ChemComm

COMMUNICATION

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2020, **56**, 12407

Received 1st August 2020,
Accepted 10th September 2020

DOI: 10.1039/d0cc05258b

rsc.li/chemcomm

Hybridizing physical and data-driven prediction methods for physicochemical properties†

Fabian Jirasek, *‡ Robert Bamler and Stephan Mandt

We present a generic way to hybridize physical and data-driven methods for predicting physicochemical properties. The approach 'distills' the physical method's predictions into a prior model and combines it with sparse experimental data using Bayesian inference. We apply the new approach to predict activity coefficients at infinite dilution and obtain significant improvements compared to the physical and data-driven baselines and established ensemble methods from the machine learning literature.

Prediction methods for physicochemical properties are indispensable for process design and optimization in chemical engineering since experimental studies are expensive and tedious. The most widely used approaches are group-contribution methods (GCMs) that model the properties of pure components or mixtures based on the structural groups that build up the components.[1–4] GCMs can also be used for predicting properties of mixtures of which the composition is unknown.[5–7] The most successful GCMs for mixtures are the different versions of UNIFAC[8–10] that model the excess Gibbs energy based on binary group-interaction parameters. The group-contribution concept greatly reduces the number of model parameters and the amount of data needed for fitting GCMs. However, the practical applicability of UNIFAC is still restricted, mainly due to necessary group-interaction parameters that have not been fitted yet. Another successful approach is the quantum chemistry-based COSMO-RS,[11] which describes the properties of mixtures referring to the polarization charge densities of the constituent components, and which depends only on a small number of adjustable parameters.[12] However, expensive COSMO calculations are required for each component.

Department of Computer Science, University of California, Donald Bren Hall, Irvine, CA 92697, USA. E-mail: fabian.jirasek@mv.uni-kl.de

† Electronic supplementary information (ESI) available: Data and data preprocessing; brief review on Bayesian matrix completion and detailed model descriptions; additional results. See DOI: 10.1039/d0cc05258b

‡ Present address: Laboratory of Engineering Thermodynamics (LTD), TU Kaiserslautern, Erwin-Schrödinger-Straße 44, 67663 Kaiserslautern, Germany.

In previous work,[13] we have introduced a novel, purely data-driven approach to predict physicochemical properties of mixtures. Specifically, we considered activity coefficients at infinite dilution γ_{ij}^{∞} in binary mixtures at a constant temperature, but this approach generalizes to other properties. The data for γ_{ij}^{∞} can be represented as a matrix whose rows and columns correspond to solutes i and solvents j, respectively. For γ_{ij}^{∞} at 298.15 ± 1 K, which we studied in our previous work, the matrix containing the available experimental data from one of the largest databases for physicochemical properties, the Dortmund Data Bank,[14] is very sparse, *cf.* Fig. S.1 (ESI†). The data set covers 240 solutes and 250 solvents, but only 4094 entries are observed. The prediction of the unobserved entries, *i.e.*, the prediction of γ_{ij}^{∞} for not yet studied mixtures, can be framed as a matrix completion problem.[15,16]

The basis of our previously introduced approach[13] is a probabilistic matrix completion method (MCM). We modeled $\ln \gamma_{ij}^{\infty}$ (the logarithm of γ_{ij}^{∞} is used for scaling purposes) as a stochastic function of initially unknown features of the solutes i and solvents j, specifically as the dot product of two vectors:

$$\ln \gamma_{ij}^{\infty} = u_i \cdot v_j + \varepsilon_{ij} \quad (1)$$

where u_i and v_j are learned feature vectors for solute i and solvent j, respectively, and the random variable ε_{ij} captures both measurement noise and inaccuracies of the model. The feature vectors of all considered solutes and solvents can be aggregated to two feature matrices U and V, respectively. Rather than selecting features based on physical considerations, the data-driven approach infers useful features from available experimental data on $\ln \gamma_{ij}^{\infty}$ alone, using the laws of probability theory and (approximate) Bayesian inference.[17–19] The inferred features can then be used to predict $\ln \gamma_{ij}^{\infty}$ for mixtures for which no experimental data are available, *cf.* eqn (1).

While the purely data-driven approach[13] already outperforms the state-of-the-art physical method for predicting activity coefficients modified UNIFAC (Dortmund)[10,20] (to which we simply refer as UNIFAC in the following) in terms of average predictive performance, it leaves substantial room for improvement

as it ignores available physical knowledge about the mixtures. In thermodynamics, such knowledge is often abundant, *e.g.*, in pure component properties or physical laws and models. In this paper, we therefore propose a hybrid physics-based/data-driven prediction method that combines the best of both worlds. We show that the framework of probabilistic models and Bayesian inference provides a principled way to incorporate scientific domain knowledge into machine learning (ML) models by specifying a so-called prior probability distribution over model parameters. Specifically, we propose to use model distillation[21] to extract physical domain knowledge from UNIFAC in a format that can be used to construct an informative prior distribution for the MCM.

In the following, we describe the details of our proposed hybrid approach. Once again, we consider predicting $\ln \gamma_{ij}^{\infty}$ in binary mixtures at 298.15 ± 1 K as a prime example and evaluate the predictive performance on the same data set as in our previous work.[13] As physical base method, we use the current publicly available version of UNIFAC.[10,20] As data-driven base method, we adopt the Bayesian MCM from our previous work.[13] We compare the performance of our hybrid method to the performances of the constituent base methods as well as two established ML ensemble methods.

Fig. 1 summarizes our proposed hybrid method, which we call whisky. Just like the manufacturing of whisky, our whisky method involves a distillation step, in which we distill knowledge from an existing model into a prior distribution using an approach known as model distillation in the ML literature,[21] and a maturation step, in which we allow the prior to mature by combining it with experimental data. Both steps are based on a probabilistic MCM similar to our previous work[13] to fit model parameters (*i.e.*, feature matrices U and V) to a data set of $\ln \gamma_{ij}^{\infty}$. The difference between the distillation and maturation step is that they operate on different data sets. The distillation step fits an MCM to all predictions for $\ln \gamma_{ij}^{\infty}$ at 298.15 K that can be obtained with UNIFAC, denoted as $\ln \gamma^{\text{UNIFAC}}$. Thus, the distillation step extracts the physical knowledge encoded in UNIFAC, which is implicitly exposed via its predictions for $\ln \gamma_{ij}^{\infty}$, into parameters of an MCM. By contrast, the maturation step builds upon the results of the distillation step and refines the parameters by fitting an MCM to the available experimental data, denoted as $\ln \gamma^{\text{exp}}$.

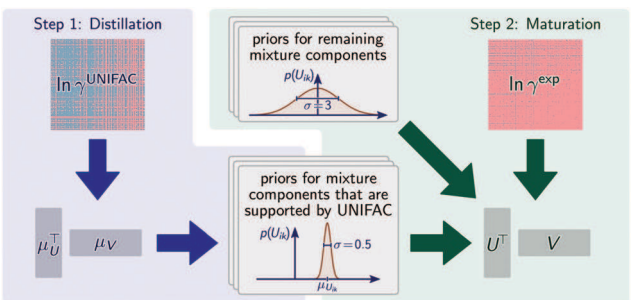


Fig. 1 Scheme of the whisky method. We first fit an MCM to UNIFAC predictions for $\ln \gamma_{ij}^{\infty}$ (distillation, purple). We then use the fitted parameters from the distillation step to construct informative priors for the component feature matrices U and V and fit the model to experimental data on $\ln \gamma_{ij}^{\infty}$ using these priors (maturation, green). $\ln \gamma^{\text{UNIFAC}}$ and $\ln \gamma^{\text{exp}}$ denote the available data sets from UNIFAC[20] and experiments,[14] respectively.

The two different data sets $\ln \gamma^{\text{UNIFAC}}$ and $\ln \gamma^{\text{exp}}$ are illustrated in the two blue/red matrices in Fig. 1. Here, rows and columns correspond to solutes and solvents, respectively, and blue or red entries indicate binary mixtures for which data points are available or absent, respectively. As can be seen, UNIFAC predictions are available for a lot more mixtures than experimental observations ($\ln \gamma^{\text{UNIFAC}}$ has more blue entries than $\ln \gamma^{\text{exp}}$, *cf.* Fig. S.1 and S.3, ESI†), meaning that the distillation step trains on a larger data set. While the experimental data set is more sparse, it is considered more reliable than the UNIFAC predictions.

The main novelty of our proposed whisky method lies in the way how it combines physical knowledge with experimental data. We realize the interface between distillation and maturation (purple and green parts of Fig. 1) by specifying an informative prior distribution over model parameters. To understand the role of the prior, it is instructive to recall the principles of Bayesian inference on which our MCM builds. Bayesian inference describes the relationship between three probability distributions, called prior, likelihood, and posterior. The prior is a probability distribution over model parameters that encodes a-priori knowledge, *i.e.*, information on the model parameters before the model is fitted to the training data. In a purely data-driven approach, no a-priori information is used, and the prior is typically a very broad (*i.e.*, noninformative) probability distribution. The likelihood encodes how model parameters manifest themselves in physically observable quantities, *i.e.*, the data to which the model is trained. Together, prior and likelihood define a probabilistic model over observable quantities, such as $\ln \gamma_{ij}^{\infty}$ here. Bayesian inference takes such a probabilistic model and compares its predictions to actual observed data. The task of Bayesian inference is to find the so-called posterior probability distribution over model parameters that are consistent with the observed quantities and the a-priori knowledge.

This framework of probabilistic modeling and Bayesian inference provides a principled way of hybridizing different methods using probability distributions as interfaces. Our approach, illustrated in Fig. 1, follows the principle that 'one man's ceiling is another man's floor'. In analogy to this proverb, the posterior of the distillation step, which encodes knowledge after seeing the UNIFAC predictions, can be turned into a prior of the maturation step, which encodes knowledge before seeing the experimental data. Specifically, we construct a physically informed prior for the maturation step by taking the posterior means μ_U and μ_V from the distillation step, and we form Gaussian prior distributions with a rather small standard deviation of σ = 0.5 around these means. Thus, this choice of prior encodes physical knowledge from the UNIFAC model. At the same time, the nonzero prior standard deviation allows the maturation step to overrule prior knowledge if the experimental data provide enough evidence to justify this.

For some of the considered mixture components (eight solutes and 41 solvents), UNIFAC is not applicable. Since the

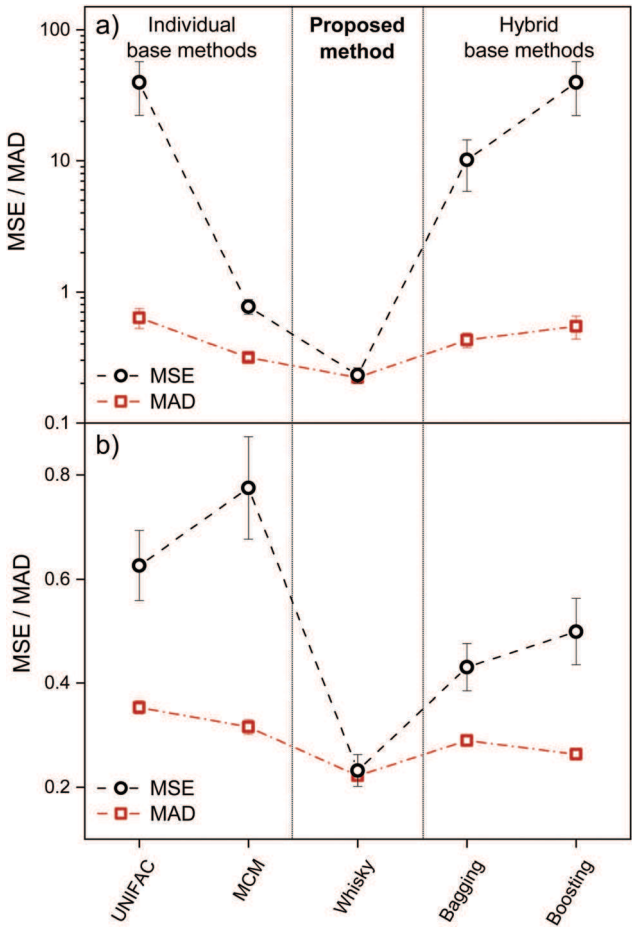


Fig. 2 Mean square error (MSE) and mean absolute deviation (MAD) for the prediction of $\ln \gamma_{ij}^{\infty}$ using the individual base methods UNIFAC and data-driven MCM, the proposed hybrid whisky method, and the hybrid baselines bagging and boosting. Lower is better for both metrics. Error bars show the standard errors of the means. (a) Considering all applicable data points. (b) Ignoring the worst eight outliers of UNIFAC.

distillation step does not provide any information about these components, we use a broader (*i.e.*, less informative) Gaussian prior here in the maturation step, with a standard deviation of $\sigma = 3$ centered around zero. For the task of (approximate) Bayesian inference, we use the Stan framework[22] and resort to variational inference.[18,19] More details on the proposed whisky method are given in the ESI.†

In Fig. 2(a), we compare the overall performance of the whisky method for predicting $\ln \gamma_{ij}^{\infty}$ with the performances of the base methods UNIFAC[20] and MCM[13] (without the informative prior), and with two alternative hybrid approaches, bootstrap aggregation (aka bagging)[23] and boosting.[24] We compare mean absolute deviation (MAD) and mean square error (MSE).

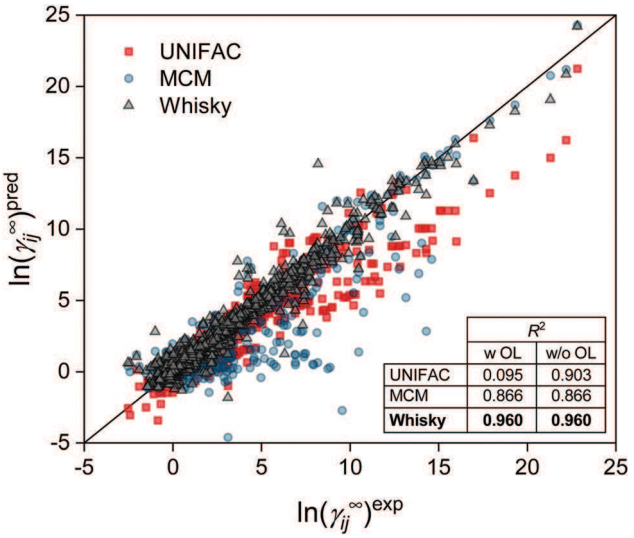
Bagging is realized here by averaging the predictions from UNIFAC and the data-driven MCM for each data point; boosting is implemented by training an MCM to the matrix of the residuals of UNIFAC. Bagging and boosting are described in detail in the ESI.† To simulate predictive performances, the predictions with MCM, whisky, and boosting (and the MCM contribution of bagging) are obtained by using leave-one-out cross-validation,[25] *i.e.*, by training the models to all experimental data points except for one, which is then used as a test data point and predicted. The training set of UNIFAC is not disclosed; hence, no statements on whether the UNIFAC results are obtained by regression or prediction can be made here.

Fig. 2(a) demonstrates that the proposed whisky method outperforms all other methods in both MAD and MSE. The poor scores of UNIFAC, bagging, and boosting can mainly be attributed to only a handful of data points that are extremely poorly predicted by UNIFAC as shown in Fig. S.8 (ESI†). However, even if we, as an example, ignore the worst eight outliers of UNIFAC (marked in Fig. S.8, ESI†) for the evaluation, the proposed whisky method still performs significantly better than all baselines, *cf.* Fig. 2(b).

If the worst eight UNIFAC outliers are ignored (Fig. 2b), the results show that the hybrid baselines – bagging and boosting – also improve the predictions of the base methods UNIFAC and MCM: bagging and boosting have smaller MAD and MSE values than the base methods. Bagging is widely used if the available base methods for a specific problem tend to overfit, *i.e.*, if they fit the training data but do not generalize well to unobserved data.[25] By contrast, boosting is commonly applied in ML to tackle the opposite problem of underfitting, which arises if the base methods are not expressible enough for a specific problem.[24] The observation that our proposed whisky method performs better than both bagging and boosting indicates that the base methods UNIFAC and data-driven MCM tend to overfit to parts of the data set. At the same time, they also seem to underfit on other combinations, so that neither bagging nor boosting is universally applicable. This may in part be explained by the fact that the experimental data set is very imbalanced: while we have data for at least 86 different binary mixtures for each of the 5% most common solutes, we only have six or fewer data points for each of the 50% most uncommon solutes (see also Fig. S.1, ESI†). The proposed whisky approach seems more robust to such an imbalanced data set than the other hybrid approaches.

In Fig. 3, we compare the predictions of the whisky method with those of the data-driven MCM and UNIFAC in a parity plot. Points on the diagonal line correspond to perfect predictions. The whisky method reliably reduces outliers of both base methods. By contrast, both bagging and boosting, shown in Fig. S.7 (ESI†), only partially compensate for outliers of the data-driven MCM but severely suffer from outliers of UNIFAC. The whisky method also yields the highest coefficient of determination R^2 (with $R^2 = 1$ being optimal) of all compared methods, irrespective of whether the worst eight UNIFAC outliers (OL) are considered or not (see table insets).

Another major advantage of the proposed whisky method is its broader applicability compared to the other hybrid approaches. For a fair comparison, Fig. 2 and 3 consider only data points that can be predicted with UNIFAC, which is also a prerequisite for applying bagging and boosting. By contrast, the whisky method (and the data-driven MCM) can be used to predict $\ln \gamma_{ij}^{\infty}$ for any binary mixture of the considered solutes and solvents. In Fig. S.9 (ESI†), we compare the performance of



	R^2	
	w OL	w/o OL
UNIFAC	0.095	0.903
MCM	0.866	0.866
Whisky	**0.960**	**0.960**

Fig. 3 Parity plot of the predictions (pred) for $\ln\gamma_{ij}^{\infty}$ with the proposed whisky method over the corresponding experimental values (exp) and comparison to UNIFAC and data-driven MCM. Coefficients of determination R^2 (higher is better, 1 implies perfect correlation) are given, both including and excluding the worst eight UNIFAC outliers (OL).

the whisky method with the data-driven MCM for predicting all available experimental data points. Again, we observe a significant improvement with the proposed whisky method.

In conclusion, we introduce a novel approach to hybridize physical and data-driven prediction methods for physicochemical properties. In this paper, we focused on predicting activity coefficients at infinite dilution, but the approach can directly be transferred to other properties. The proposed method is termed whisky, reflecting its similarities with the manufacturing of whisky as it combines model distillation with maturation. As a Bayesian approach, it incorporates physical knowledge in the form of a prior belief, and allows to combine it with empirical data evidence in a theoretically well-motivated and convenient way. The proposed method outperforms all considered baselines in predicting activity coefficients at infinite dilution in binary mixtures: the physical gold standard UNIFAC,[10,20] the purely data-driven MCM from our previous work,[13] and two established machine learning ensemble methods, bagging and boosting. We further show that the whisky method is more robust to outliers in the base methods and has a broader applicability than the hybrid baselines. We demonstrate that probabilistic machine learning is perfectly suited for incorporating physical knowledge (that is often abundant in thermodynamics) in powerful data-driven models. We emphasize the generic nature of the proposed whisky approach that opens perspectives to a new generation of hybrid prediction methods for physicochemical properties beyond purely data-driven or purely physical approaches. The transfer to further mixture properties and other physical and data-driven base methods is straightforward. We expect additional improvements if explicit physical information is incorporated and exciting insights by elucidating relations between the learned component features and physical component descriptors.

F. J. greatly acknowledges financial support from the German Academic Exchange Service (DAAD). This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0021. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Furthermore, this work was supported by the National Science Foundation under Grants 1928718, 2003237, and by Qualcomm.

Conflicts of interest

There are no conflicts to declare.

Notes and references

1 A. Fredenslund, *Fluid Phase Equilib.*, 1989, **52**, 135–150.
2 L. Constantinou and R. Gani, *AIChE J.*, 1994, **40**, 1697–1710.
3 Y. Nannoolal, J. Rarey, D. Ramjugernath and W. Cordes, *Fluid Phase Equilib.*, 2004, **226**, 45–63.
4 R. A. Gardas and J. A. P. Coutinho, *AIChE J.*, 2009, **55**, 1274–1290.
5 F. Jirasek, J. Burger and H. Hasse, *Ind. Eng. Chem. Res.*, 2018, **57**, 7310–7313.
6 F. Jirasek, J. Burger and H. Hasse, *Ind. Eng. Chem. Res.*, 2019, **58**, 9155–9165.
7 F. Jirasek, J. Burger and H. Hasse, *AIChE J.*, 2020, **66**, e16826.
8 A. Fredenslund, R. L. Jones and J. M. Prausnitz, *AIChE J.*, 1975, **21**, 1086–1099.
9 A. Fredenslund, J. Gmehling and P. Rasmussen, *Vapor-Liquid Equilibria using UNIFAC, a Group-Contribution Method*, Elsevier, Amsterdam, The Netherlands, 1977.
10 U. Weidlich and J. Gmehling, *Ind. Eng. Chem. Res.*, 1987, **26**, 1372–1381.
11 A. Klamt, *J. Phys. Chem.*, 1995, **99**, 2224–2235.
12 A. Klamt, F. Eckert and W. Arlt, *Annu. Rev. Chem. Biomol. Eng.*, 2010, **1**, 101–122.
13 F. Jirasek, R. A. S. Alves, J. Damay, R. A. Vandermeulen, R. Bamler, M. Bortz, S. Mandt, M. Kloft and H. Hasse, *J. Phys. Chem. Lett.*, 2020, **11**, 981–985.
14 U. Onken, J. Rarey-Nies and J. Gmehling, *Int. J. Thermophys.*, 1989, **10**, 739–747.
15 E. J. Candès and B. Recht, *Found. Comput. Math.*, 2009, **9**, 717–772.
16 R. Mazumder, T. Hastie and R. Tibshirani, *J. Mach. Learn. Res.*, 2010, **11**, 2287–2322.
17 K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.
18 D. M. Blei, A. Kucukelbir and J. D. McAuliffe, *J. Am. Stat. Assoc.*, 2017, **112**, 859–877.
19 A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman and D. M. Blei, *J. Mach. Learn. Res.*, 2017, **18**, 1–45.
20 D. Constantinescu and J. Gmehling, *J. Chem. Eng. Data*, 2016, **61**, 2738–2748.
21 G. Hinton, O. Vinyals and J. Dean, 2015, arXiv:1503.02531, preprint.
22 B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li and A. Riddell, *J. Stat. Software*, 2017, **76**, 1–32.
23 L. Breiman, *Mach. Learn.*, 1996, **24**, 123–140.
24 R. E. Schapire, *Mach. Learn.*, 1990, **5**, 197–227.
25 T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2001.

