Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

ESI for Prediction of Parameters of Group Contribution Models of Mixtures by Matrix Completion

Fabian Jirasek,[*,†] Nicolas Hayer,[†] Rima Abbas,[‡] Bastian Schmid,[‡] and Hans Hasse[†]

[†] *Laboratory of Engineering Thermodynamics (LTD), TU Kaiserslautern, Erwin-Schrödinger-Str. 44, 67663 Kaiserslautern, Germany*

[‡] *DDBST GmbH, Marie-Curie-Str. 10, 26129 Oldenburg, Germany*

E-mail: fabian.jirasek@mv.uni-kl.de

UNIFAC Model

UNIFAC Equations

For predicting the logarithmic activity coefficient $\ln \gamma_i$ of a component i in a mixture, the UNIFAC model considers the sum of an entropic contribution, called combinatorial part, $\ln \gamma_i^{\mathrm{C}}$ and an energetic contribution, called residual part, $\ln \gamma_i^{\mathrm{R}}$:[1]

$$\ln \gamma_i = \ln \gamma_i^{\mathrm{C}} + \ln \gamma_i^{\mathrm{R}} \tag{S1}$$

The combinatorial part $\ln \gamma_i^{\mathrm{C}}$ is thereby calculated by:

$$\ln \gamma_i^{\mathrm{C}} = 1 - V_i + \ln V_i - \frac{z}{2} q_i \left(1 - \frac{V_i}{F_i} + \ln \frac{V_i}{F_i}\right) \tag{S2}$$

with

$$V_i = \frac{r_i}{\sum_j r_j x_j} \tag{S3}$$

$$F_i = \frac{q_i}{\sum_j q_j x_j} \tag{S4}$$

where r_i and q_i are the relative Van der Waals volume and surface area of component i, respectively, x_i is the mole fraction of component i in the mixture, and z is the coordination number, which is set to $z = 10$ in basically all cases and was also used here. Eqs. S1 - S4 are identical to the equations used in the UNIQUAC model;[2] the difference between UNIQUAC and UNIFAC is that UNIQUAC is based on *component-specific* parameters, whereas they are derived from *group-specific* parameters in UNIFAC. Specifically, in UNIFAC, the relative Van der Waals volume r_i and surface area q_i of the component i are calculated from the group volume and group surface parameters, R_k and Q_k, respectively, which are tabulated for multiple structural groups k,[3–8] as follows:

$$r_i = \sum_k \nu_k^{(i)} R_k \tag{S5}$$

$$q_i = \sum_k \nu_k^{(i)} Q_k \tag{S6}$$

where $\nu_k^{(i)}$ denotes the frequency of group k in one molecule of component i.

The residual part $\ln \gamma_i^{\mathrm{R}}$ of UNIFAC is calculated by:

$$\ln \gamma_i^{\mathrm{R}} = \sum_k \nu_k^{(i)} \left(\ln \Gamma_k - \ln \Gamma_k^{(i)} \right) \tag{S7}$$

where Γ_k is the group activity coefficient of group k in the mixture and $\Gamma_k^{(i)}$ is the group activity coefficient of group k in the pure component i. Both Γ_k and $\Gamma_k^{(i)}$ are calculated similar to the residual part in the UNIQUAC model by:

$$\ln \Gamma_k = Q_k \left(1 - \ln \left(\sum_m \Theta_m \Psi_{mk} \right) - \sum_m \frac{\Theta_m \Psi_{km}}{\sum_n \Theta_n \Psi_{nm}} \right) \tag{S8}$$

where Θ_m is the surface fraction of group m in the mixture:

$$\Theta_m = \frac{Q_m X_m}{\sum_n Q_n X_n} \tag{S9}$$

and X_m is the group mole fraction of group m, which is related to the mole fractions x_j of components j:

$$X_m = \frac{\sum_j \nu_m^{(j)} x_j}{\sum_j \sum_n \nu_n^{(j)} x_j} \tag{S10}$$

The parameters Ψ_{nm} and Ψ_{mn} in Eq. S8 contain the group-interaction parameters of UNIFAC, A_{nm} and A_{mn}, between the groups m and n:

$$\Psi_{nm} = \exp\left(-\frac{A_{nm}}{T}\right); \qquad \Psi_{mn} = \exp\left(-\frac{A_{mn}}{T}\right) \tag{S11}$$

UNIFAC Group-Interaction Parameters

In Figure S1, the current availability of group-interaction parameters of the public UNIFAC[8] and the commercial UNIFAC-TUC[9] is indicated.

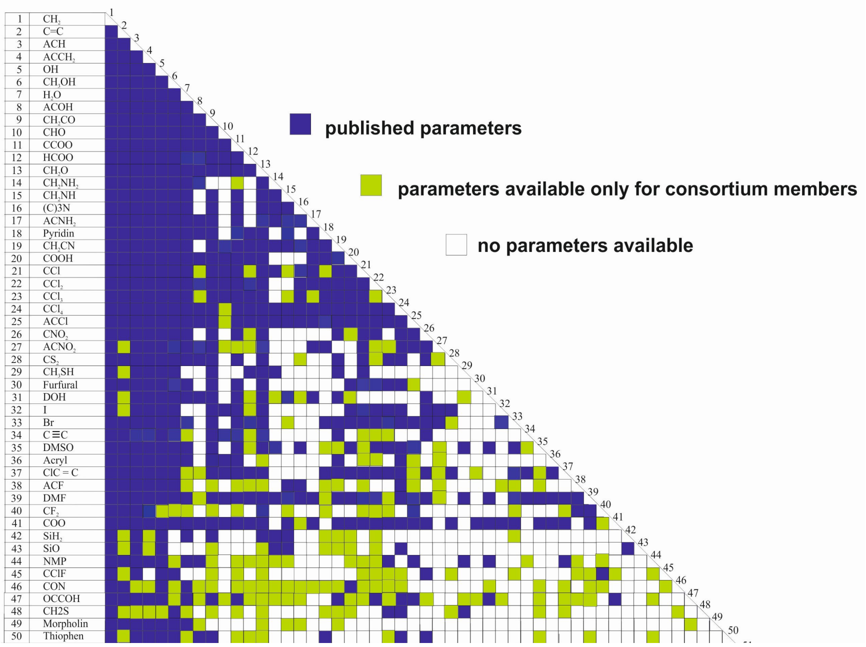

Figure S1: Matrix representing the availability of group-interaction parameters of the public UNIFAC[8] (blue) and the commercial UNIFAC-TUC[9] (green) up to main group 50. White cells: no parameters available.

Model Details

Bayesian Matrix Completion

The model of the present work is similar to our recently introduced approach,[10] in which we have combined a matrix completion method (MCM) from machine learning with the UNIQUAC model.[2,11] In contrast to our previous work,[10] the group-interaction parameters among *structural groups* G and G' (and not components), specifically between main groups of UNIFAC, are predicted here. Figure S2 shows an overview of the proposed UNIFAC-MCM model as well as of the training and evaluation procedure.

We have trained the model on pseudo-data for logarithmic activity coefficients $\ln \gamma_{GG'}$ in hypothetical binary mixtures of groups, G and G' , which we have generated with the UNIFAC model using the current public parameterization[8] as described in the manuscript. Note that although these pseudo-data were generated based on an *inconsistent* set of group-interaction parameters, the pseudo-data themselves are not inconsistent, because, as we

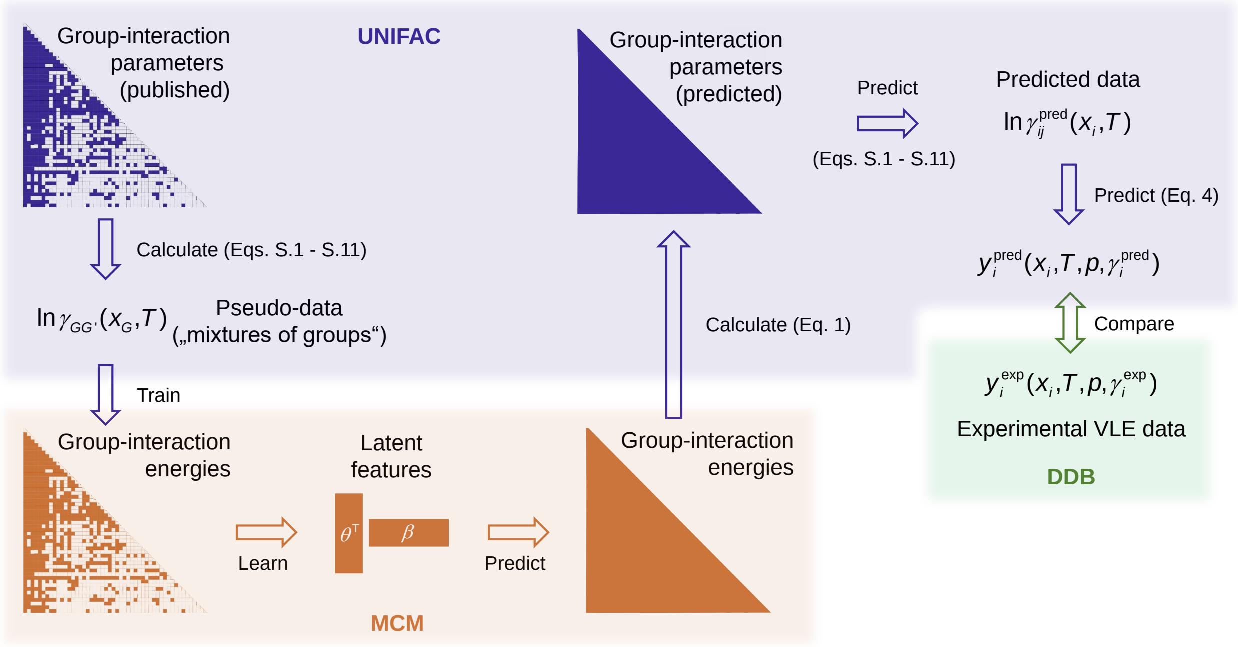
Group-interaction parameters (published)
UNIFAC
Group-interaction parameters (predicted)
Predict
(Eqs. S.1 - S.11)
Predicted data
$\ln \gamma_{ij}^{\mathrm{pred}}(x_i, T)$
Predict (Eq. 4)
Calculate (Eqs. S.1 - S.11)
$y_i^{\mathrm{pred}}(x_i, T, p, \gamma_i^{\mathrm{pred}})$
$\ln \gamma_{GG'}(x_G, T)$
Pseudo-data („mixtures of groups“)
Calculate (Eq. 1)
Compare
Train
$y_i^{\mathrm{exp}}(x_i, T, p, \gamma_i^{\mathrm{exp}})$
Group-interaction energies
Latent features
Group-interaction energies
Experimental VLE data
DDB
θ^{T}
β
Learn
Predict
MCM

Figure S2: Scheme representing the training and evaluation of UNIFAC-MCM. Besides based on the vapor-phase composition y, the results were also evaluated based on deviations in the temperature T and the pressure p from the experimental vapor-liquid equilibrium (VLE) data from the Dortmund Data Bank (DDB).

describe in the manuscript, very similar activity coefficients can be obtained by different combinations of group-interaction parameters. This makes the values of the group-interaction parameters less informative, whereas the generated pseudo-data contain the structure that is recovered by the MCM during the training.

We have thereby employed a Bayesian approach to matrix completion, which consists of multiple steps as described in the following.

First, we have specified a generative probabilistic model for $\ln \gamma_{GG'}$ as a nonlinear function f of the groups G and G', the temperature T, and the mole fraction x_G of group G in the hypothetical mixture. This function is basically defined by the UNIFAC equations, cf. Eqs S1 - S11, the correlation of the group-interaction parameters $A_{GG'}$ and $A_{G'G}$ via the group-interaction energies $U_{GG'}$, U_{GG}, and $U_{G'G'}$, cf. Eqs. (1) - (2) in the manuscript, and an embedded matrix factorization for the unlike group-interaction energies $U_{GG'}$ between the groups, cf. Eq. (3) in the manuscript. The function furthermore considers the following parameters:

- group-specific parameters considered in the UNIFAC model, specifically the group volume parameters R_G and $R_{G'}$ and the group surface parameters Q_G and $Q_{G'}$, which were adopted from the latest public parameter table of UNIFAC, cf. Table S1;
- initially unknown (latent) feature vectors θ_G, $\theta_{G'}$, β_G, and $\beta_{G'}$ of the groups, which are used for modeling the unlike group-interaction energies $U_{GG'}$ between the groups, as well as the like group-specific group-interaction energies U_{GG} and $U_{G'G'}$.

The length K of the feature vectors, which controls the number of features that are considered for each group, is in principle a hyperparameter of the model, which can be adjusted during model selection. However, in this work, we have not carried out a comprehensive hyperparameter screening, but have simply adopted the hyperparameters from our previous work,[10] which includes setting $K = 3$ here.

θ_G, $\theta_{G'}$, β_G, $\beta_{G'}$, U_{GG}, and $U_{G'G'}$ constitute the parameters of the model that were inferred during the training. For the training, the generative model defines a probability distribution over all used pseudo-data for $\ln \gamma_{GG'}$ by specifying a stochastic process for generating hypothetical data for $\ln \gamma_{GG'}$ conditioned on θ_G, $\theta_{G'}$, β_G, $\beta_{G'}$, U_{GG}, and $U_{G'G'}$, which are initially unknown, R_G, $R_{G'}$, Q_G, and $Q_{G'}$, which were adopted from Refs. ,[1,12–14] and the temperature and the mole fraction of G in the mixture. The generative process therefore draws θ_G, $\theta_{G'}$, β_G, $\beta_{G'}$, U_{GG}, and $U_{G'G'}$ from a normal so-called prior distribution with zero mean and a standard deviation of one. The type of distribution used as prior as well as the mean and the standard deviation are also hyperparameters of the model, but were, as K, also set as in our previous work.[10] Then, the generative process models the probability of the training data $\ln \gamma_{GG'}$ as a Cauchy so-called likelihood distribution with scale $\lambda = 0.2$ centered around the outcome of the function f with the θ_G, $\theta_{G'}$, β_G, $\beta_{G'}$, U_{GG}, and $U_{G'G'}$ drawn from the prior and the fixed parameters and conditions. Again, the type of distribution used as likelihood as well as the scale are hyperparameters, which were set as in our previous work.[10] We can write the likelihood as follows:

$$\ln \gamma_{GG'}(T, x_G) = \text{Cauchy}(f(T, x_G, R_G, R_{G'}, Q_G, Q_{G'}, \theta_G, \theta_{G'}, \beta_G, \beta_{G'}, U_{GG}, U_{G'G'}), \lambda) + \epsilon_{GG'} \tag{S12}$$

where the function f includes the UNIFAC equations, Eqs S1 - S11, as well as Eqs. (1) - (3) in the manuscript and $\epsilon_{GG'}$ captures the deviations between the model results and the pseudo-data $\ln \gamma_{GG'}(T, x_G)$ for training the model. In the next step, the parameters that were to be learned, i.e., θ_G, $\theta_{G'}$, β_G, $\beta_{G'}$, U_{GG}, and $U_{G'G'}$, were concurrently inferred for all groups G based on the set of training data, which requires the inversion of the generative model. Since full Bayesian inference is intractable except for very simple cases, we resorted to Gaussian mean-field variational inference[15–17] for this purpose. Simply put, we can understand this procedure as a comparison of the generated hypothetical $\ln \gamma_{GG'}$ to the training data, i.e., the pseudo-data for $\ln \gamma_{GG'}$ as obtained with UNIFAC using the latest public parameterization, to subsequently adjust the initially unknown parameters. This results in the so-called posterior, which constitutes a probability distribution for all inferred parameters.

Finally, we used the means of the approximated posterior distributions over θ_G, $\theta_{G'}$, β_G, $\beta_{G'}$, U_{GG}, and $U_{G'G'}$ to predict the group-interaction parameters $A_{GG'}$ and $A_{G'G}$ for all possible combinations of groups according to Eqs. (1) - (3) in the manuscript. The predicted $A_{GG'}$ and $A_{G'G}$ were, in turn, used for predicting the activity coefficients of *components* $\ln \gamma_i$ in binary mixtures with Eqs. S1 - S11, which were finally used for predicting vapor-liquid equilibrium (VLE) phase diagrams. Our approach thereby basically changes Eq. S11 to:

$$\Psi_{nm} = \exp\left(-\frac{\theta_n \cdot \beta_m + \theta_m \cdot \beta_n - U_{mm}}{T}\right); \qquad \Psi_{mn} = \exp\left(-\frac{\theta_n \cdot \beta_m + \theta_m \cdot \beta_n - U_{nn}}{T}\right) \tag{S13}$$

The predicted VLE phase diagrams were compared to experimental data from the Dortmund Data Bank (DDB) [18] as discussed in the manuscript.

For performing the task of Bayesian inference, the Stan framework[19] was used.

Scope of UNIFAC-MCM

Since UNIFAC-MCM yields a complete set of group-interaction parameters for the first 50 main groups of UNIFAC, the approach allows modeling any binary and multi-component mixture whose components can be built from these groups. The scope of the new approach is thereby much larger than we can demonstrate here, simply due to missing experimental data for a more comprehensive assessment. This is also indicated in Figure S3, which shows the number of binary systems from our data set for which VLE data are available in the DDB[18] and which contain the respective combination of UNIFAC main groups.

While there are several group-interaction parameters that are required for modeling a large number of binary systems for our data set (dark-colored cells in Figure S.3), approximately 80% of all possible main group combinations are not represented in the data set (white cells in Figure S.3). The lack of experimental data inevitably prevents the parameterization of UNIFAC, both in its public and commercial versions, in the ordinary way, as only those parameters can be fitted for which respective training data are available. With UNIFAC-MCM, on the other hand, this problem is solved; UNIFAC-MCM yields predictions also for the 80% of group-interaction parameters from Figure S.3 for which classical UNIFAC versions cannot achieve this based on the studied data set.

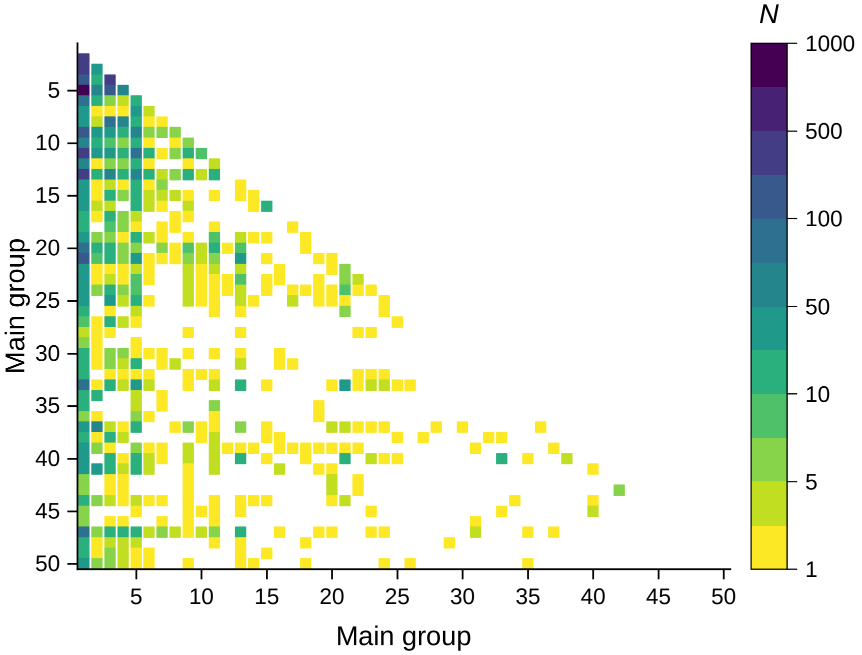

Figure S3: Heat map showing the number N of binary systems for which VLE data are available in the DDB and which contain the respective combination of UNIFAC main groups. White cells indicate that no VLE data are available for the given combination of groups.

Additional Results

In Figure S4, the results of UNIFAC-MCM for the prediction of the VLE data are plotted in histogram representations, which show the number of binary systems that are predicted with a defined relative deviation from the experimental mole fraction of the low-boiling component in the vapor phase Δy. In the left panel, the results of UNIFAC-MCM on the complete horizon are shown. In the middle panel, the results for those systems from the complete horizon are shown that can not be modeled by the public UNIFAC version, but by the commercial UNIFAC-TUC; here, the UNIFAC-MCM predictions are compared to those of UNIFAC-TUC. In the right panel, the results for those systems from the complete horizon are shown that can only be modeled by UNIFAC-MCM.

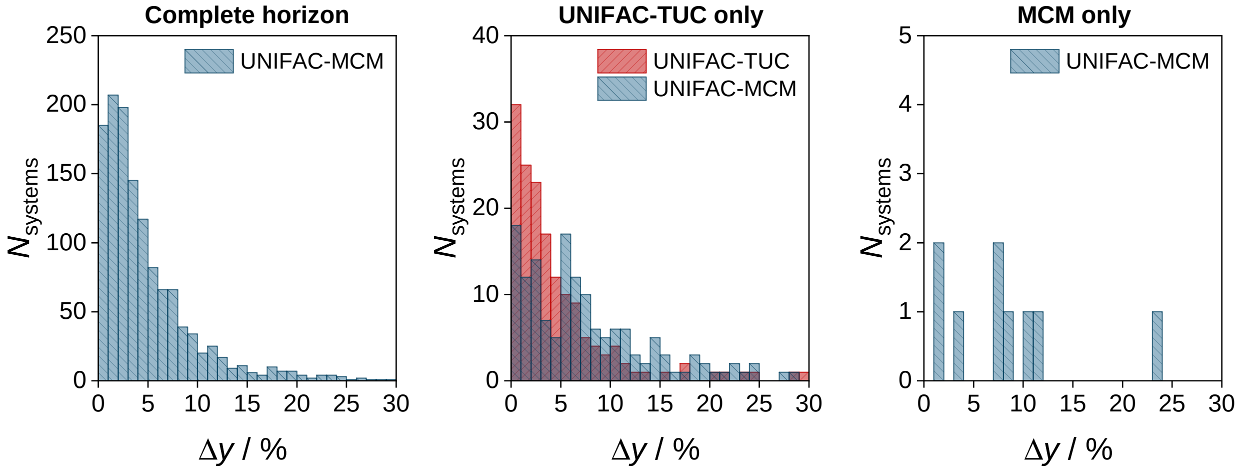

Figure S4: Histogram representations of number of systems that are predicted by UNIFAC-MCM with a defined relative deviation from the experimental vapor mole fractions of the low-boiling components Δy. Left: for the complete horizon (2,246 systems). Middle: for those systems that can not be predicted with public UNIFAC (169 systems). Right: for those systems that can not be predicted with UNIFAC-TUC (9 systems).

Table S1: UNIFAC main groups G considered in the present work and the respective group volume and group surface parameters, R_G and Q_G, used. Some main groups include multiple sub groups, such that R_G and Q_G could have been chosen differently, whereby, however, no large impact is expected; in such cases, usually one of the 'intermediate' sub groups was chosen randomly here (e.g., 'CH2').

G	R_G	Q_G	G	R_G	Q_G
1	0.6744	0.54	26	1.7818	1.56
2	1.1167	0.867	27	1.4199	1.104
3	0.5313	0.4	28	2.057	1.65
4	1.0396	0.66	29	1.651	1.368
5	1	1.2	30	3.168	2.484
6	1.4311	1.432	31	2.4088	2.248
7	0.92	1.4	32	1.264	0.992
8	0.8952	0.68	33	0.9492	0.832
9	1.4457	1.18	34	1.0613	0.784
10	0.998	0.948	35	2.8266	2.472
11	1.6764	1.42	36	2.3144	2.052
12	1.242	1.188	37	0.791	0.724
13	0.9183	0.78	38	0.6948	0.524
14	1.3692	1.236	39	3.0856	2.736
15	1.207	0.936	40	1.0105	0.92
16	0.9597	0.632	41	1.38	1.2
17	1.06	0.816	42	1.4443	1.0063
18	2.8332	1.833	43	1.303	0.7639
19	1.6434	1.416	44	3.981	3.2
20	1.3013	1.224	45	2.2287	1.916
21	1.238	0.952	46	1.9637	1.488
22	2.0606	1.684	47	1.8952	1.592
23	2.6401	2.184	48	1.3863	1.06
24	3.39	2.91	49	3.474	2.796
25	1.1562	0.844	50	2.6908	1.86

References

(1) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AIChE Journal* **1975**, *21*, 1086–1099.

(2) 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* **1975**, *21*, 116–128.

(3) 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* **1979**, *18*, 714–722.

(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* **1982**, *21*, 118–127.

(5) 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* **1983**, *22*, 676–678.

(6) Tiegs, D.; Rasmussen, P.; Gmehling, J.; Fredenslund, A. Vapor-liquid equilibria by UNIFAC group contribution. 4. Revision and extension. *Industrial & Engineering Chemistry Research* **1987**, *26*, 159–161.

(7) 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* **1991**, *30*, 2352–2355.

(8) Wittig, R.; Lohmann, J.; Gmehling, J. Vapor- liquid equilibria by UNIFAC group contribution. 6. Revision and extension. *Industrial & Engineering Chemistry Research* **2003**, *42*, 183–188.

(9) The UNIFAC Consortium. 2022; `http://www.unifac.org`.

(10) 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* **2022**, *13*, 4854–4862.

(11) Maurer, G.; Prausnitz, J. On the Derivation and Extension of the UNIQUAC Equation. *Fluid Phase Equilibria* **1978**, *2*, 91–99.

(12) Magnussen, T.; Rasmussen, P.; Fredenslund, A. UNIFAC parameter table for prediction of liquid-liquid equilibriums. *Industrial & Engineering Chemistry Process Design and Development* **1981**, *20*, 331–339.

(13) Wienke, G.; Gmehling, J. Prediction of octanol-water partition coefficients, Henry coefficients and water solubilities using UNIFAC. *Toxicological & Environmental Chemistry* **1998**, *65*, 57–86.

(14) Yan, W.; Topphoff, M.; Rose, C.; Gmehling, J. Prediction of vapor–liquid equilibria in mixed-solvent electrolyte systems using the group contribution concept. *Fluid Phase Equilibria* **1999**, *162*, 97–113.

(15) Zhang, C.; Bütepage, J.; Kjellström, H.; Mandt, S. Advances in Variational Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2018**, *41*, 2008–2026.

(16) Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. Variational inference: A Review for Statisticians. *Journal of the American Statistical Association* **2017**, *112*, 859–877.

(17) Kucukelbir, A.; Tran, D.; Ranganath, R.; Gelman, A.; Blei, D. M. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research* **2017**, 1–45.

(18) Dortmund Data Bank (DDB). 2022; `http://www.ddbst.com`.

(19) 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* **2017**, 1–32.