNIFAC 2.0 over UNIFAC 1.0 in both error scores. Even more importantly, the new method not only improves accuracy for data points

within the predictive range of UNIFAC 1.0, but it also maintains this accuracy for data points beyond the scope of UNIFAC 1.0. Additionally, Fig. 2 shows that the accuracy of the true predictions with UNIFAC 2.0 obtained by strictly withholding the test data during the training (open symbols) is only marginally smaller than that of the UNIFAC 2.0 version that has been trained on all data points (closed symbols); this holds for both the "UNIFAC 1.0 horizon" and the "UNIFAC 2.0 only" subset.

The activity coefficients obtained by UNIFAC 2.0 can be used directly to predict phase equilibria of mixtures, which are at the core of the design and optimization of thermal separation processes in chemical engineering. In Fig. 3, we show isothermal vapor-liquid phase diagrams for two ternary mixtures predicted by UNIFAC 2.0 as examples. Although no data on multi-component mixtures were used for training UNIFAC 2.0, the underlying physical framework of UNIFAC also enables predictions for such mixtures. Excellent accuracy is found.

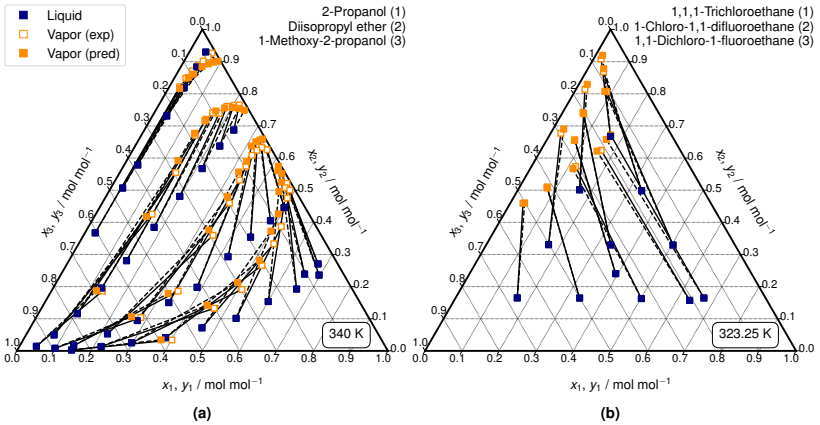

Fig. 3. Prediction of isothermal vapor-liquid phase diagrams for ternary mixtures with UNIFAC 2.0 (pred) and comparison to experimental data (exp) from the DDB. The temperature and the composition of the liquid phase were specified, and the composition of the corresponding vapor phase in equilibrium was predicted.

4 Conclusions

We introduce UNIFAC 2.0, a hybrid model based on a MCM embedded into the UNIFAC framework, thereby combining machine learning with established physical models and addressing their limitations. UNIFAC 2.0 shows significantly superior performance than the original model and even maintains high predictive accuracy in cases where UNIFAC 1.0 is not applicable. The hybridization approach can easily be extended to other physical thermodynamic models.

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. Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, Paul and Horsfall, Paul, Goodman, N.D.: Pyro: Deep universal probabilistic programming. Journal of Machine Learning Research (2018)
2. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: A review for statisticians. Journal of the American Statistical Association **112**(518), 859–877 (2017). https://doi.org/10.1080/01621459.2017.1285773
3. Fredenslund, A., Jones, R.L., Prausnitz, J.M.: Group-contribution estimation of activity coefficients in nonideal liquid mixtures. AIChE Journal **21**(6), 1086–1099 (1975). https://doi.org/10.1002/aic.690210607
4. Gmehling, J., Rasmussen, P., Fredenslund, A.: Vapor-liquid equilibriums by unifac group contribution. revision and extension. 2. Industrial & Engineering Chemistry Process Design and Development **21**(1), 118–127 (1982). https://doi.org/10.1021/i200016a021
5. Hansen, H.K., Rasmussen, P., Fredenslund, A., Schiller, M., Gmehling, J.: Vapor-liquid equilibria by unifac group contribution. 5. revision and extension. Industrial & Engineering Chemistry Research **30**(10), 2352–2355 (1991). https://doi.org/10.1021/ie00058a017
6. Hayer, N., Jirasek, F., Hasse, H.: Prediction of henry's law constants by matrix completion. AIChE Journal **68**(9), 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**(3), 981–985 (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**(17), 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**(82), 12407–12410 (2020). https://doi.org/10.1039/D0CC05258B
10. 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 : PCCP **25**(2), 1054–1062 (2023). https://doi.org/10.1039/d2cp04478a
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization, http://arxiv.org/pdf/1412.6980.pdf
12. Macedo, E.A., Weidlich, U., Gmehling, J., Rasmussen, P.: Vapor-liquid equilibriums by unifac group contribution. revision and extension. 3. Industrial & Engineering Chemistry Process Design and Development **22**(4), 676–678 (1983). https://doi.org/10.1021/i200023a023
13. Mohs, A., Jakob, A., Gmehling, J.: Analysis of a concept for predicting missing group interaction parameters of the unifac model using connectivity indices. AIChE journal **55**(6), 1614–1625 (2009)

14. Ramlatchan, A., Yang, M., Liu, Q., Li, M., Wang, J., Li, Y.: A survey of matrix completion methods for recommendation systems. Big Data Mining and Analytics **1**(4), 308–323 (2018). https://doi.org/10.26599/BDMA.2018.9020008
15. Skjold-Jorgensen, S., Kolbe, B., Gmehling, J., Rasmussen, P.: Vapor-liquid equilibria by unifac group contribution. revision and extension. Industrial & Engineering Chemistry Process Design and Development **18**(4), 714–722 (1979). https://doi.org/10.1021/i260072a024
16. Tiegs, D., Rasmussen, P., Gmehling, J., Fredenslund, A.: Vapor-liquid equilibria by unifac group contribution. 4. revision and extension. Industrial & Engineering Chemistry Research **26**(1), 159–161 (1987). https://doi.org/10.1021/ie00061a030
17. Wittig, R., Lohmann, J., Gmehling, J.: Vapor−liquid equilibria by unifac group contribution. 6. revision and extension. Industrial & Engineering Chemistry Research **42**(1), 183–188 (2003). https://doi.org/10.1021/ie0205061