Electronic Supplementary Information for Prediction of Pair Interactions in Mixtures by Matrix Completion

Marco Hoffmann, Nicolas Hayer, Maximilian Kohns, Fabian Jirasek,* and Hans Hasse

Laboratory of Engineering Thermodynamics, RPTU Kaiserslautern

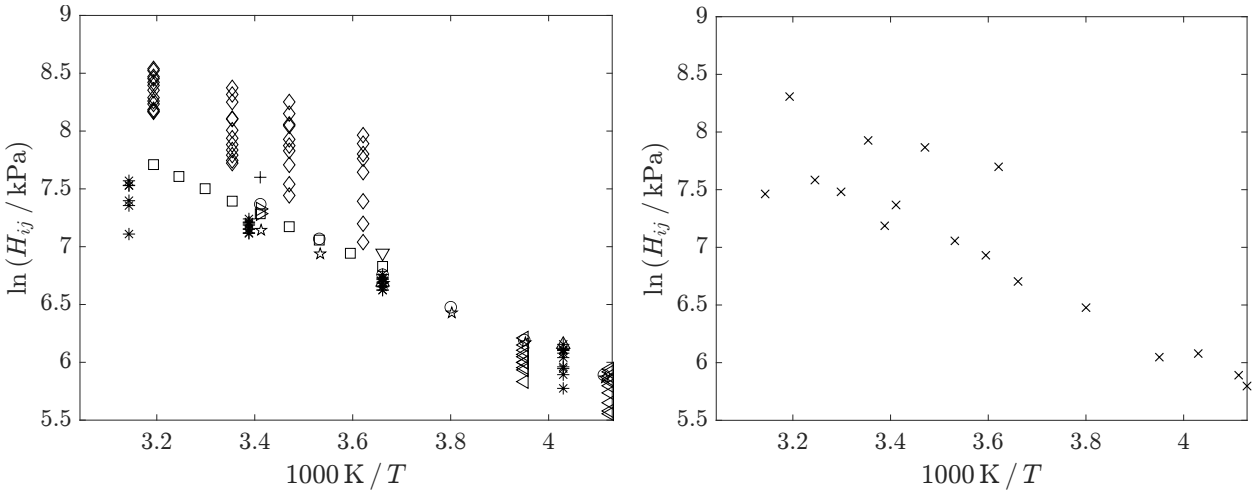
E-mail: fabian.jirasek@rptu.de

Data

In the following, additional information on the pre-processing of the experimental data, the selection of molecular models, and the final assembly of the data basis is provided.

Experimental Data

At the time of access, the Dortmund Data Bank[1] (DDB) provided approximately 65 000 Henry's law constants for 366 solutes and 1106 solvents at various temperatures. In a pre-processing, the data was consolidated as proposed in one of our earlier works:[2] if multiple data points were available at the same temperature, the median of all these data points within a range of ±0.5 K, e.g. 298.15 ± 0.5 K, was taken. This step is illustrated for the binary system acetylene - acetone in Fig. 1.



(a) Original data from literature

(b) Consolidated data

Figure 1: Visualization of the data consolidation for the binary system acetylene - acetone. Left: the original experimental data sets corresponding to the following authors: Bodor et al.[3] (∗); Bodor et al.[4] (△); Eck[5] (+); Hannaert et al.[6] (▷); Hölemann and Hasselmann[7] (◇); Horiuti[8] (□); Kvasenkov and Shleinikov[9] (☆); Otsuka and Takada[10] (▽); Shenderei and Ivanovskii[11] (◁); Usyukin and Shleinikov[12] (○). Right: consolidated experimental data points (×) after pre-processing.

As the binary interaction parameter is assumed to be temperature-independent, only one experimental Henry's law constant is required for fitting. For a large share of the binary systems, the experimental data set contains Henry's law constants at 298.15 K. Thus, if available, data points at this temperature were chosen. This comes in handy for the simulations, as the Henry's law constants for different solutes in the same solvent at the same temperature can be obtained in a single run when Widom's test particle method[13] is used to sample the chemical potential. For the other binary systems, data points at temperatures close to 298.15 K were chosen.

Molecular Models

At the time of access, the MolMod database[14] contained 164 molecular models for 124 different molecules. For the molecules for which more than one molecular model was available, a preliminary study was conducted to find out which model is best suited for the prediction of Henry's law constants. All binary systems of this molecular model for which experimen-

tal data on the Henry's law constant was available were simulated. For this purpose, the molecular model that gave the best predictions (based on the MAE of $\ln(H_{ij}/\mathrm{kPa})$) was selected.

Final Data Basis

Molecular simulations were performed for all binary systems in the intersection of the experimental data and molecular models (as described above). Because Widom's test particle method,[13] which was used for the calculation of H_{ij}, is prone to errors when dealing with large test particles and/or a high-density solvent, the result of every simulation was manually checked for convergence. All questionable and faulty results were subsequently omitted (and with them the respective experimental data point). For this reason, also water dropped out. Furthermore, components for which only one data point was remaining were discarded, since a minimum of two data points per component are required for the leave-one-out analysis that was used for the assessment of the results (see section Matrix Completion Method in the paper). The final data basis includes 213 binary systems consisting of 34 solutes and 15 solvents. The corresponding molecular models are listed in Tab. 1, which also assigns them unique IDs for reference.

Molecular Simulation

In this work, molecular dynamics (MD) simulations were carried out with 1000 solvent molecules. First, 1024 Monte-Carlo relaxation loops were performed for an initial energy minimisation, followed by 10 000 time steps in the NVT ensemble and 50 000 time steps in the NpT ensemble for equilibration. Thereafter, 1 000 000 time steps in the NpT ensemble were carried out for the data production, during which 5000 molecules of each considered solute were inserted to calculate Henry's law constants every 1000 time steps. The center of mass cut-off mode was chosen with a cut-off radius of 15.05 Å. Long-range LJ interactions

were accounted for with the formulations proposed by Lustig.[15] The reaction field method[16] was employed for the longe-range electrostatic interactions. To solve Newton's equation of motion, the Gear predictor-corrector integrator was used. The time step was $\Delta t = 1.92\,\mathrm{fs}$.

To allow for a quick equilibration, the density of the solvent was initialised according to the correlation

$$\rho^{\mathrm{S}}_{j,\mathrm{liq}} / \left(\mathrm{mol\,l^{-1}}\right) = a_2 \cdot (T/\mathrm{K})^2 + a_1 \cdot (T/\mathrm{K}) + a_0 \tag{1}$$

with the parameters given in Tab. 2. The pressure in the NpT steps was set to 105 % of the pressure resulting from the Antoine's equation

$$\log_{10}\left(p^{\mathrm{S}}_j/\mathrm{Pa}\right) = A - \frac{B}{(T/\mathrm{K}) + C} \tag{2}$$

for the respective solvent. The parameters are again listed in Tab. 2. Both sets of parameters of Eq. (1) and (2) reported in Tab. 1 were determined in the present work from a fit to molecular simulation data from the original publications of the molecular models given in Tab. 1.

Table 1: List of the molecular models that were used for the molecular simulations in this work.

Name	**CAS-Number**	**Reference**	**Solute-ID**	**Solvent-ID**
1,1-Difluoroethane	75-37-6	17	11	-
1,1-Dimethylhydrazine	57-14-7	18	-	11
Acetone	67-64-1	19	-	2
Acetonitrile	75-05-8	20	-	1
Acetylene	74-86-2	21	27	-
Ammonia	7664-41-7	22	4	-
Argon	7440-37-1	21	21	-
Benzene	71-43-2	23	-	6
Carbon dioxide	124-38-9	24	13	-
Carbon disulfide	75-15-0	21	3	-
Carbon monoxide	630-08-0	17	20	-
Carbon tetrachloride	56-23-5	23	-	9
Carbon tetrafluoride	75-73-0	21	12	-
Chlorine	7782-50-5	25	22	-
Chlorodifluoromethane	75-45-6	17	5	-
Chlorotrifluoromethane	75-72-9	17	8	-
Cyclohexane	110-82-7	26	-	7
Cyclohexanol	108-93-0	27	-	13
Cyclohexanone	108-94-1	28	-	12
Cyclopropane	75-19-4	29	34	-
Dichlorodifluoromethane	75-71-8	17	6	-
Difluoromethane	75-10-5	17	29	-
Ethane	74-84-0	25	17	-
Ethanol	64-17-5	30	1	3
Ethylene	74-85-1	25	16	-
Ethylene oxide	75-21-8	31	-	4
Fluoromethane	593-53-3	17	33	-
Formic acid	64-18-6	32	-	5
Hydrogen	1333-74-0	33 [1]	26	-

[1]The publication lists several models for hydrogen. Here, model A was used.

Table 1 (cont.): List of the molecular models that were used for the molecular simulations in this work.

Name	**CAS-Number**	**Reference**	**Solute-ID**	**Solvent-ID**
Hydrogen chloride	7647-01-0	34	2	-
Isobutane	75-28-5	26	7	-
Isopropanol	67-63-0	35	-	8
Krypton	7439-90-9	21	23	-
Methane	74-82-8	21	14	-
Methyl chloride	74-87-3	17	10	-
Methyl ether	115-10-6	26	9	-
Methylhydrazine	60-34-4	18	-	15
Neon	7440-01-9	21	30	-
Nitrogen	7727-37-9	25	19	-
Nitrous oxide	10024-97-2	36 [2]	24	-
Octamethylcyclotetrasiloxane	556-67-2	37	-	14
Oxygen	7782-44-7	25	15	-
Propylene	115-07-1	21	18	-
Sulfur dioxide	7446-09-5	26	32	-
Sulfur hexafluoride	2551-62-4	21	31	-
Toluene	108-88-3	34	-	10
Trifluoromethane	75-46-7	17	28	-
Xenon	7440-63-3	21	25	-

[2]The publication lists several models for nitrous oxide. Here, the 2CLJQ model was used.

Table 2: Parameters of Eq. (1) and Eq. (2) for all studied solvents obtained from molecular simulation data of the pure solvents reported in the original publications, see Tab. 1.

Name	Solvent-ID	a_2	a_1	a_0	A	B	C
1,1-Dimethylhydrazine	11	-6.94e-05	3.32e-02	9.16e+00	9.89	1644.08	1.47
Acetone	2	-8.95e-05	4.09e-02	9.23e+00	9.17	1139.46	-55.94
Acetonitrile	1	-8.20e-05	2.92e-02	1.76e+01	9.70	1544.03	-33.58
Benzene	6	-4.47e-05	2.09e-02	8.50e+00	10.04	2084.68	63.77
Carbon tetrachloride	9	-2.40e-05	2.52e-03	1.38e+01	8.55	852.86	-116.08
Cyclohexane	7	-4.85e-05	2.67e-02	5.07e+00	9.04	1237.39	-46.72
Cyclohexanol	13	-2.43e-05	1.05e-02	8.29e+00	8.61	969.44	-166.25
Cyclohexanone	12	-1.69e-05	4.65e-03	9.39e+00	9.37	1674.31	-45.97
Ethanol	3	-7.46e-05	2.74e-02	1.57e+01	9.97	1465.02	-56.56
Ethylene oxide	4	-9.16e-05	2.23e-02	2.13e+01	9.38	1127.57	-25.33
Formic acid	5	-7.57e-05	2.87e-02	2.40e+01	10.29	2214.90	46.15
Isopropanol	8	-6.22e-05	2.28e-02	1.19e+01	10.45	1771.50	-35.07
Methylhydrazine	15	-2.57e-05	-1.54e-03	2.16e+01	9.42	1418.84	-43.64
Octamethylcyclotetrasiloxane	14	-6.64e-06	1.02e-03	3.47e+00	8.87	1359.57	-99.07
Toluene	10	-1.64e-05	3.65e-04	1.08e+01	9.21	1413.38	-48.46

Validity of the Linear Correlation

For a substantial number of investigated systems, the experimental binary interaction parameter ξ_{ij}^{exp} lies outside of the interval [0.95,1.05] in which molecular simulations were carried out. Especially systems with the solute Neon or Hydrogen exhibit extreme values for ξ_{ij}^{exp}. We therefore performed additional molecular simulations for all 16 systems containing either of these solutes, this time using $\xi_{ij} = 0.5$ and $\xi_{ij} = 1.5$. We then compared the obtained Henry's law constants with those predicted by the linear correlation, cf. Eq. (6) in the manuscript, which was fitted to three simulation data points obtained from simulations with $\xi_{ij} = 0.95$, 1.00, and 1.05. On average, the resulting difference in $\ln(H_{ij}/\mathrm{kPa})$ amounts to 0.18 for $\xi_{ij} = 0.5$ and 0.09 for $\xi_{ij} = 1.5$. Considering that in our previous work,[2] we found a mean standard deviation of about 0.1 for the reported experimental values of $\ln(H_{ij}/\mathrm{kPa})$ in the literature, these differences are acceptable. In addition, more than 90 % of the values for ξ_{ij}^{exp} lie in the interval [0.8,1.2], for which the extrapolation errors are much smaller.

Fitted Binary Interaction Parameters

The fitted binary interaction parameters ξ_{ij}^{exp} for all 213 binary systems investigated in this work are reported in the machine-readable file *xi_exp.csv* available with the ESI. The solute (rows) and solvent (columns) IDs refer to those defined in Tab. 1.

Predicted Binary Interaction Parameters

The completed matrix of binary interaction parameters $\xi_{ij}^{\mathrm{pred,full}}$ and their standard deviations are depicted in Fig. 2 and reported in the machine-readable files *xi_pred_full.csv* and *std_xi_pred_full.csv* available with the ESI. The values are inferred with the MCM that was trained on all fitted binary interaction parameters ξ_{ij}^{exp}. The solute (rows) and solvent (columns) IDs refer to those defined in Tab. 1.

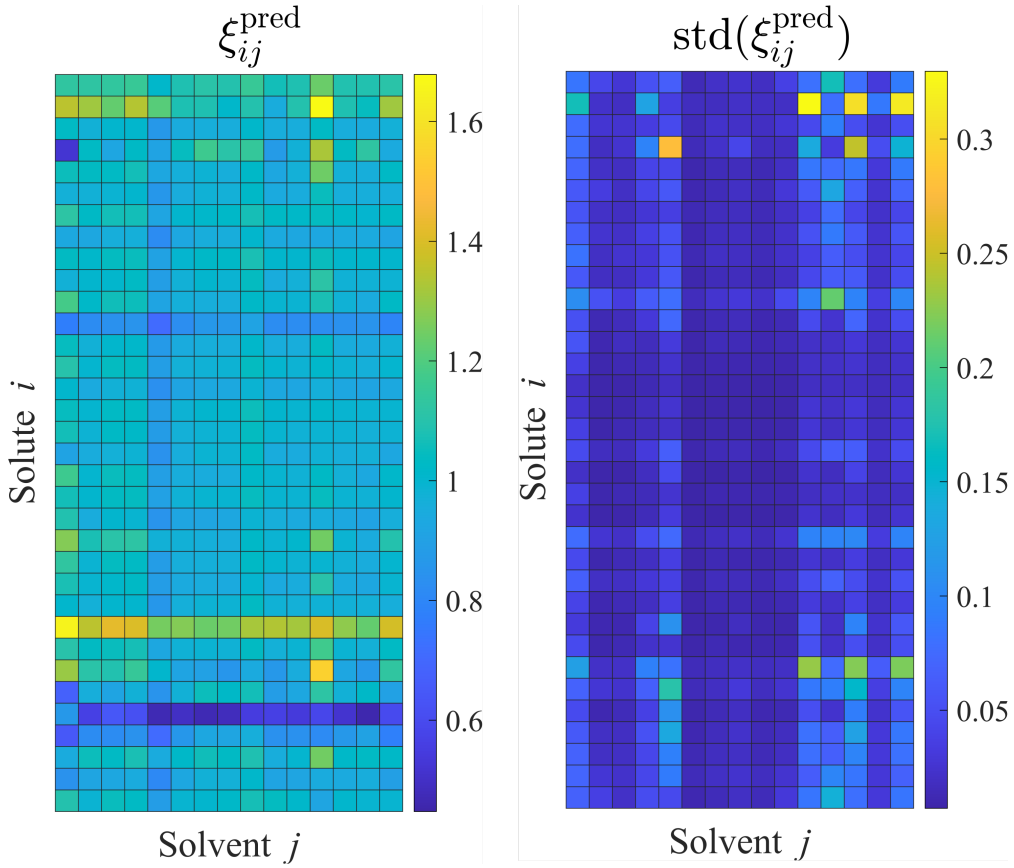


Figure 2: Heatmaps of the binary interaction parameters ξ_{ij}^{pred} predicted with the MCM trained on all ξ_{ij}^{exp} values. Left: mean values. Right: standard deviations. Solute and solvent models are sorted by their ID as listed in Tab. 1.

Model Uncertainty

The values predicted from the MCM are the mean of 1000 samples drawn from the posterior, i.e., the probability distribution of the model parameters after the training. This way, the standard deviation is obtained for each parameter, capturing the model uncertainty in the parameters of both the solute and the solvent.

Fig. 3 helps to understand how the model uncertainty depends on the number of available data points in the training data. The top panel shows the uncertainty of ξ_{ij}^{pred} as a function of the number of available training data points for a certain solute (i.e., systems in the same row of the matrix). The bottom panel shows the same but for the solvents (i.e., systems in the same column of the matrix). As expected, both plots indicate that, on average, an increase in the number of training data points leads to a decrease in the model uncertainty.

Furthermore, we have investigated how the model uncertainty differs between the predictions made with the model trained by LOO (where the predicted data point is not part of the training set) and the model trained on the complete data set (where the predicted data point is part of the training set). For 209 of the 213 data points, the standard deviation of the 'full' model is lower than that of the LOO model. On average, the uncertainty decreases by 18 %.

In Fig. 4, we show how the difference between the predicted ξ_{ij}^{pred} and the experimental ξ_{ij}^{exp}, i.e., the prediction error, depends on the number of training data points in similar plots as those shown in Fig. 3. Here, similar trends can be observed with, on average, higher numbers of training data points, leading to decreasing prediction errors. However, the trends are not as significant as in Fig. 3, indicating that the MCM also performs well in situations with few training data.

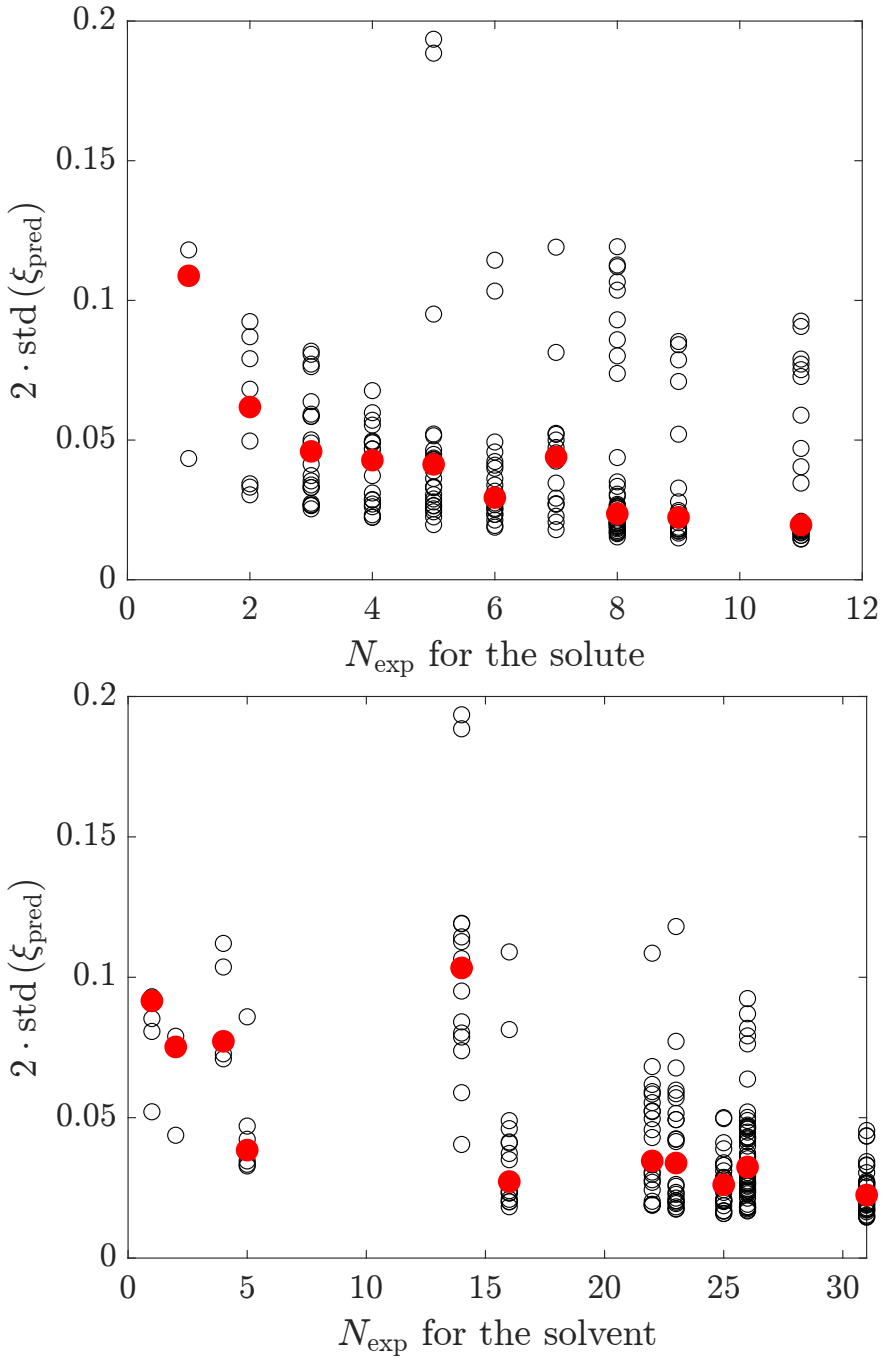


Figure 3: Influence of the number of training data points N_{exp} on the standard deviation describing the model uncertainty of ξ_{pred} from the LOO-trained model. The open circles show the standard deviation for individual systems, the red dots mark the median of the standard deviations at a specific N_{exp}. Top panel: solutes. Bottom panel: solvents. Note that for clarity, values higher than 0.2 are cut-off.

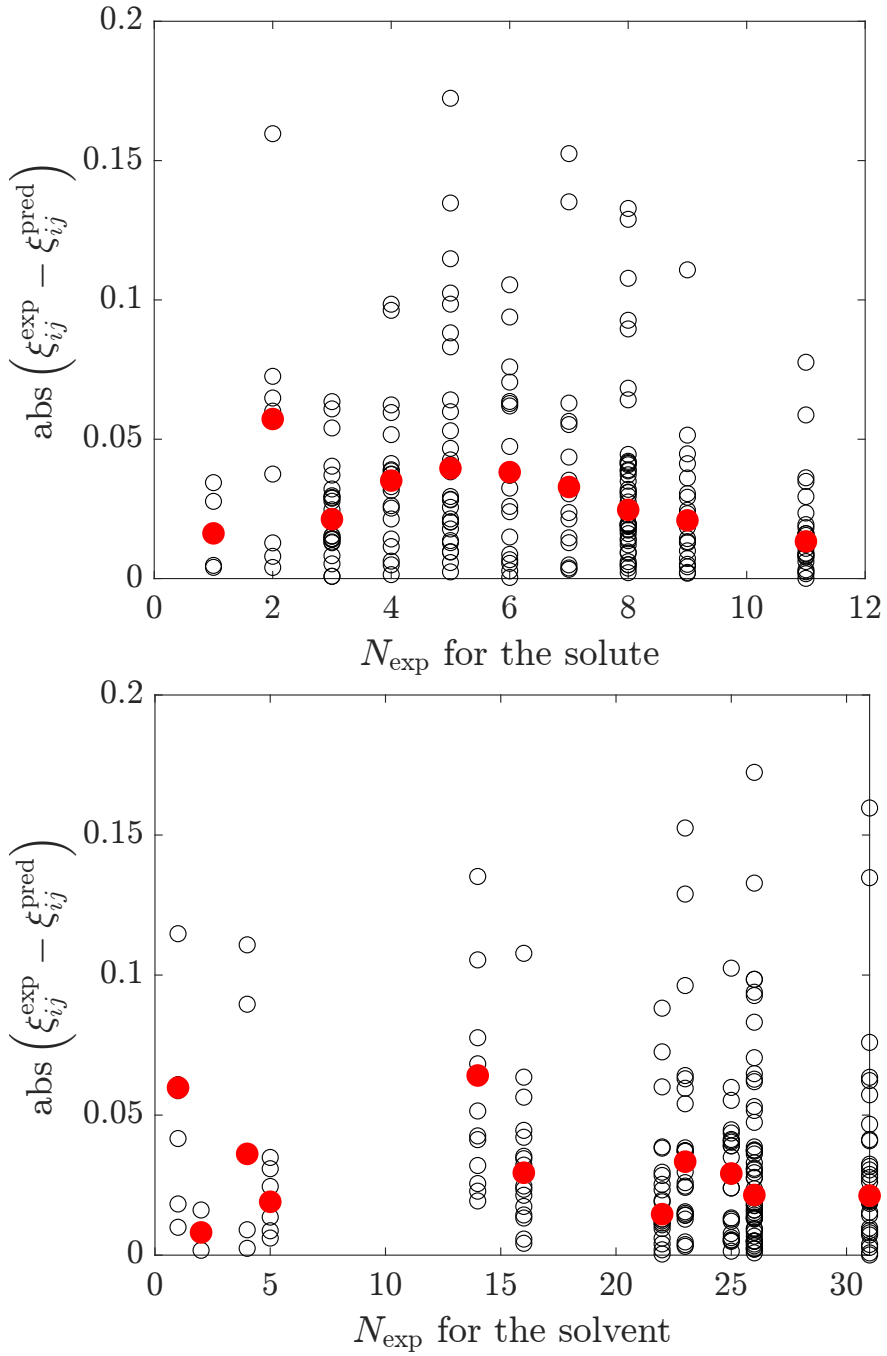


Figure 4: Influence of the number of training data points N_{exp} on the absolute difference between the predicted ξ_{ij}^{pred} from the LOO-trained model and the experimental ξ_{ij}^{exp}. The open circles show the absolute difference for individual systems, the red dots mark the median of the absolute differences at a specific N_{exp}. Top panel: solutes. Bottom panel: solvents. Note that for clarity, values higher than 0.2 are cut-off.

Application to Molecular Simulation

In Fig. 5 molecular simulation results of temperature-dependent Henry's law constants are shown for four binary systems and compared to consolidated experimental data. The plot includes results using ξ_{ij}^{exp}, ξ_{ij}^{pred} and $\xi_{ij}^{\mathrm{pred,full}}$.

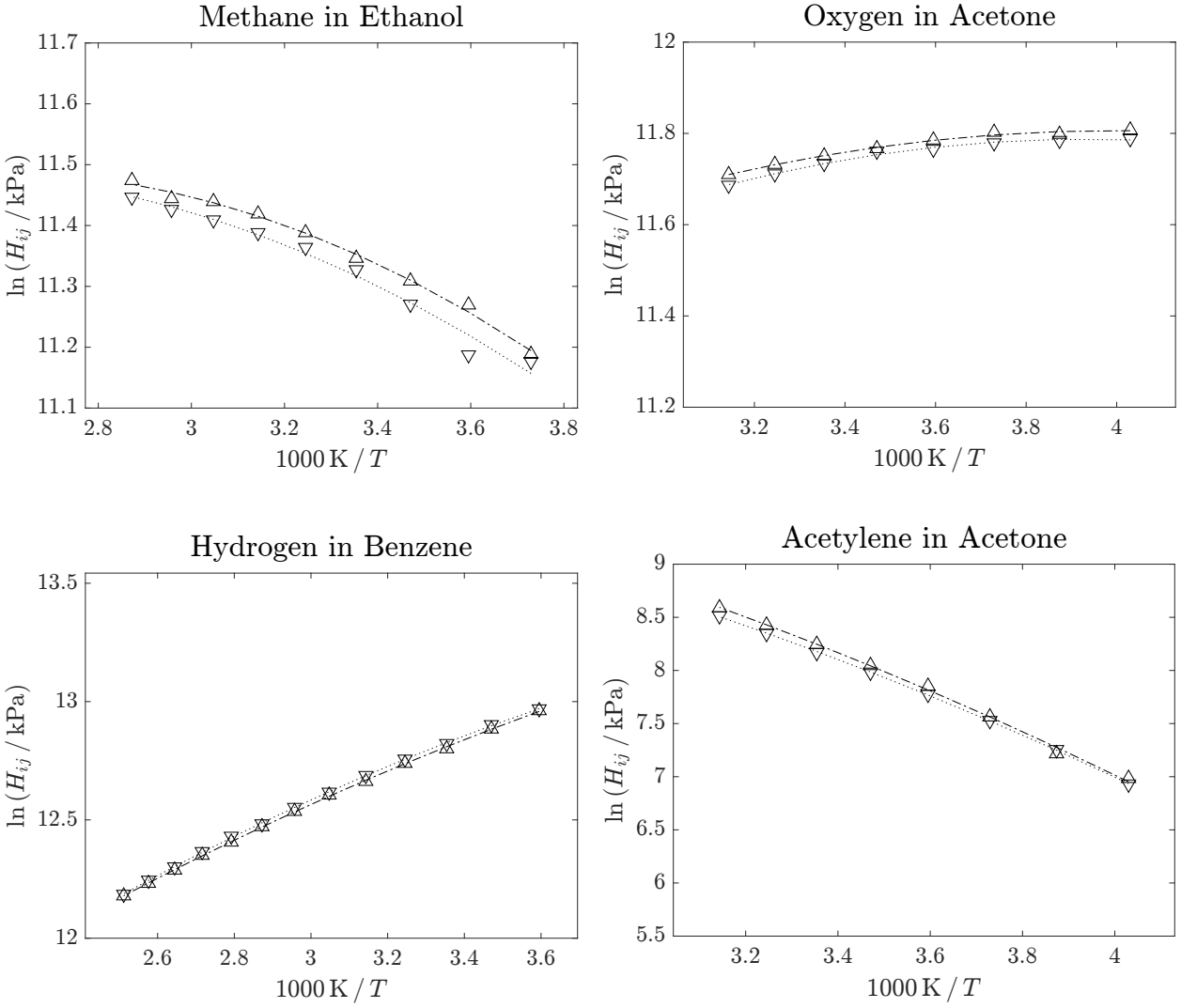


Figure 5: Results of molecular simulations of temperature-dependent Henry's law constants in four binary systems using ξ_{ij}^{pred} ($\triangle$, dash-dotted) predicted by the MCM with LOO and $\xi_{ij}^{\mathrm{pred,full}}$ ($\triangledown$, dotted) predicted by the MCM trained on all ξ_{ij}^{exp}. Error bars of the statistical uncertainty of molecular simulations are omitted for clarity. Lines are guides for the eye. Axis limits are chosen as in Fig. 5 in the paper for comparability.

References

(1) Dortmund Data Bank. www.ddbst.com, 2022.

(2) Hayer, N.; Jirasek, F.; Hasse, H. Prediction of Henry's law constants by matrix completion. *AIChE Journal* **2022**, *68*.

(3) Bodor, E.; Bor, G.; Maleczkine, S.; Mesko, G.; Mohai, B.; Siposs, G. *Veszpr.Veg.Egyet.Közlem.* **1957**, *1*, 63–76.

(4) Bodor, E.; Mohai, B.; Pfeifer, G. *Veszpr.Veg.Egyet.Közlem.* **1959**, *3*, 205–209.

(5) Eck, J. US Patent No. US 2664997. US-Patent, 1954.

(6) Hannaert, H.; Haccuria, M.; Mathieu, M. *Ind.Chim.Belge* **1967**, *32*, 156–164.

(7) Hölemann, P.; Hasselmann, R. *Chem.Ing.Tech. CIT* **1953**, *25*, 466–468.

(8) Horiuti, J. *Sci.Papers Inst.Phys.Chem.Res. (Japan)* **1931**, *17*, 125–256.

(9) Kvasenkov, I.; Shleinikov, V. *J.Appl.Chem.USSR* **1969**, *42*, 1727–1730.

(10) Otsuka, E.; Takada, M. *Nenryo-Kyokai-shi* **1963**, *42*, 229–237.

(11) Shenderei, E.; Ivanovskii, F. *Zh.Prikl.Khim.* **1964**, *37*, 1557–1562.

(12) Usyukin, I.; Shleinikov, V. *Novosti Neft.Gaz.Tekhn.Neftepererabotka Neftekhim.* **1961**, 33–39.

(13) Widom, B. Some Topics in the Theory of Fluids. *The Journal of Chemical Physics* **1963**, *39*, 2808–2812.

(14) Stephan, S.; Horsch, M. T.; Vrabec, J.; Hasse, H. MolMod – an open access database of force fields for molecular simulations of fluids. *Molecular Simulation* **2019**, *45*, 806–814.

(15) Lustig, R. Angle-Average for the Powers of the Distance between two Separated Vectors. *Molecular Physics* **1988**, *65*, 175–179.

(16) Nymand, T. M.; Linse, P. Ewald summation and reaction field methods for potentials with atomic charges, dipoles, and polarizabilities. *The Journal of Chemical Physics* **2000**, *112*, 6152–6160.

(17) Stoll, J.; Vrabec, J.; Hasse, H. A set of molecular models for carbon monoxide and halogenated hydrocarbons. *The Journal of Chemical Physics* **2003**, *119*, 11396–11407.

(18) Elts, E.; Windmann, T.; Staak, D.; Vrabec, J. Fluid phase behavior from molecular simulation: Hydrazine, Monomethylhydrazine, Dimethylhydrazine and binary mixtures containing these compounds. *Fluid Phase Equilibria* **2012**, *322-323*, 79–91.

(19) Windmann, T.; Linnemann, M.; Vrabec, J. Fluid Phase Behavior of Nitrogen + Acetone and Oxygen + Acetone by Molecular Simulation, Experiment and the Peng–Robinson Equation of State. *Journal of Chemical Engineering Data* **2013**, *59*, 28–38.

(20) Deublein, S.; Metzler, P.; Vrabec, J.; Hasse, H. Automated development of force fields for the calculation of thermodynamic properties: acetonitrile as a case study. *Molecular Simulation* **2013**, *39*, 109–118.

(21) Vrabec, J.; Stoll, J.; Hasse, H. A Set of Molecular Models for Symmetric Quadrupolar Fluids. *The Journal of Physical Chemistry B* **2001**, *105*, 12126–12133.

(22) Eckl, B.; Vrabec, J.; Hasse, H. An optimised molecular model for ammonia. *Molecular Physics* **2008**, *106*, 1039–1046.

(23) Guevara-Carrion, G.; Janzen, T.; Muñoz-Muñoz, Y. M.; Vrabec, J. Mutual diffusion of binary liquid mixtures containing methanol, ethanol, acetone, benzene, cyclohexane, toluene, and carbon tetrachloride. *The Journal of Chemical Physics* **2016**, *144*, 124501.

(24) Merker, T.; Engin, C.; Vrabec, J.; Hasse, H. Molecular model for carbon dioxide optimized to vapor-liquid equilibria. *The Journal of Chemical Physics* **2010**, *132*, 234512.

(25) Stöbener, K.; Klein, P.; Horsch, M.; Küfer, K.; Hasse, H. Parametrization of two-center Lennard-Jones plus point-quadrupole force field models by multicriteria optimization. *Fluid Phase Equilibria* **2016**, *411*, 33–42.

(26) Eckl, B.; Vrabec, J.; Hasse, H. Set of Molecular Models Based on Quantum Mechanical Ab Initio Calculations and Thermodynamic Data. *The Journal of Physical Chemistry B* **2008**, *112*, 12710–12721.

(27) Merker, T.; Guevara-Carrión, G.; Vrabec, J.; Hasse, H. *High Performance Computing in Science and Engineering '08*; Springer Berlin Heidelberg, pp 529–541.

(28) Merker, T.; Vrabec, J.; Hasse, H. Molecular simulation study on the solubility of carbon dioxide in mixtures of cyclohexane + cyclohexanone. *Fluid Phase Equilibria* **2012**, *315*, 77–83.

(29) Muñoz-Muñoz, Y. M.; Guevara-Carrion, G.; Llano-Restrepo, M.; Vrabec, J. Lennard-Jones force field parameters for cyclic alkanes from cyclopropane to cyclohexane. *Fluid Phase Equilibria* **2015**, *404*, 150–160.

(30) Schnabel, T.; Vrabec, J.; Hasse, H. Henry's law constants of methane, nitrogen, oxygen and carbon dioxide in ethanol from 273 to 498 K: Prediction from molecular simulation. *Fluid Phase Equilibria* **2005**, *233*, 134–143.

(31) Eckl, B.; Vrabec, J.; Hasse, H. On the application of force fields for predicting a wide variety of properties: Ethylene oxide as an example. *Fluid Phase Equilibria* **2008**, *274*, 16–26.

(32) Schnabel, T.; Cortada, M.; Vrabec, J.; Lago, S.; Hasse, H. Molecular model for formic acid adjusted to vapor–liquid equilibria. *Chemical Physics Letters* **2007**, *435*, 268–272.

(33) Köster, A.; Thol, M.; Vrabec, J. Molecular Models for the Hydrogen Age: Hydrogen, Nitrogen, Oxygen, Argon, and Water. *Journal of Chemical Engineering Data* **2018**, *63*, 305–320.

(34) Huang, Y.-L.; Heilig, M.; Hasse, H.; Vrabec, J. Vapor-liquid equilibria of hydrogen chloride, phosgene, benzene, chlorobenzene, ortho-dichlorobenzene, and toluene by molecular simulation. *AIChE Journal* **2011**, *57*, 1043–1060.

(35) Muñoz-Muñoz, Y. M.; Guevara-Carrion, G.; Vrabec, J. Molecular Insight into the Liquid Propan-2-ol + Water Mixture. *The Journal of Physical Chemistry B* **2018**, *122*, 8718–8729.

(36) Kohns, M.; Werth, S.; Horsch, M.; von Harbou, E.; Hasse, H. Molecular simulation study of the CO 2 -N 2 O analogy. *Fluid Phase Equilibria* **2017**, *442*, 44–52.

(37) Thol, M.; Rutkai, G.; Köster, A.; Dubberke, F. H.; Windmann, T.; Span, R.; Vrabec, J. Thermodynamic Properties of Octamethylcyclotetrasiloxane. *Journal of Chemical Engineering Data* **2016**, *61*, 2580–2595.