Balancing Molecular Information and Empirical Data in the Prediction of Physico-Chemical Properties

Johannes Zenn[1,2,3][0009−0001−5834−3474], Dominik Gond[4][0000−0003−3958−6945], Fabian Jirasek[4][0000−0002−2502−5701], and Robert Bamler[1,2][0000−0002−3135−8107]

[1] AI Center Tübingen
[2] University of Tübingen
{johannes.zenn,robert.bamler}@uni-tuebingen.de
[3] IMPRS-IS
[4] Laboratory of Engineering Thermodynamics (LTD), RPTU Kaiserslautern
{dominik.gond,fabian.jirasek}@rptu.de

Abstract. Predicting the physico-chemical properties of pure substances and mixtures is a central task in thermodynamics. Established prediction methods range from fully physics-based ab-initio calculations, over descriptor-based methods, to representation-learning methods, which, in extreme cases, may completely ignore molecular structure and extrapolate only from existing experimental data (e.g., matrix completion methods). In this work, we propose a general method for combining molecular descriptors with representation learning using the so-called expectation maximization algorithm from the probabilistic machine-learning literature, which uses uncertainty estimates to trade off between the two approaches. The proposed hybrid model uses graph neural networks to exploit chemical structure information, but it detects and corrects unreliable structure-based predictions by more specialized representation-learning based predictions. Our method significantly improves predictive accuracy over the current state of the art in the example problem of predicting activity coefficients in binary mixtures, showcasing its potential to advance the prediction of physico-chemical properties in general.

This document is an extended abstract to Ref. 22 by the same authors, available at: https://arxiv.org/abs/2406.08075

Keywords: thermodynamics · machine learning · activity coefficients.

1 Introduction

Information on physico-chemical properties is crucial for the design and optimization of processes in many industries, including chemistry, pharmacy, and biotechnology. As an important example, activity coefficients of the components in a mixture, which describe the deviation from the ideal mixture, are the basis for modeling any phase and reaction equilibria. As measuring such thermodynamic properties for all relevant mixtures would be infeasible [14], prediction methods for thermodynamic properties of mixtures are paramount.

Recently, research on such prediction methods has split into two branches: (a) descriptor-based methods, which correlate information on the molecules to be modeled (e.g., molecular weight, surface area, or composition in terms of structural groups) with properties of interest; and (b) so-called matrix completion methods (MCMs) [11,13,10,9,4] from the machine-learning literature, which ignore chemical structure and fit individual abstract representation vectors for each mixture component that appears in a set of available experimental data.

In statistics parlance, descriptor-based models are called "*parametric*" whereas MCMs are called "*nonparametric* in the components" (despite having many parameters, but each parameter in a nonparametric model is only relevant to a single component and can thus be seen as a parameter of the component rather than a parameter of the model). Nonparametric models are more flexible, and, empirically, MCMs were shown to predict activity coefficients more accurately [11,13,12] than the (parametric) group-contribution method UNIFAC [6,21,3], which is still considered the gold standard for property prediction in many (industrial) fields [7,15]. But nonparametric models cannot exploit structural similarities across components nor extrapolate to new components.

In our work [22], we propose a new method for predicting activity coefficients (and thermodynamic properties in general) in binary mixtures that combines the strengths of both the parametric (descriptor-based) and the nonparametric (representation-based) approach, while avoiding their respective weaknesses. We jointly fit probabilistic variants of both model types using the so-called variational expectation maximization (variational EM) [5,1] algorithm, which finds a compromise based on the two models' uncertainty estimates. Our evaluation shows that this uncertainty-based trade-off indeed improves predictive accuracy.

2 Method

As an example prediction problem of physico-chemical properties, we consider activity coefficients in binary mixtures at infinite dilution at 298.15 ($\pm$1) K (our method can be extended to arbitrary concentrations and temperatures by following [12]). We start from a dataset of 4094 experimentally measured activity coefficients $\gamma_{i,j}^{\infty}$ between $M = 240$ distinct solutes i and $N = 250$ distinct solvents j taken from the Dortmund Data Bank (DDB) [19]. Our goal is to predict activity coefficients for mixtures where no experimental data is available, including mixtures that involve at least one component that does not appear in the experimental data (referred to as "out-of-domain prediction" in the following).

Our proposed prediction method uses a probabilistic model that combines a parametric with a nonparametric part. The parametric part uses graph neural networks (GNNs) [8,20,2] that map the molecular graph structures of the solute i and solvent j, respectively, and output (so-called conditional prior distributions over) representation vectors $u_i, v_j \in \mathbb{R}^K$ (the dimension K is a modeling choice). The nonparametric model part is a probabilistic matrix completion method (MCM) [11], which learns (a so-called variational distribution over) u_i and v_j for each solute and solvent, respectively, that appears in the data.

Crucially, we train both parts jointly with the so-called variational expectation maximization (variational EM) [5,1] algorithm. Variational EM is a so-called empirical Bayes method, i.e., it is similar to Bayesian inference [18] with the exception that not only the posterior distribution but also the prior distribution is (albeit to a lesser degree) informed by the data. Our main paper [22] discusses variational EM in detail. In this extended abstract, we focus on showing how easily variational EM can be implemented in practice and how effective it is empirically for the prediction of thermodynamic quantities. The combination of both models parts defines a latent variable model with the joint distribution

$$p_\theta(\mathbf{u}, \mathbf{v}, \ln \boldsymbol{\gamma}^\infty \,|\, \mathbf{r}, \mathbf{s}) = \Big(\prod_{i=1}^{M} p_\theta(u_i|r_i)\Big)\Big(\prod_{j=1}^{N} p_\theta(v_j|s_j)\Big)\Big(\prod_{(i,j)\in\mathcal{D}} p(\ln \gamma_{i,j}^\infty|u_i, v_j)\Big). \quad (1)$$

Here, boldface symbols $\mathbf{u}$, $\mathbf{v}$, $\mathbf{r}$, $\mathbf{s}$, and $\boldsymbol{\gamma}^\infty$ on the left-hand side denote the collection of all representation vectors u_i and v_j, chemical structures r_i and s_j, and activity coefficients $\gamma_{i,j}^\infty$ in the experimental data $\mathcal{D}$. On the right-hand side, $p_\theta(u_i|r_i)$ and $p_\theta(v_j|s_j)$ are the conditional prior distributions that are parameterized by the GNNs, and $p(\ln \gamma_{i,j}^\infty|u_i, v_j)$ is a simple (fixed) likelihood function.

Variational EM is an approximate variant of EM [5], which jointly learns model parameters (i.e., neural network weights) θ that maximize the marginal likelihood $p_\theta(\ln \boldsymbol{\gamma}^\infty \,|\, \mathbf{r}, \mathbf{s}) = \iint p_\theta(\mathbf{u}, \mathbf{v}, \ln \boldsymbol{\gamma}^\infty \,|\, \mathbf{r}, \mathbf{s})\, \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{v}$ of the data while also finding the posterior $p_\theta(\mathbf{u}, \mathbf{v} \,|\, \mathbf{r}, \mathbf{s}, \ln \boldsymbol{\gamma}^\infty) = p_\theta(\mathbf{u}, \mathbf{v}, \ln \boldsymbol{\gamma}^\infty \,|\, \mathbf{r}, \mathbf{s}) \,/\, p_\theta(\ln \boldsymbol{\gamma}^\infty \,|\, \mathbf{r}, \mathbf{s})$. Variational EM makes this task computationally feasible by introducing so-called variational distributions $q_\phi(u_i)$ and $q_\phi(v_j)$, which will end up approximating the (marginal) posterior distributions, and which we choose to be Gaussian distributions with diagonal covariance matrices, where the free parameters ϕ are the means and variances. Instead of maximizing the marginal likelihood, variational EM then maximizes a lower bound to it called the evidence lower bound (ELBO),

$$\begin{aligned} \mathrm{ELBO}(\theta, \phi) = &\sum_{(i,j)\in\mathcal{D}} \mathbb{E}_{q_\phi(u_i)\, q_\phi(v_j)}\big[\ln p(\ln \gamma_{i,j}^\infty \,|\, u_i, v_j)\big] \\ &- \sum_{i=1}^{M} D_{\mathrm{KL}}\big(q_\phi(u_i) \,\big\|\, p_\theta(u_i \,|\, r_i)\big) - \sum_{j=1}^{N} D_{\mathrm{KL}}\big(q_\phi(v_j) \,\big\|\, p_\theta(v_j \,|\, s_j)\big). \end{aligned} \quad (2)$$

Here, $\mathbb{E}[\,\cdot\,]$ denotes the expectation value and D_{KL} the Kullback-Leibler (KL) divergence [16,18], which quantifies how much the variational distributions differ from the conditional priors, and which can be calculated analytically for normal distributions [18]. Our main paper [22] discusses the ELBO in detail. In this extended abstract, we only highlight that (i) maximizing the ELBO over both θ and ϕ fits the variational distributions q_ϕ to the data in a nonparametric way that is, however, regularized by the (parametric) GNNs, where the regularization is stronger in cases where the GNNs estimate their own uncertainty as low; and (ii) the ELBO can be easily maximized with standard stochastic gradient descent using automatic differentiation, i.e., with the same techniques that are usually used for neural network training (see explicit pseudocode in our main paper [22]).

3 Evaluation Setup and Results

We evaluate our proposed method "GNN MCM" and a simpler variant, "MoFo MCM", which replaces the GNNs with neural networks that operate on the molecular formula. As baselines, we compare to modified UNIFAC (Dortmund) [21,3] (called "UNIFAC" below), a fully nonparametric MCM [11], and a fully parametric GNN-based method [17]. For training setup, detailed results, and further details see our full paper [22].

Table 1. Comparison of the proposed GNN MCM with the fully parametric model from [17].

model	MAE	MSE
GNN MCM	$\mathbf{0.1542}_{\pm \mathbf{0.0046}}$	$\mathbf{0.0905}_{\pm \mathbf{0.0071}}$
Medina et al. [17]	$0.1973_{\pm 0.0067}$	$0.1196_{\pm 0.0074}$

We evaluate in-domain predictions (both solute and solvent appear in the experimental data) using q_ϕ and out-of-domain predictions (at least one of solute or solvent is new) using the conditional priors. We further investigate two ablation studies: (1) using only the conditional priors of the same trained models for in-domain predictions; and (2) removing uncertainty estimates and training by simple maximum likelihood estimation (MLE) rather than variational EM.

Figure 1 and Table 1 summarize our results by showing the mean absolute error (MAE) and mean squared error (MSE) of the predicted logarithmic activity coefficients of all evaluated models. Light hatched bars in Figure 1 represent models trained and evaluated only on data points that can be modeled by UNIFAC. Comparisons to [17] are shown in Table 1 using the same dataset as in [17].

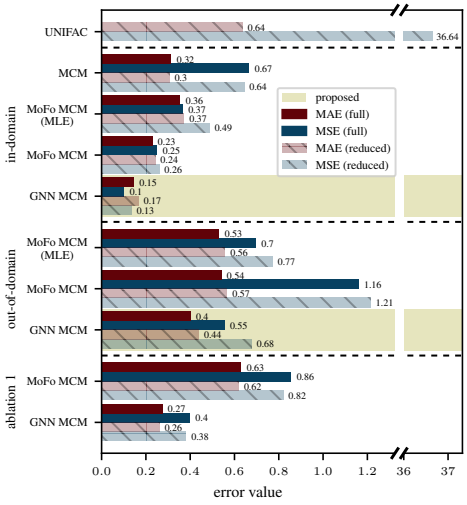

Fig. 1. Prediction errors of our models, baselines, and ablations. The proposed GNN MCM has best predictive accuracy for both in- & out-of-domain predictions.

The proposed GNN MCM makes more accurate predictions than all considered baselines, both in terms of MAE and MSE, and for both in-domain and out-of-domain predictions. Comparing the last two rows of Figure 1 (ablation 1) to rows four and five shows that the nonparametric fits are indeed useful where they are available (i.e., for in-domain predictions). Ablation 2 shows that variational EM significantly improves in-domain predictions over MLE-training, but the picture is less clear for out-of-domain predictions. See full paper [22] for further discussion.

Conclusions. We propose a generic method for predicting physico-chemical properties that uses variational EM to combine the generalization capability of structure-based methods with the flexibility of nonparametric representation-learning. Our method significantly improves upon the state of the art in the example problem of predicting activity coefficients in binary mixtures, and it can easily be applied to similar physico-chemical prediction problems as well.

Acknowledgements

Johannes Zenn thanks Tim Xiao for helpful comments and discussions. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors would like to acknowledge support of the 'Training Center Machine Learning, Tübingen' with grant number 01|S17054. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Johannes Zenn. Robert Bamler acknowledges funding by the German Research Foundation (DFG) for project 448588364 of the Emmy Noether Program. Fabian Jirasek gratefully acknowledges funding by DFG for project 528649696 of the Emmy Noether Program and for project 497201843 in the Priority Program Molecular Machine Learning (SPP 2363).

References

1. Beal, M.J., Ghahramani, Z.: The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. Bayesian statistics **7**, 453–464 (2003)
2. Bronstein, M.M., Bruna, J., Cohen, T., Veličković, P.: Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478 (2021)
3. Constantinescu, D., Gmehling, J.: Further development of modified unifac (dortmund): Revision and extension 6. Journal of Chemical & Engineering Data **61**(8), 2738–2748 (2016). https://doi.org/10.1021/acs.jced.6b00136, https://doi.org/10.1021/acs.jced.6b00136
4. Damay, J., Jirasek, F., Kloft, M., Bortz, M., Hasse, H.: Predicting activity coefficients at infinite dilution for varying temperatures by matrix completion. Industrial & Engineering Chemistry Research **60**(40), 14564–14578 (2021)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (methodological) pp. 1–38 (1977)
6. 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/https://doi.org/10.1002/aic.690210607, https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.690210607
7. Gmehling, J., Constantinescu, D., Schmid, B.: Group contribution methods for phase equilibrium calculations. Annual Review of Chemical and Biomolecular Engineering **6**(Volume 6, 2015), 267–292 (2015)
8. Gori, M., Monfardini, G., Scarselli, F.: A new model for learning in graph domains. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. vol. 2, pp. 729–734. IEEE (2005)
9. 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 **1**, 886–897 (2022)

10. Hayer, N., Jirasek, F., Hasse, H.: Prediction of henry's law constants by matrix completion. AIChE Journal **68**(9), e17753 (2022)
11. Jirasek, F., Alves, R.A., 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)
12. 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)
13. Jirasek, F., Bamler, R., Mandt, S.: Hybridizing physical and data-driven prediction methods for physicochemical properties. Chemical Communications **56**(82), 12407–12410 (2020)
14. 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)
15. Jirasek, F., Hayer, N., Abbas, R., Schmid, B., Hasse, H.: Prediction of parameters of group contribution models of mixtures by matrix completion. Phys. Chem. Chem. Phys. **25**, 1054–1062 (2023)
16. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**, 79 – 86 (1951)
17. Medina, E.I.S., Linke, S., Stoll, M., Sundmacher, K.: Graph neural networks for the prediction of infinite dilution activity coefficients. Digital Discovery (2022)
18. Murphy, K.P.: Probabilistic machine learning: an introduction. MIT press (2022)
19. 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**(5), 739–747 (1989). https://doi.org/10.1007/BF00507993
20. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: Computational capabilities of graph neural networks. IEEE Transactions on Neural Networks **20**(1), 81–102 (2008)
21. Weidlich, U., Gmehling, J.: A modified unifac model. 1. prediction of vle, he, and .gamma..infin. Industrial & Engineering Chemistry Research **26**(7), 1372–1381 (1987). https://doi.org/10.1021/ie00067a018, https://doi.org/10.1021/ie00067a018
22. Zenn, J., Gond, D., Jirasek, F., Bamler, R.: Balancing molecular information and empirical data in the prediction of physico-chemical properties. arXiv e-prints pp. arXiv–2406 (2024)