Supporting Information for Prediction of Activity Coefficients by Similarity-Based Imputation using Quantum-Chemical Descriptors

Nicolas Hayer, Thomas Specht, Justus Arweiler, Dominik Gond, Hans Hasse, and Fabian Jirasek*

Laboratory of Engineering Thermodynamics, RPTU Kaiserslautern,
Erwin-Schrödinger-Str. 44, 67663 Kaiserslautern, Germany

E-mail: fabian.jirasek@rptu.de

Outliers of Modified UNIFAC (Dortmund)

The predictions of modified UNIFAC (Do)[1] include eight extreme outliers, cf. Table S.1, which can be attributed to poorly fitted group-interaction parameters. Specifically, all of the relevant solutes contain main group 42 ("CY-CH2"), while all of the relevant solvents contain main group 18 ("PYRIDINE").

The few observed outliers would drastically increase the mean absolute error (MAE) and mean squared error (MSE) and thus lead to a false impression of the predictive performance of modified UNIFAC (Dortmund); they were therefore removed from this study. By removing the eight listed outliers, the MAE decreases from 0.6477 to 0.3340.

Table S.1: Binary systems with available experimental $\ln \gamma_{ij}^{\infty}$ at 298.15 K where modified UNIFAC (Dortmund) predictions deviate significantly from the experimental data, likely due to inaccurate group-interaction parameters. Identifiers (DDB no.) are the original ones from the DDB.[2]

Solute i		Solvent j	
DDB no.	Name	DDB no.	Name
50	Cyclohexane	19	2-Methylpyridine
50	Cyclohexane	144	Pyridine
50	Cyclohexane	433	Quinoline
51	Cyclopentane	433	Quinoline
52	Cyclohexene	144	Pyridine
159	Tetrahydrofuran	144	Pyridine
401	Ethylcyclohexane	19	2-Methylpyridine
401	Ethylcyclohexane	144	Pyridine

Results of the Hyperparameter Variations

The predictive performance of each SBM variant is shown in Fig. 3 in the manuscript, focusing on the MAE. By displaying the MSE in a similar way, Fig. S.1 supports the choice of the final model, represented by the orange dots.

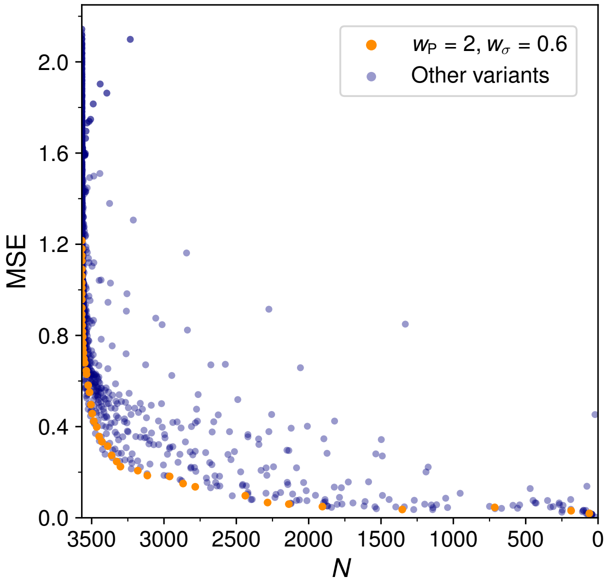

Figure S.1: Mean squared error (MSE) of the predicted $\ln \gamma_{ij}^{\infty}$ over the number of predictable experimental data points N for all tested SBM variants. The results of the best-performing SBM (as specified with the weights w) are highlighted in orange.

In Fig. S.2, a detailed analysis of the results is shown, focusing on the influence of different hyperparameter choices. From this analysis, we derive the following heuristics:

- A stronger weighting of the polar regions in the σ-profiles ($w_{\mathrm{P}} = 2$) enhances both objectives, i.e., the predictive performance and the scope, especially near the Pareto knee, cf. Fig. S.2a.

- Relying solely on surface area similarity ($w_\sigma = 0$) results in poor predictions, cf. Fig. S.2b.

- Focusing solely on charge distribution similarity ($w_\sigma = 1$) achieves comparatively good results at lower thresholds, where both the SBM scope and the MAE are generally large. However, increasing the threshold yields only small improvements in predictive accuracy, cf. Fig. S.2c. In contrast, many model variants incorporating surface area similarity ($w_\sigma < 1$) perform significantly better here.

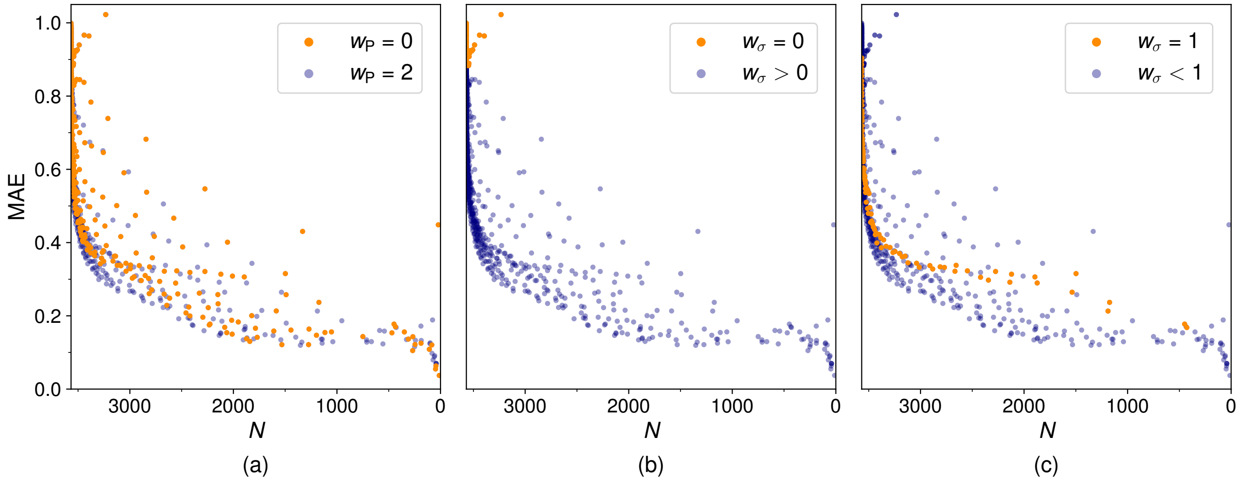

Figure S.2: Mean absolute error (MAE) of the predicted $\ln \gamma_{ij}^{\infty}$ over the number of predictable experimental data points N for all tested SBM variants. In each panel, a subset of SBM variants is highlighted in orange: (a) equal weighting of the polar and non-polar regions in the σ-profiles ($w_\mathrm{P} = 0$), (b) using only the surface area similarity ($w_\sigma = 0$), (c) using only the charge distribution similarity ($w_\sigma = 1$).

Scope of the Proposed Similarity-Based Method

The performance of the final SBM, characterized through $w_\sigma = 0.6$, and $w_\mathrm{P} = 2$, can be influenced by varying the threshold ξ. Fig. S.3 illustrates the resulting trade-off between scope and predictive accuracy, focusing not only on the scope in terms of the number of predictable data points from the database (Fig. S.3a), but also on the number of predictable data points from all possible solute-solvent combinations (Fig. S.3b).

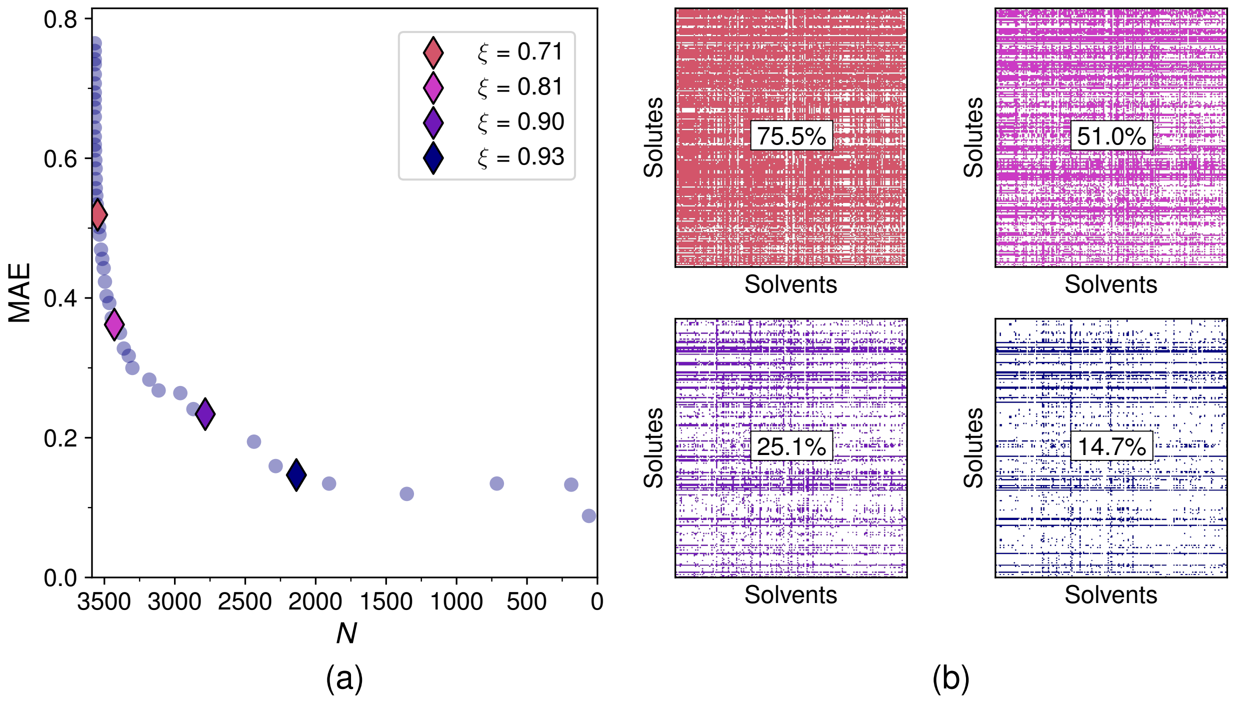

Figure S.3: Influence of the threshold ξ on the predictive performance and the scope of the SBM. (a) Mean absolute error (MAE) of the SBM for the prediction of $\ln \gamma_{ij}^\infty$. N represents the number of predictable experimental data points. Four distinct thresholds are highlighted. (b) Matrices representing all predictable solute-solvent combinations for the four ξ values highlighted in (a).

In general, a large ξ results in a small scope and vice versa. Thereby, the percentage of predictable data points from the experimental database is always higher than the scope concerning the entire solute-solvent matrix. For example, the selected threshold of $\xi = 0.93$ yields an SBM that can predict 59.9% of the experimental database but can only populate 14.7% of the solute-solvent matrix. The SBM is limited in extrapolating into the sparse region of the matrix, which is not surprising since it relies on experimental data points of similar mixtures. In comparison, modified UNIFAC (Do)[1] is capable of predicting 83.7% of the experimentally studied mixtures and 69.9% of the entire matrix, COSMO-SAC-dsp[3] achieves 89.7% and 85.4%, respectively, whereas COSMO-SAC[4] can calculate $\ln \gamma_{ij}^{\infty}$ for all considered binary mixtures.

By reducing ξ, the scope of the SBM could be extended so that almost any solute-solvent combination can be predicted. However, this would lead to a significant loss of predictive accuracy. Therefore, we continue to use the SBM with $\xi = 0.93$ as its prediction accuracy is within the range of experimental uncertainty, and we accept the limited scope. The unprecedented performance of this SBM makes it a precious tool in chemical process engineering, even if it is not applicable in all cases.

Parity Plots of the Considered Methods

In Fig. S.4, parity plots compare the predicted $\ln \gamma_{ij}^{\infty}$ values for the SBM with $\xi = 0.93$, modified UNIFAC (Dortmund), COSMO-SAC, and COSMO-SAC-dsp with the experimental data from the DDB[2]. The plots include only mixtures predictable by all methods, ensuring consistency with the data points used in the histograms in Fig. 3 of the manuscript.

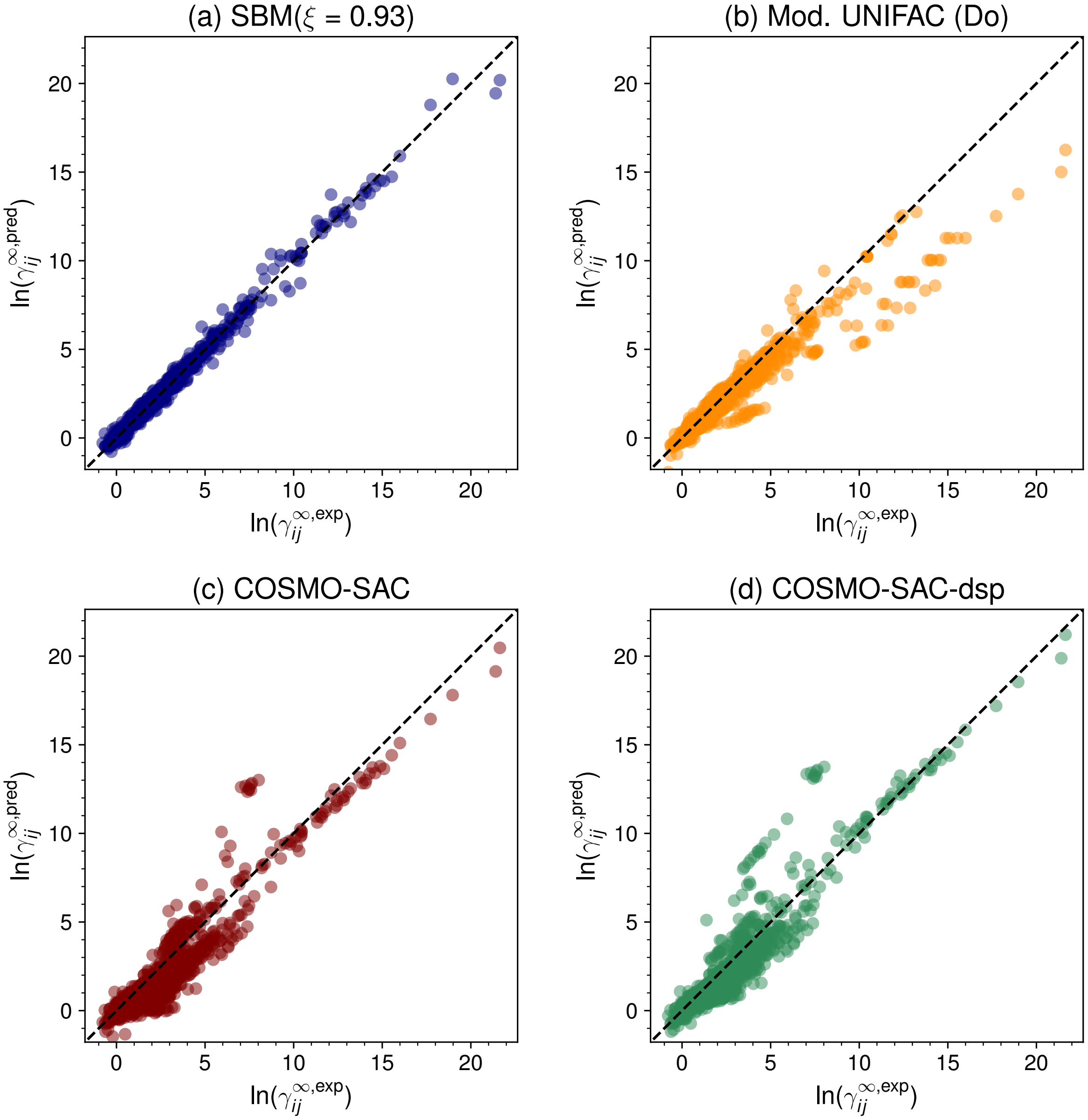

Figure S.4: Parity plots of the predictions (pred) for $\ln \gamma_{ij}^{\infty}$ with the SBM with $\xi = 0.93$, modified UNIFAC (Dortmund), COSMO-SAC, and COSMO-SAC-dsp over experimental data (exp) from the DDB, considering only the 1,748 mixtures that all methods can describe.

Fig. S.4 shows that modified UNIFAC (Dortmund) performs less accurately for mixtures with highly non-ideal behavior, characterized by large $\ln \gamma_{ij}^{\infty}$ values. In contrast, COSMO-SAC and COSMO-SAC-dsp predict these mixtures more accurately but exhibit larger overall deviations, as indicated by the wider scatter of points. The SBM demonstrates consistently high accuracy over the entire range of $\ln \gamma_{ij}^{\infty}$ values, with points closely aligned along the diagonal, underscoring its robustness even for non-ideal mixtures.

Case Studies of Similar Components

The calculated similarity scores S_{ij} between two components are not only the basis of the proposed SBM for the prediction of activity coefficients at infinite dilution. They also offer the possibility to identify the most similar components for a target component, exemplified in Tables S.2 and S.3. Here, the S_{ij} of the final SBM, obtained through the grid search, are used.

Table S.2: Lists of top 10 components among the solutes most similar to ethanol or n-butane.

Solutes Similar to Ethanol		Solutes Similar to n-Butane	
Solute	S_{ij}	Solute	S_{ij}
1-Propanol	0.909	Cyclopentane	0.962
2-Propanol	0.869	2-Methylpropane	0.961
Methanol	0.850	Pentane	0.930
1-Butanol	0.835	2-Methylbutane	0.929
N-Methylformamide	0.829	Propane	0.917
2-Butanol	0.808	Methylcyclopentane	0.910
1-Pentanol	0.806	Cyclohexane	0.894
tert-Butanol	0.801	3-Methylpentane	0.890
2-Methyl-1-propanol	0.776	2-Methylpentane	0.886
Cyclohexanol	0.770	2,3-Dimethylbutane	0.881

Table S.3: Lists of top 10 components among the solvents most similar to water or chlorobenzene.

Solvents Similar to Water		Solvents Similar to Chlorobenzene	
Solvent	S_{ij}	Solvent	S_{ij}
Methanol	0.716	Bromobenzene	0.977
Formamide	0.702	Iodobenzene	0.948
Ethanol	0.636	Fluorobenzene	0.921
1,2-Ethanediol	0.623	1-Chloronaphthalene	0.869
1-Propanol	0.599	1-Bromonaphthalene	0.861
2-Propanol	0.592	1,1-Dichloroethane	0.770
1,3-Propanediol	0.591	Toluene	0.746
1-Butanol	0.572	Methoxybenzene	0.743
1,4-Butanediol	0.572	Indene	0.742
1,2-Propanediol	0.569	Diiodomethane	0.740

Not surprisingly, many alkanes are similar to n-butane, and many halogenobenzenes are similar to chlorobenzene. In contrast, water has no similar components, as it is unique in being an extremely polar and rather small molecule.

Literature Cited

(1) Constantinescu, D.; Gmehling, J. Further Development of Modified UNIFAC (Dortmund): Revision and Extension 6. *Journal of Chemical & Engineering Data* **2016**, *61*, 2738–2748.

(2) Dortmund Data Bank. 2023; `www.ddbst.com`.

(3) Hsieh, C.-M.; Lin, S.-T.; Vrabec, J. Considering the dispersive interactions in the COSMO-SAC model for more accurate predictions of fluid phase behavior. *Fluid Phase Equilibria* **2014**, *367*, 109–116.

(4) Hsieh, C.-M.; Sandler, S. I.; Lin, S.-T. Improvements of COSMO-SAC for vapor–liquid and liquid–liquid equilibrium predictions. *Fluid Phase Equilibria* **2010**, *297*, 90–97.