Electronic Supplementary Material (ESI) for Digital Discovery.

Electronic Supplementary Information (ESI)

Database for liquid phase diffusion coefficients at infinite dilution at 298 K and matrix completion methods for their prediction

Oliver Großmann, Daniel Bellaire, Nicolas Hayer, Fabian Jirasek, Hans Hasse

Laboratory of Engineering Thermodynamics (LTD), TU Kaiserslautern, Erwin-Schrödinger-Straße 44, 67663, Kaiserslautern, Germany

Email: fabian.jirasek@mv.uni-kl.de

Contents

S.1 Data curation

In the following, we describe the criteria that we have applied for deciding whether to adopt a data point to our database or not. First, all data points that were labeled in the Dortmund Data Bank (DDB) to be of poor quality were omitted. Furthermore, we have excluded all solutes and solvents without a well-defined molecular composition, such as polymers and pseudocomponents (e.g. seawater, jet fuel, bitumen). In cases where we found data points to be erroneously labeled in the DDB, e.g., when predicted data was reported as experimental data, or in cases where the reported type of diffusion coefficient was unclear, we have excluded that data as well.

Moreover, the consistency of the reported diffusion coefficients was assessed in two ways. First, for mixtures for which multiple data points at similar concentrations (differences below 0.02 mol/mol) were reported by different authors, those deviating by more than one standard deviation from the mean were excluded. Second, for mixtures for which data points were measured over a range of concentrations, we have removed those data points that deviated more than one standard deviation from the fitted curve describing the concentration dependence of D_{ij} (cf. description of the fitting procedure in Section 2 of the manuscript).

Going beyond the formal data curation steps described above, we note that the matrix completion methods (MCMs) developed in this work can be used to obtain information on erroneous data: MCMs basically analyze data sets for (hidden) structure, which they will not be able to find in the case of erroneous data; hence, such data points are likely to be outliers in the MCM predictions. Therefore, it is interesting to analyze the outliers in the MCM predictions closer in order to find out whether the deviation might stem from errors in the data. However, this requires applying methods beyond the MCMs and was not in the scope of the present work.

S.2 Semiempirical models

In this section, the considered semiempirical models studied in the present work are briefly presented. Further, we show exactly how D_{ij}^{∞} is calculated in each of these cases from some pure-component properties of the solutes and solvents. The pure-component properties needed for this purpose were calculated for the studied temperature $T = 298.15$ K by DIPPR correlations, which are provided in the DIPPR database.[1] For the solutes i and solvents j, these include, depending on the model, some combination of the molar masses M_i and M_j, the parachors P_i and P_j, and the saturated liquid phase molar volumes $\tilde{v}_i$ and $\tilde{v}_j$ at the respective normal boiling temperatures of solute and solvent, as well as the viscosity η_j of the solute.

§ The parachor is used here as defined by Quayle: $P_i = \sqrt[4]{\gamma_i} v_i$, where γ_i and v_i are the surface tension and liquid molar volume of pure component i at the studied temperature, respectively.[2]

With the exception of SEGWE, the semiempirical models considered here need information on the saturated liquid phase molar volume $\tilde{v}_i$ of the solute i at its normal boiling temperature. However, for carbon dioxide this value is not defined since its triple point pressure is above the ambient pressure; therefore, the liquid molar volume at the triple point was used instead here. Similarly, also for perylene and 3-hydroxyaniline $\tilde{v}_i$ at the normal boiling temperature cannot be measured since both components decompose before reaching the respective temperatures; therefore, we have used a hypothetical value for $\tilde{v}_i$ for these components, which was calculated with the group contribution method of Schröder.[3]

While the four semiempirical models have been developed as general-purpose correlations that aim at describing a diverse set of mixtures and components, there are still some restrictions in the scope of these models, which we briefly mention here. All authors have limited their models to moderate viscosities and have excluded data for viscous solvents (e.g., polymers) from their training sets. Further, none of the semiempirical models were trained on data of mixtures containing electrolytes, i.e., neither mixtures with salts as solutes nor with ionic liquids as solutes or solvents should be expected to be predicted with high accuracy.

S.2.1 Wilke and Chang, 1955

One of the first widely applicable correlations for diffusion coefficients in liquids was developed by Wilke and Chang.[4] According to the model of Wilke and Chang, D_{ij}^{∞} is calculated by:

$$\left(\frac{D_{ij}^{\infty}}{\mathrm{m^2/s}}\right) = 7.4 \times 10^{-12} \sqrt{\phi_j \left(\frac{M_j}{\mathrm{g/mol}}\right)} \frac{1}{\left(\frac{\tilde{v}_i}{\mathrm{cm^3/mol}}\right)^{0.6}} \frac{\left(\frac{T}{\mathrm{K}}\right)}{\left(\frac{\eta_j}{\mathrm{mPa\ s}}\right)} \tag{S.1}$$

where ϕ_j is a solvent-specific factor, which was introduced to improve the description of diffusion coefficients in associating solvents; for some common solvents, values for ϕ_j have been reported.[4] However, in this work, values for ϕ_j were fitted for each solvent individually to experimental D_{ij}^{∞} from our database (cf. Section 3.2.4).

S.2.2 Reddy and Doraiswamy, 1967

Reddy and Doraiswamy sought to improve on the Wilke-Chang correlation by eliminating the factor ϕ_j and considering the molar volume $\tilde{v}_j$ of the solvent instead.[5] They also changed the exponent of both $\tilde{v}_i$ and $\tilde{v}_j$ to $\frac{1}{3}$, an idea that was previously introduced by Scheibel,[6] resulting in Equation S.2:

$$\left(\frac{D_{ij}^{\infty}}{\mathrm{m^2/s}}\right) = K_{\mathrm{RS}} \frac{\sqrt{\frac{M_j}{\mathrm{g/mol}}}}{\sqrt[3]{\left(\frac{\tilde{v}_i}{\mathrm{cm^3/mol}}\right)\left(\frac{\tilde{v}_j}{\mathrm{cm^3/mol}}\right)}} \frac{\left(\frac{T}{\mathrm{K}}\right)}{\left(\frac{\eta_j}{\mathrm{mPa\ s}}\right)}. \tag{S.2}$$

The empirical constant K_{RS} depends on the ratio of $\tilde{v}_i$ to $\tilde{v}_j$:

$$K_{\mathrm{RS}} = \begin{cases} 10 \times 10^{-12}, & \text{for } \frac{\tilde{v}_j}{\tilde{v}_i} \leq 1.5 \\ 8.5 \times 10^{-12}, & \text{for } \frac{\tilde{v}_j}{\tilde{v}_i} > 1.5 \end{cases} \tag{S.3}$$

S.2.3 Tyn and Calus, 1975

Tyn and Calus found that the ratio of the parachors P_i and P_j correlates strongly with D_{ij}^{∞},[7] and therefore proposed the following equation:

$$\left(\frac{D_{ij}^{\infty}}{\mathrm{m}^2/\mathrm{s}}\right) = 8.93 \times 10^{-12} \sqrt[6]{\frac{\left(\frac{\tilde{v}_i}{\mathrm{cm}^3/\mathrm{mol}}\right)}{\left(\frac{\tilde{v}_j}{\mathrm{cm}^3/\mathrm{mol}}\right)^2}} \left(\frac{P_j}{P_i}\right)^{0.6} \frac{\left(\frac{T}{\mathrm{K}}\right)}{\left(\frac{\eta_j}{\mathrm{mPa\ s}}\right)}. \tag{S.4}$$

The Tyn and Calus model is subject to the following restrictions:[7]

- For the solute water, the authors suggest that water should be treated as a dimer, i.e., the values of $\tilde{v}_i$ and P_i should be doubled. In this work, we have used the values $\tilde{v}_{water} = 37.4$ cm^3/mol and $P_{water} = 105.2$ cm^3 g$^{1/4}$/(s$^{1/2}$ mol) for the water dimer, as recommended by Poling.[8]
- When the solute is an organic acid, the dimer value of $2\tilde{v}_i$ and $2P_i$ should be used in solvents other than water, methanol, and butanol. In the present work, we have followed this suggestion.
- For nonpolar solutes in monohydroxy alcohol solvents, the values of $\tilde{v}_j$ and P_j should be multiplied by the factor $8\eta_j$, with the solvent viscosity η_j in units of mPa s, which was done accordingly in the present work.

S.2.4 SEGWE (Stokes-Einstein Gierer-Wirtz Estimation)

In a recent work of Evans et al., the Stokes-Einstein equation[9] was extended by introducing the Gierer-Wirtz[10] correction to loosen the assumption of the Stokes-Einstein theory that the solvent is a continuum fluid.[11] Consequently, they named their model SEGWE (Stokes-Einstein Gierer-Wirtz Estimation), which calculates D_{ij}^{∞} as:

$$D_{ij}^{\infty} = \frac{k_{\mathrm{B}}}{6\pi} \frac{\left(\frac{3\alpha}{2} + \frac{1}{1+\alpha}\right)}{\sqrt[3]{\frac{3M_i}{4\pi\varrho_{\mathrm{eff}}N_{\mathrm{A}}}}} \frac{T}{\eta_j} \tag{S.5}$$

where ϱ_{eff} is the effective density and α is the ratio of the solvent and solute radii, r_j and r_i, respectively. Further, k_{B} and N_{A} are the Boltzmann and Avogadro constants, respectively. Assuming that all molecules are hard spheres, α can also be expressed in terms of the molar masses M_j and M_i:

$$\alpha = \frac{r_j}{r_i} = \sqrt[3]{\frac{M_j}{M_i}}. \tag{S.6}$$

The effective density ϱ_{eff}, which can be considered either as a solvent-specific parameter or fitted to a global value, was fitted by the original authors to diffusion coefficient data at 25 °C for 109 combinations of 44 solutes and 5 solvents, yielding a global value of 619 kg/m^3.[11]

In the present work, we use ϱ_{eff} as a solvent-specific parameter, which we have fitted individually to the respective data on D_{ij}^{∞} for each solvent from our database; as described above, we thereby followed a leave-one-out strategy (cf. Section 3.2.4).

S.2.5 Effect of fitting the model parameters with a leave-one-out strategy

Both the Wilke-Chang and SEGWE models contain a solvent-specific fit parameter, called ϕ_j and $\varrho_{\text{eff},j}$, respectively. For a fair comparison to the MCMs, these were fitted to the new database using a leave-one-out strategy in the present work: i.e., for the prediction of each experimental D_{ij}^{∞}, a $\phi_j^{(i)}$ (or $\varrho_{\text{eff},j}^{(i)}$) was fitted to all available experimental data in that particular solvent *minus* the data point $i+j$ that is to be predicted. The optimum $\phi_j^{(i),*}$ was chosen for the minimum in the rRMSE:

$$\phi_j^{(i),*} = \underset{\phi_j^{(i)}}{\arg\min} \sum_{k \neq i} \left(\frac{D_{kj}^{\infty,\text{pred}}\left(\phi_j^{(i)}\right) - D_{kj}^{\infty,\text{exp}}}{D_{kj}^{\infty,\text{exp}}} \right)^2 \tag{S.7}$$

However, it is also possible to apply the Wilke-Chang and SEGWE models in a purely predictive manner: for Wilke-Chang this means using the (few) parameter values of ϕ_j supplied by the original authors, for SEGWE the global value $\varrho_{\text{eff}} = 619$ kg/m^3 is used.

For both models, there is only a small difference in the overall performance when comparing the purely predictive approach to that with the fitted parameter. The effect is shown in Figure S.1. For SEGWE, the rMAE and rRMSE decrease from 0.213 and 0.285 in the predictive approach to 0.193 and 0.276 in the fitted approach, respectively. For Wilke-Chang, the rMAE decreases from 0.227 to 0.209, while, surprisingly, the rRMSE slightly increases from 0.304 to 0.314. This paradoxical effect is due to the large number of solvents in which data is available only for very little mixtures (i.e. solvents that have been measured in combinations with few solutes). In such cases, the leave-one-out strategy will lead to a good fit of ϕ_j to the (limited) available data, while the left-out point may therefore be grossly mispredicted, resulting in a high rRMSE.

S.2.6 Mixtures poorly described by semiempirical models

In this section, we take a closer look at those mixtures from our database, for which D_{ij}^{∞} is only poorly described by the semiempirical models, and we try to specify those groups of solutes and solvents for which this is the case. We thereby focus on SEGWE, but also briefly touch upon the other models.

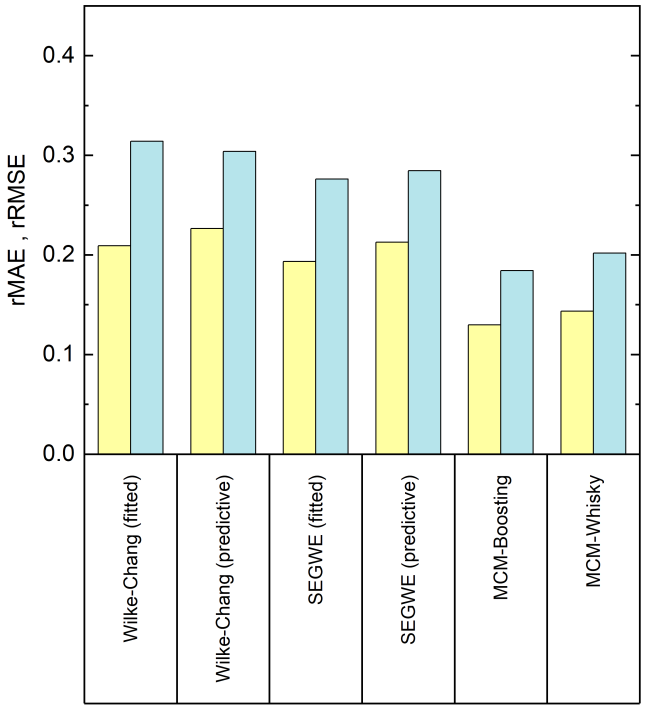

Figure S.1: Relative mean absolute error (rMAE, yellow) and relative root mean-squared error (rRMSE, blue) of the predicted D_{ij}^{∞} for the experimental data from the reduced database. We compare the developed MCMs to the semiempirical models Wilke-Chang and SEGWE in two variants: a purely predictive one and a one that was fitted to the database of this work using a leave-one-out strategy.

For discussing the performance of SEGWE in detail, we refer to Figure 7 in the manuscript, which shows the residuals of the SEGWE predictions from the experimental data. One solute that SEGWE is apparently struggling to describe accurately is water (solute $i = 27$, cf. Figure 7). In our reduced database, there are eight mixtures with the solute water; the relative deviations of the SEGWE predictions from the experimental data for D_{ij}^{∞} for these eight mixtures are shown in Figure S.2.

We find the largest positive relative deviations for mixtures in which strong hydrogen bonding occurs, namely the mixtures (water + ethanol) and (water + 1-propanol). Slightly smaller, but still large positive relative deviations are found for mixtures of water with solvents in which weaker hydrogen bonds are formed (acetone, butyl acetate, *N*-methyl-2-pyrrolidone and methyl isopropyl ketone, cf. Figure S.2). This is not astonishing as the developers of SEGWE have explicitly excluded data for mixtures with "aggregating components" in the development of SEGWE.[11] Aggregation leads to lower diffusion coefficients; an effect which is not described by SEGWE, which, as a consequence, overpredicts

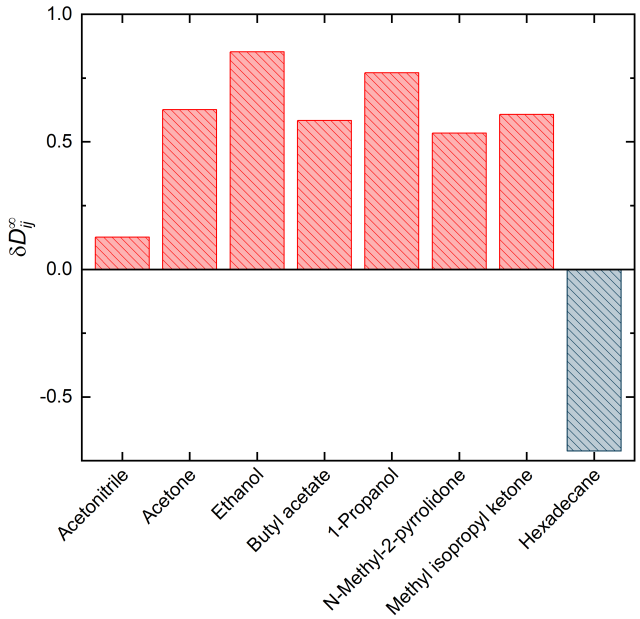

Figure S.2: Relative deviations $\delta D_{ij}^{\infty} = \left(D_{ij}^{\infty,\mathrm{pred}} - D_{ij}^{\infty,\mathrm{exp}}\right) / D_{ij}^{\infty,\mathrm{exp}}$ of the SEGWE predictions for D_{ij}^{∞} of the solute water in different solvents from the experimental data from the reduced database.

D_{ij}^{∞} in such mixtures, cf. Figure S.2.

High positive relative deviations of the SEGWE predictions from the experimental data are also found for many other hydrogen bonding systems in our database.

Furthermore, SEGWE mispredicts D_{ij}^{∞} in mixtures where the molecular mass in relation to the molecule size strongly differs between both components. This is in particular the case if one of the components contains heavy atoms, and the other does not. The reason for this is that in the development of SEGWE, it was assumed that both solute and solvent can be modeled as hard spheres, and that both spheres have an equal ratio of mass to volume – the so-called effective density ϱ_{eff} of the mixture.

An instructive example for this case is the result for the solute carbon dioxide ($i = 39$) in Figure 7 of the manuscript. Carbon dioxide has a relatively large molecular mass in relation to its small molecular volume, which leads to a rather high effective density compared to, e.g., typical organic solvents. Accordingly, we find SEGWE to significantly underestimate D_{ij}^{∞} for basically all mixtures with carbon dioxide from the reduced database (cf. Figure 7 in the manuscript), and even for all mixtures with carbon dioxide from the full database (not shown here).

Two other examples for solutes in our database with rather high effective

densities are methyl iodide ($i = 19$), which is due to the heavy iodine atom, and the fully fluorinated hexafluorobenzene ($i = 30$); we find that SEGWE also underestimates the diffusion in all mixtures containing these two solutes. Returning to Figure S.2 as a last example, we can likewise explain the significant underestimation of the experimental D_{ij}^{∞} in the mixture (water + hexadecane) by the higher effective density of water in relation to that of hexadecane (and the absence of significant attractive forces in the mixture to counteract this effect).

Finally, we briefly touch on the limitations of the models of Wilke and Chang,[4] Reddy and Doraiswamy,[5] and Tyn and Calus.[7] Due to their similar nature they are all subject to similar restrictions, so that they will be discussed together here. Despite the original authors' intention to provide general-purpose correlations that work in nonpolar and polar mixtures alike, all three models have been found to struggle significantly with hydrogen bonding mixtures (as it is also the case for SEGWE). Hence, they overpredict D_{ij}^{∞} for hydrogen bonding solvents, such as methanol, ethanol and 1-propanol. Further, the Wilke-Chang model is inaccurate in the prediction of the diffusion of water in organic solvents, which has been described before in the literature.[12] Accordingly, we find a significant overestimation of D_{ij}^{∞} by the Wilke-Chang model for nearly all mixtures from the reduced database in which water is the solute, with the exception of the mixture (water + hexadecane). This trend is not observed for the models of Tyn and Calus or Reddy and Doraiswamy.

Lastly, we note that MCMs can be used to identify such systematic deviations in the predictions of (semiempirical) models, and that MCMs can also predict them, which is used in the hybrid MCM based on "boosting" for improving the performance of the semiempirical models, cf. Figure 6 in the manuscript.

S.3 Maximum errors in the predictive performance of the studied models

In Figure S.3, we present the relative maximum absolute error, defined by Equation S.8, of the predictions for D_{ij}^{∞} with the four semiempirical models and the three MCMs studied in this work on the reduced database is shown. We find similar results as in Figure 6 in the manuscript, namely that the performance of the data-driven MCM suffers from some drastic mispredictions (leading to the high relative maximum absolute error seen here), and that both hybrid MCMs outperform the semiempirical models in this statistic too. Again, we find that MCM-Boosting performs slightly better than MCM-Whisky in terms of the relative maximum absolute error.

$$\text{relative maximum absolute error} = \max_{i,j} \left| \frac{D_{ij}^{\infty,\text{pred}} - D_{ij}^{\infty,\text{exp}}}{D_{ij}^{\infty,\text{exp}}} \right| \tag{S.8}$$

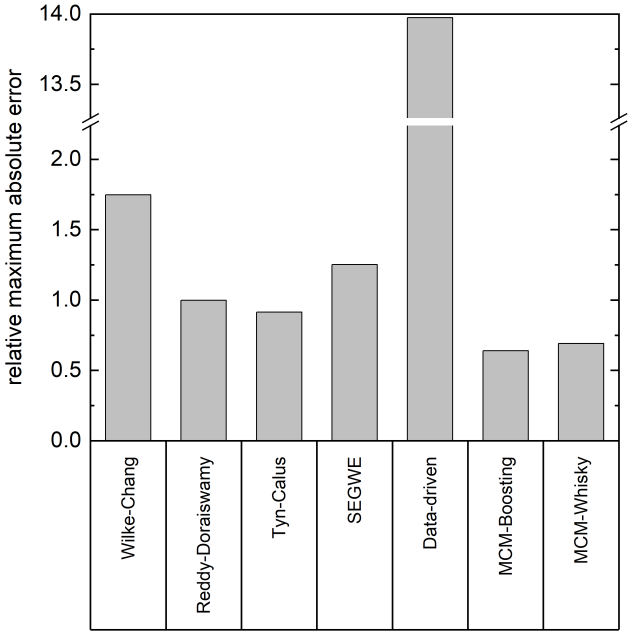

Figure S.3: Relative maximum absolute error of the predicted D_{ij}^{∞} with the studied semiempirical models and the developed MCMs for the experimental data from the reduced database.

S.4 Complete predictions from MCM-Whisky

Analogous to the MCM-Boosting results in Figure 10 of the main manuscript, we show in Figure S.4 the completed D_{ij}^{∞} matrix from the MCM-Whisky predictions together with the uncertainties of those predictions.

S.5 Supplementary tabular files

Additional Supplementary Information is provided in a machine readable format in the form of .csv files. The data is provided in two separate folders, named "full" and "reduced", representing the full database and the reduced database that we provide. In each folder, the following files are found:

- *List_Solutes.csv* and *List_Solvents.csv*: Here, we give information on all 208 (45) solutes and all 51 (23) solvents, respectively, appearing in the full (reduced) database. Specifically, we report the respective identifiers used in the Dortmund Data Bank (DDB), i.e., the DDB No., which are also used for identifying the components in the other tables, and the CAS Registry Numbers.

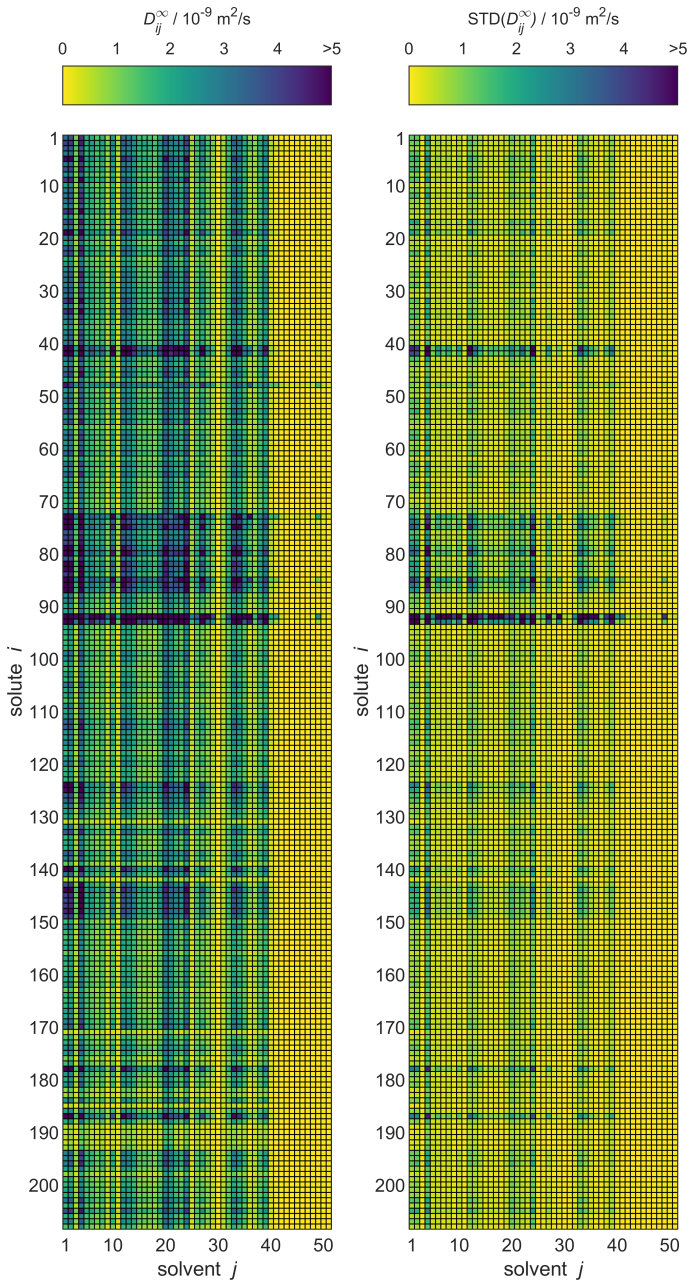

Figure S.4: Predictions of D_{ij}^{∞} by MCM-Whisky (left) and the uncertainties of the predictions (right) for all solutes i and solvents j (identified by numbers, see Table S.1) from the full database. The color code indicates the values of D_{ij}^{∞}.

- *DataBase.csv*: Here, we report the numerical values of the full (reduced) database of experimental D_{ij}^{∞} as described in Section 2. The database covers 208 (45) solutes and 51 (23) solvents and includes a total of 353 (166) experimental data points. Data values that were directly adopted from the proprietary DDB without modification are censored in the table.

- *SEGWE*: Here, we report the numerical values of SEGWE for the prediction of D_{ij}^{∞} with a fixed $\varrho_{\text{eff}} = 619$ kg/m^3 on the full (reduced) database. These values of SEGWE were used in the hybridization of the developed MCMs (cf. Manuscript Section 3.2).

- *Boosting_ Predictions.csv*: Here, we report the predictions of D_{ij}^{∞} with the hybrid MCM "MCM-Boosting" developed in this work (cf. Section 3.2.2). The results were obtained here after training the model on all data on D_{ij}^{∞} in the full (reduced) database. The predictions are listed for all 10,608 (1,035) solute-solvent combinations and include a large number of novel data points on D_{ij}^{∞}. In the same table, the the model uncertainty of each predicted D_{ij}^{∞} is listed next to the predicted value in the form of standard deviations.

- *Boosting_ LV_ Solutes.csv* and *Boosting_ LV_ Solvents.csv*: Here, we report the results of the training of MCM-Boosting on the full (reduced) database of D_{ij}^{∞}, which are the feature vectors $\boldsymbol{u_i}$ and $\boldsymbol{v_j}$ of the solutes and solvents, respectively. The length of the feature vectors is $K = 2$.

- *Whisky_ Predictions.csv*: Here, we report the predictions of D_{ij}^{∞} with the hybrid MCM "MCM-Whisky" developed in this work (cf. Section 3.2.2). The results were obtained here after training the model on all data on D_{ij}^{∞} in the full (reduced) database. The predictions are listed for all 10,608 (1,035) solute-solvent combinations and include a large number of novel data points on D_{ij}^{∞}. In the same table, the the model uncertainty of each predicted D_{ij}^{∞} is listed next to the predicted value in the form of standard deviations.

- *Whisky_ LV_ Solutes.csv* and *Whisky_ LV_ Solvents.csv*: Here, we report the results of the training of MCM-Whisky on the full (reduced) database of D_{ij}^{∞}, which are the feature vectors $\boldsymbol{u_i}$ and $\boldsymbol{v_j}$ of the solutes and solvents, respectively. The length of the feature vectors is $K = 2$.

An excerpt of this information is also provided in written form in Tables S.1-S.5.

S.6 Tabular material

Table S.1: Table of all components, subdivided into solutes and solvents, encountered in the data base on D_{ij}^{∞} developed in this work. All components are listed by their consecutive number, as used in all figures throughout this work, together with their DDB identification number.

Cons. No.	DDB No.	Name	Cons. No.	DDB No.	Name
		Solutes			
1	2	Acetamide	105	3063	L-Ascorbic acid
2	3	Acetonitrile	106	3215	4-Hydroxy-3-methoxybenzaldehyde
3	4	Acetone	107	3258	2,2-Bis(hydroxymethyl)-1,3-propanediol
4	8	1,2-Ethanediol	108	3347	D-(+)-Saccharose
5	11	Ethanol	109	3410	1,3,5-Triisopropylbenzene
6	12	Diethyl ether	110	3468	DL-Phenylalanine
7	15	Formic acid	111	3523	1,4-Diaminobenzene
8	17	Aniline	112	3715	Benzenesulfonic acid
9	21	Ethyl acetate	113	3717	p-Toluenesulfonic acid
10	24	Benzyl alcohol	114	3724	L-Alanine
11	25	Ethylbenzene	115	3725	L-Serine
12	26	Bromobenzene	116	3729	Glycine
13	27	Chlorobenzene	117	3731	L-(+)-Aspartic acid
14	30	Nitrobenzene	118	3732	L-Glutamic acid
15	31	Benzene	119	3865	Piperazine
16	39	1-Butanol	120	3988	beta-Alanine
17	40	2-Butanone	121	3989	4-Aminobutyric acid
18	41	n-Butane	122	3990	5-Aminovaleric acid
19	46	Butyl chloride	123	3991	6-Aminohexanoic acid
20	47	Chloroform	124	4490	Potassium thiocyanate
21	49	3-Methylphenol	125	4577	Potassium chloride
22	50	Cyclohexane	126	4591	Cadmium chloride
23	72	N,N-Dimethylformamide	127	4592	Nickel chloride
24	77	2,6-Dimethylpyridine	128	4596	Ferrocene
25	78	Dodecane	129	4707	(+-)-alpha-Aminobutyric acid
26	79	Benzaldehyde	130	4708	alpha-Aminoisobutanoic acid
27	80	Butyl acetate	131	4771	Buckminsterfullerene
28	84	Acetic acid	132	4776	2-Acetoxy benzoic acid
29	85	Furfural	133	4792	Sodium nitrate
30	89	Hexane	134	4795	D-Mannose
31	91	Heptane	135	4801	!D-Xylose
32	99	Methyl iodide	136	4817	1,2,6-Hexanetriol
33	108	1-Methylnaphthalene	137	4911	Sodium chloride
34	110	Methanol	138	4955	Magnesium chloride
35	112	3-Methylpentane	139	4960	Magnesium sulfate
36	123	Naphthalene	140	4965	Potassium nitrite
37	129	1-Octene	141	5261	1,2-Ethanediol-D2 (deuterioglycol)
38	138	Phenol	142	5949	3,4-Dihydroxy benzoic acid
39	140	1-Propanol	143	6317	Iron(III) sulfate
40	141	Propionic acid	144	6319	Ammonium chloride
41	145	Nitric acid	145	6325	Ammonium sulfate
42	146	Hydrogen chloride	146	6326	Lead nitrate
43	147	Salicylic acid methyl ester	147	6353	Sodium perchlorate
44	153	tert-Butanol	148	6355	Potassium chlorate

Continued on next page

Continued from previous page

Cons. No.	DDB No.	Name	Cons. No.	DDB No.	Name
45	157	Tetrachloromethane	149	6372	Sodium thiocyanate
46	161	Toluene	150	6465	N-Acetyl-p-aminophenol
47	168	Trichloroethylene	151	6529	Di-tert-butylsulfide
48	174	Water	152	7467	Titanium tetra-tert.butyloxide
49	230	Glycerol	153	7533	15-Crown-5 (15C5)
50	235	Butyric acid	154	7847	L-Valine
51	237	Propane	155	7848	L-Isoleucine
52	250	Cyclohexanone	156	7852	L-Tryptophane
53	269	Caprylic acid	157	7949	L-Cystine
54	284	N-Methyl-2-pyrrolidone	158	9329	7-Aminoheptanoic acid
55	297	Hexafluorobenzene	159	10334	Tris(2,4-pentanedionato)chromium
56	308	2-Methyl-2,4-pentanediol	160	10571	Phenylphosphonic acid
57	322	o-Xylene	161	11004	D-Galactose
58	367	2,3-Dimethylbutane	162	11201	Sodium caprylate
59	372	Acetophenone	163	11202	Sodium dodecyl sulfate
60	425	Benzoic acid	164	11722	L-Threonine
61	430	Methyl isopropyl ketone	165	12706	D-Glucose
62	516	Hexadecane	166	13599	L-Lysine
63	546	Monoethanolamine	167	16447	1-Ethyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
64	598	Trifluoroacetic acid	168	16583	1-Butyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
65	750	p-Chlorotoluene	169	16584	1-Ethyl-3-methylimidazolium ethylsulfate
66	766	1,2-Dihydroxybenzene	170	16731	Cadmium perchlorate
67	809	2-Methoxyphenol	171	17118	(-)-Epicatechin
68	810	o-Chlorophenol	172	17231	Calcium-L-lactate
69	812	p-Chlorophenol	173	17273	D-(-)-Arabinose
70	817	1,3-Dihydroxybenzene	174	17617	tert-Butan(ol-D)
71	894	2,2"-Diethanolamine (DEA)	175	18690	Monosodium glutamate
72	925	Anthracene	176	18840	Lysozyme
73	1050	Carbon dioxide	177	18842	L-3,4-Dihydroxyphenylalanine
74	1051	Methane	178	18845	1-Butyl-3-methylimidazolium methylsulfate
75	1052	Oxygen	179	18857	Monosodium L-aspartate
76	1053	Ethylene	180	19687	L-Arginine
77	1054	Ethane	181	20036	1-Butyl-3-methylimidazolium octyl sulfate
78	1055	Propylene	182	20046	alpha-Cyclodextrin
79	1056	Nitrogen	183	20047	beta-Cyclodextrin
80	1058	Argon	184	22696	5-Hydroxymethylfurfural
81	1059	Chlorine	185	23228	Isoquercitrin
82	1060	Krypton	186	23325	(.+-.)-.beta.-Aminobutyric acid
83	1061	Dinitrogen monoxide	187	26695	[EMIM] methylsulfate
84	1062	Xenon	188	26828	Platinum (II) acetylacetonate
85	1063	Hydrogen	189	33333	Gallic acid monohydrate
86	1064	Ethyne	190	33334	(+)-Catechin hydrate
87	1065	Hydrogen sulfide	191	33340	Peonidin-3-glucoside chloride
88	1086	2,2,2-Trifluoroethanol	192	33341	Malvidin-3,5-diglucoside chloride
89	1090	2,2-Dimethylpentane	193	34501	2-Hydroxypropyl-beta-cyclodextrin
90	1143	1,3-Butanediol	194	34550	1,8-Bis(trimethylammonium)octane dibromide

Continued on next page

Continued from previous page

Cons. No.	DDB No.	Name	Cons. No.	DDB No.	Name
91	1264	alpha-Aminotoluene	195	34551	1,10-Bis(trimethylammonium)decane dibromide
92	1292	Helium	196	34552	1,12-Bis(trimethylammonium)dodecane dibromide
93	1293	Neon	197	36721	o-Sulfanilic acid
94	1594	Pyrene	198	37864	2-Hydroxypropyl-alpha-cyclodextrin
95	1642	1,4-Dihydroxybenzene	199	40775	DL-m-Tyrosine
96	1645	1,2,3-Trihydroxybenzene	200	40777	DL-o-Tyrosine
97	2186	Diisopropanolamine	201	40779	D,L-beta-Aminoisobutyric acid
98	2187	Methyldiethanolamine	202	43996	m-Sulfanilic acid
99	2245	Phosphoric acid	203	46014	p-Phenolsulfonic acid
100	2501	1,2-Benzenediamine	204	49211	beta-Cyclodextrin, sulfated sodium salt
101	2506	3-Methoxyphenol	205	51976	Lithium acetylacetonate
102	2542	Perylene	206	54011	N-Methylphenothiazine
103	2945	3-Hydroxyaniline	207	54491	L-Histidine methyl ester dihydrochloride
104	2994	DL-Tyrosine	208	61801	Tetrasodium tetraphenylporphyrintetrasulfonate
			Solvents		
1	3	Acetonitrile	27	161	Toluene
2	4	Acetone	28	174	Water
3	11	Ethanol	29	250	Cyclohexanone
4	12	Diethyl ether	30	282	1,2-Propanediol
5	21	Ethyl acetate	31	284	N-Methyl-2-pyrrolidone
6	25	Ethylbenzene	32	297	Hexafluorobenzene
7	26	Bromobenzene	33	367	2,3-Dimethylbutane
8	27	Chlorobenzene	34	430	Methyl isopropyl ketone
9	30	Nitrobenzene	35	451	Carbonic acid dimethyl ester
10	31	Benzene	36	516	Hexadecane
11	39	1-Butanol	37	887	Deuterium oxide
12	40	2-Butanone	38	982	Perdeuteromethanol
13	46	Butyl chloride	39	1090	2,2-Dimethylpentane
14	47	Chloroform	40	3410	1,3,5-Triisopropylbenzene
15	50	Cyclohexane	41	4331	Hexamethyltetracosane
16	60	Decane	42	16447	1-Ethyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
17	72	N,N-Dimethylformamide	43	16583	1-Butyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
18	78	Dodecane	44	16810	1-Octyl-3-methylimidazolium tetrafluoroborate
19	80	Butyl acetate	45	18162	1-Hexyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
20	89	Hexane	46	18174	1-Hexyl-3-methylimidazolium tetrafluoroborate
21	91	Heptane	47	18642	1-Ethyl-3-methylimidazolium bis(pentafluoroethylsulfonyl)imide
22	97	2,2,4-Trimethylpentane	48	18988	1-Ethyl-3-methylimidazolium trifluoromethylsulfonate
23	110	Methanol	49	20138	1-Butyl-3-methylimidazolium dicyanamide
24	112	3-Methylpentane	50	22417	1-Ethyl-3-methylimidazolium trifluoroacetate
25	140	1-Propanol	51	22674	1-Butyl-3-methylpyridinium tetrafluoroborate
26	157	Tetrachloromethane			

Table S.2: Latent variables u_i of the solutes for both hybrid MCMs, for the data set of the reduced data base.

		MCM-Boosting		MCM-Whisky	
i	**Name**	u_{i1}	u_{i2}	u_{i1}	u_{i2}
1	Acetonitrile	0.0259	−0.3137	1.1140	1.0072
2	Acetone	−0.0383	−0.6594	0.8325	1.0875
3	Ethanol	−0.1208	−0.2232	1.2558	0.6697
4	Ethyl acetate	0.0183	−1.0212	1.2218	0.3069
5	Benzyl alcohol	−0.0248	0.0109	0.9067	−0.2790
6	Ethylbenzene	0.0853	−0.3729	1.0475	−0.0265
7	Chlorobenzene	0.0830	−0.6592	1.0640	−0.0972
8	Benzene	0.0225	−0.7974	1.0589	0.6732
9	1-Butanol	0.0652	0.5200	0.8152	−0.3123
10	Butyl chloride	−0.1005	−1.3470	0.9289	1.0876
11	3-Methylphenol	0.0168	0.0857	0.9696	−0.3696
12	Cyclohexane	−0.0247	−0.1559	0.9850	0.0256
13	Dodecane	−0.1310	0.4674	0.7540	−0.2834
14	Benzaldehyde	−0.0026	−0.3415	0.9643	0.0233
15	Butyl acetate	0.0583	−0.5567	1.0636	−0.1444
16	Acetic acid	0.1207	0.3451	1.0522	0.0817
17	Hexane	−0.0562	0.4757	1.0284	0.1294
18	Heptane	−0.1007	−0.3165	1.0001	−0.0528
19	Methyl iodide	0.0453	−1.8181	1.1878	0.5537
20	Methanol	−0.0717	0.6005	1.2734	0.8127
21	Naphthalene	0.1071	−0.6001	1.0165	−0.0713
22	Phenol	0.1714	−0.1511	1.1026	−0.1210
23	1-Propanol	0.0338	0.4886	1.0087	−0.1563
24	Propionic acid	0.0481	0.1907	1.0585	0.0166
25	Tetrachloromethane	0.0657	−0.7153	0.9538	−0.2629
26	Toluene	0.0294	−0.4151	1.0385	0.1623
27	Water	0.0160	1.0622	1.2625	1.2361
28	Glycerol	−0.0547	0.2778	0.9945	−0.4034
29	Butyric acid	0.0493	0.1375	0.9714	−0.0868
30	Hexafluorobenzene	0.0554	−1.2698	1.0224	−0.0083
31	2-Methyl-2,4-pentanediol	0.1493	−0.1078	0.7891	−0.4680
32	Acetophenone	0.0366	−0.2193	0.8595	−0.0045
33	Methyl isopropyl ketone	−0.0220	−0.3874	1.0480	0.2781
34	Hexadecane	−0.0178	0.6118	0.6210	−0.4563
35	p-Chlorotoluene	0.1212	−0.5070	1.0570	−0.1735
36	1,2-Dihydroxybenzene	0.0735	0.3086	0.8749	−0.5178
37	p-Chlorophenol	0.0550	−0.0153	0.9082	−0.4891
38	1,3-Dihydroxybenzene	0.1110	0.5870	1.0430	−0.8759
39	Carbon dioxide	0.0024	−2.1624	1.0677	2.6978
40	Pyrene	0.0411	−0.3815	0.7835	−0.4026
41	1,4-Dihydroxybenzene	−0.0929	0.7952	1.0314	−0.9901
42	1,2,3-Trihydroxybenzene	0.0466	0.5331	0.8596	−0.8712
43	Perylene	−0.0153	−0.2712	0.6542	−0.6323
44	3-Hydroxyaniline	0.0221	0.5058	0.9982	−0.6519
45	Di-tert-butylsulfide	−0.0455	−0.5260	0.8416	0.0409

Table S.3: Latent variables v_j of the solvents for both hybrid MCMs, for the data set of the reduced data base.

		MCM-Boosting		MCM-Whisky	
j	**Name**	v_{j1}	v_{j2}	v_{j1}	v_{j2}
1	Acetonitrile	0.0808	0.1552	1.0329	0.3479
2	Acetone	−0.0143	0.3749	1.2012	0.0947
3	Ethanol	0.0235	0.4952	−0.0895	0.4689
4	Ethyl acetate	0.0245	−0.4662	0.5537	0.1355
5	Benzene	0.0041	0.3100	0.7094	0.2268
6	1-Butanol	−0.0171	0.6077	−0.7692	0.4649
7	Butyl chloride	−0.0616	0.0769	1.0076	0.0935
8	Chloroform	−0.1435	0.0860	0.7535	0.2162
9	Cyclohexane	−0.0421	0.4390	0.3997	0.2822
10	Dodecane	0.1043	−0.6007	0.1037	0.2835
11	Butyl acetate	−0.0016	0.3710	0.5711	0.2338
12	Hexane	−0.0297	0.0871	1.2860	0.0933
13	Heptane	−0.0449	0.1479	1.1070	0.1109
14	Methanol	−0.0325	0.3837	0.5793	0.4010
15	1-Propanol	0.0604	0.5545	−0.6209	0.5466
16	Tetrachloromethane	0.0471	−0.0292	0.2200	0.2178
17	Toluene	−0.0317	0.0448	0.7320	0.1470
18	Water	0.0019	0.1685	0.0015	0.2546
19	1,2-Propanediol	−0.0228	0.4908	−3.1392	0.5874
20	N-Methyl-2-pyrrolidone	−0.0056	0.2545	−0.4168	0.3778
21	Hexafluorobenzene	−0.0188	−0.0455	0.3585	0.1540
22	Methyl isopropyl ketone	−0.0201	0.3742	0.7499	0.1693
23	Hexadecane	−0.0470	−1.1143	−0.3064	1.1473

Table S.4: Table of the full data base on D_{ij}^{∞} developed in this work. The numerical values were determined as described in Section 2. In cases where a value was adopted directly from the Dortmund Data Base [13] and can thus be found therein, it is simply listed as "DDB".

Solute	**Solvent**	$D_{ij}^{\infty} \times 10^9$	**Solute**	**Solvent**	$D_{ij}^{\infty} \times 10^9$
DDB No.	DDB No.	m^2/s	DDB No.	DDB No.	m^2/s
2	174	1.25	39	110	1.30
3	161	2.88	39	140	0.53
3	174	DDB	39	174	DDB
4	31	2.76	40	157	1.67
4	39	DDB	41	174	DDB
4	47	2.33	46	91	3.43
4	50	2.22	46	140	DDB
4	110	DDB	47	4	3.63
4	140	DDB	47	12	DDB
4	157	1.71	49	11	DDB
4	174	DDB	49	110	DDB
8	174	DDB	49	140	DDB
11	31	DDB	49	174	DDB
11	174	1.23	50	4	3.84
12	47	DDB	50	31	DDB
15	174	1.51	50	91	DDB

Continued on next page

Continued from previous page

Solute	**Solvent**	$D_{ij}^{\infty} \times 10^9$	**Solute**	**Solvent**	$D_{ij}^{\infty} \times 10^9$
DDB No.	DDB No.	m^2/s	DDB No.	DDB No.	m^2/s
17	174	DDB	50	157	1.28
21	31	3.17	50	161	DDB
21	157	1.46	50	297	1.59
24	11	DDB	72	31	DDB
24	110	DDB	77	174	0.71
24	140	DDB	78	89	DDB
24	174	DDB	78	516	DDB
25	89	DDB	79	11	DDB
25	91	3.08	79	110	DDB
26	27	1.62	79	174	DDB
27	26	1.32	80	174	DDB
27	89	DDB	80	430	2.56
27	91	DDB	84	11	0.96
30	89	3.93	84	110	1.83
30	97	1.78	84	140	DDB
31	4	DDB	84	157	1.34
31	11	DDB	84	174	1.21
31	21	1.88	85	174	DDB
31	47	DDB	89	30	0.86
31	50	DDB	89	31	2.09
31	60	DDB	89	78	1.40
31	72	DDB	89	91	3.18
31	89	DDB	89	161	2.27
31	91	3.83	89	516	0.85
31	110	DDB	91	25	DDB
31	157	DDB	91	31	DDB
31	161	DDB	91	46	2.65
31	174	DDB	91	50	DDB
31	282	DDB	91	89	3.72
31	297	1.54	91	112	3.75
39	4	DDB	91	161	DDB
91	250	DDB	174	284	DDB
91	367	3.34	174	430	3.26
91	1090	2.74	174	516	DDB
99	110	DDB	174	4331	DDB
99	174	DDB	174	16583	0.39
108	110	DDB	174	20138	2.08
110	39	0.48	230	174	0.95
110	140	0.69	230	284	DDB
110	174	DDB	235	11	DDB
112	91	3.06	235	110	DDB
123	89	DDB	235	140	DDB
123	91	DDB	235	174	0.96
123	110	DDB	237	174	DDB
123	174	DDB	250	91	DDB
129	18162	DDB	269	174	1.56
138	11	DDB	284	174	DDB
138	110	DDB	297	31	2.27
138	140	DDB	297	50	1.63

Continued on next page

Continued from previous page

Solute	**Solvent**	$D_{ij}^{\infty} \times 10^9$	**Solute**	**Solvent**	$D_{ij}^{\infty} \times 10^9$
DDB No.	DDB No.	m^2/s	DDB No.	DDB No.	m^2/s
138	174	DDB	297	3410	0.41
138	282	DDB	308	174	DDB
140	4	DDB	308	284	DDB
140	39	0.41	322	91	DDB
140	46	DDB	367	91	2.94
140	110	1.36	372	11	DDB
140	174	1.07	372	110	DDB
141	11	DDB	425	174	1.12
141	110	DDB	430	80	2.15
141	140	DDB	430	174	1.17
141	174	1.06	516	78	0.94
145	174	DDB	516	89	2.20
146	174	3.00	546	174	1.09
147	982	1.92	598	157	1.33
153	174	0.92	750	89	DDB
157	4	3.59	750	91	DDB
157	21	1.47	766	11	DDB
157	40	3.19	766	110	DDB
157	50	DDB	766	140	DDB
161	3	3.33	766	174	DDB
161	31	DDB	809	174	DDB
161	50	DDB	810	174	DDB
161	89	3.95	812	11	DDB
161	91	3.35	812	110	DDB
161	110	DDB	812	140	DDB
168	174	DDB	812	174	DDB
174	3	DDB	817	11	DDB
174	4	DDB	817	110	DDB
174	11	1.20	817	140	DDB
174	80	2.83	817	174	DDB
174	140	0.74	817	282	DDB
894	174	0.82	2542	89	DDB
925	110	DDB	2542	91	DDB
1050	11	DDB	2945	11	DDB
1050	110	6.20	2945	110	DDB
1050	140	DDB	2945	140	DDB
1050	174	DDB	2945	174	DDB
1050	16447	DDB	2994	174	DDB
1050	16810	DDB	3063	174	1.13
1050	18162	0.37	3215	174	DDB
1050	18174	DDB	3258	174	0.77
1050	18642	DDB	3347	174	DDB
1050	18988	DDB	3410	297	1.23
1050	22417	DDB	3468	174	DDB
1050	22674	DDB	3523	174	DDB
1051	174	1.92	3715	174	DDB
1052	174	DDB	3717	174	DDB
1053	174	DDB	3724	174	0.93
1054	174	DDB	3725	174	0.92

Continued on next page

Continued from previous page

Solute	**Solvent**	$D_{ij}^{\infty} \times 10^9$	**Solute**	**Solvent**	$D_{ij}^{\infty} \times 10^9$
DDB No.	DDB No.	m^2/s	DDB No.	DDB No.	m^2/s
1055	174	DDB	3729	174	1.06
1056	174	1.99	3731	174	0.83
1058	174	DDB	3732	174	0.74
1059	174	DDB	3865	174	0.89
1060	174	DDB	3988	174	DDB
1061	174	1.76	3989	174	0.78
1062	174	DDB	3990	174	DDB
1063	174	2.91	3991	174	DDB
1064	174	DDB	4490	174	DDB
1065	174	2.07	4577	174	1.95
1086	174	1.08	4591	174	DDB
1090	91	3.09	4592	174	1.26
1143	174	DDB	4596	3	DDB
1264	174	DDB	4707	174	DDB
1292	174	6.76	4708	174	DDB
1293	174	DDB	4771	161	DDB
1594	89	DDB	4776	982	1.59
1594	91	DDB	4792	174	1.44
1594	110	DDB	4795	174	DDB
1642	11	DDB	4801	174	DDB
1642	140	DDB	4817	174	DDB
1642	174	DDB	4911	174	1.32
1645	11	DDB	4955	174	1.16
1645	110	DDB	4960	174	DDB
1645	140	DDB	4965	174	DDB
1645	174	DDB	5261	887	0.82
2186	174	DDB	5949	174	DDB
2187	174	DDB	6317	174	1.57
2245	174	1.09	6319	174	DDB
2501	174	DDB	6325	174	DDB
2506	174	DDB	6326	174	DDB
6353	174	1.48	18845	174	DDB
6355	174	DDB	18857	174	0.94
6372	174	1.48	19687	174	0.74
6465	174	0.67	20036	174	DDB
6529	110	DDB	20046	174	0.35
6529	174	DDB	20047	174	0.33
7467	157	0.97	22696	174	DDB
7533	887	0.40	23228	174	DDB
7847	174	0.74	23325	174	DDB
7848	174	0.77	26695	174	DDB
7852	110	DDB	26828	11	DDB
7949	174	DDB	33333	174	DDB
9329	174	DDB	33334	174	DDB
10334	11	DDB	33340	174	DDB
10571	174	DDB	33341	174	DDB
11004	174	DDB	34501	174	0.32
11201	174	0.86	34550	174	DDB
11202	174	0.49	34551	174	DDB

Continued on next page

Continued from previous page

Solute	**Solvent**	$D_{ij}^{\infty} \times 10^9$	**Solute**	**Solvent**	$D_{ij}^{\infty} \times 10^9$
DDB No.	DDB No.	m^2/s	DDB No.	DDB No.	m^2/s
11722	174	0.77	34552	174	DDB
12706	174	DDB	36721	174	DDB
13599	174	0.67	37864	174	0.35
16447	174	DDB	40775	174	DDB
16583	174	DDB	40777	174	DDB
16584	174	DDB	40779	174	DDB
16731	174	DDB	43996	174	DDB
17118	174	DDB	46014	174	DDB
17231	174	0.63	49211	174	0.70
17273	174	DDB	51976	11	1.24
17617	887	0.65	54011	451	1.26
18690	174	0.89	54491	174	1.05
18840	174	DDB	61801	174	0.62
18842	174	0.61			

Table S.5: Table of the reduced data base on D_{ij}^{∞} developed in this work. The numerical values were determined as described in Section 3.2.4. In cases where a value was adopted directly from the Dortmund Data Base [13] and can thus be found therein, it is simply listed as "DDB".

Solute	**Solvent**	$D_{ij}^{\infty} \times 10^9$	**Solute**	**Solvent**	$D_{ij}^{\infty} \times 10^9$
DDB No.	DDB No.	m^2/s	DDB No.	DDB No.	m^2/s
3	161	2.88	50	91	DDB
3	174	DDB	50	157	1.28
4	31	2.76	50	161	DDB
4	39	DDB	50	297	1.59
4	47	2.33	78	89	DDB
4	50	2.22	78	516	DDB
4	110	DDB	79	11	DDB
4	140	DDB	79	110	DDB
4	157	1.71	79	174	DDB
4	174	DDB	80	174	DDB
11	31	DDB	80	430	2.56
11	174	1.23	84	11	0.96
21	31	3.17	84	110	1.83
21	157	1.46	84	140	DDB
24	11	DDB	84	157	1.34
24	110	DDB	84	174	1.21
24	140	DDB	89	31	2.09
24	174	DDB	89	78	1.40
25	89	DDB	89	91	3.18
25	91	3.08	89	161	2.27
27	89	DDB	89	516	0.85
27	91	DDB	91	31	DDB
31	4	DDB	91	46	2.65
31	11	DDB	91	50	DDB
31	21	1.88	91	89	3.72

Continued on next page

Continued from previous page

Solute	**Solvent**	$D_{ij}^{\infty} \times 10^9$	**Solute**	**Solvent**	$D_{ij}^{\infty} \times 10^9$
DDB No.	DDB No.	m^2/s	DDB No.	DDB No.	m^2/s
31	47	DDB	91	161	DDB
31	50	DDB	99	110	DDB
31	89	DDB	99	174	DDB
31	91	3.83	110	39	0.48
31	110	DDB	110	140	0.69
31	157	DDB	110	174	DDB
31	161	DDB	123	89	DDB
31	174	DDB	123	91	DDB
31	282	DDB	123	110	DDB
31	297	1.54	123	174	DDB
39	4	DDB	138	11	DDB
39	110	1.30	138	110	DDB
39	140	0.53	138	140	DDB
39	174	DDB	138	174	DDB
46	91	3.43	138	282	DDB
46	140	DDB	140	4	DDB
49	11	DDB	140	39	0.41
49	110	DDB	140	46	DDB
49	140	DDB	140	110	1.36
49	174	DDB	140	174	1.07
50	4	3.84	141	11	DDB
50	31	DDB	141	110	DDB
141	140	DDB	750	91	DDB
141	174	1.06	766	11	DDB
157	4	3.59	766	110	DDB
157	21	1.47	766	140	DDB
157	50	DDB	766	174	DDB
161	3	3.33	812	11	DDB
161	31	DDB	812	110	DDB
161	50	DDB	812	140	DDB
161	89	3.95	812	174	DDB
161	91	3.35	817	11	DDB
161	110	DDB	817	110	DDB
174	3	DDB	817	140	DDB
174	4	DDB	817	174	DDB
174	11	1.20	817	282	DDB
174	80	2.83	1050	11	DDB
174	140	0.74	1050	110	6.20
174	284	DDB	1050	140	DDB
174	430	3.26	1050	174	DDB
174	516	DDB	1594	89	DDB
230	174	0.95	1594	91	DDB
230	284	DDB	1594	110	DDB
235	11	DDB	1642	11	DDB
235	110	DDB	1642	140	DDB
235	140	DDB	1642	174	DDB
235	174	0.96	1645	11	DDB
297	31	2.27	1645	110	DDB
297	50	1.63	1645	140	DDB

Continued on next page

Continued from previous page

Solute	**Solvent**	$D_{ij}^{\infty} \times 10^9$	**Solute**	**Solvent**	$D_{ij}^{\infty} \times 10^9$
DDB No.	DDB No.	m^2/s	DDB No.	DDB No.	m^2/s
308	174	DDB	1645	174	DDB
308	284	DDB	2542	89	DDB
372	11	DDB	2542	91	DDB
372	110	DDB	2945	11	DDB
430	80	2.15	2945	110	DDB
430	174	1.17	2945	140	DDB
516	78	0.94	2945	174	DDB
516	89	2.20	6529	110	DDB
750	89	DDB	6529	174	DDB

S.7 Stan code

In the following, we provide the Stan codes for the training of all MCMs used in this work: the data-driven MCM, MCM-Boosting, and MCM-Whisky. For MCM-Whisky, the codes of the two training steps, distillation and maturation, are given individually. An executable form of this code is included for download in the form of .stan files. To run the code, users will need to install an interface of their choice from the project's homepage (`https://mc-stan.org/users/interfaces/`). For further information, we refer to Stan's excellent documentation: `https://mc-stan.org/users/documentation/`.

Furthermore, we have included a wrapper code for each MCM, i.e., a MATLAB script that reads the training data from a .csv file, applies the developed MCMs for the prediction of the full matrix, and exports the result to a .csv file.

S.7.1 Data-driven MCM

```
data {
  int<lower=0> I; // number of solutes
  int<lower=0> J; // number of solvents
  int<lower=0> K; // number of latent dimensions
  real ln_D[I,J]; // matrix of logarithmic diffusion
      coefficients
  real<lower=0> sigma_0; // prior standard deviation
  real<lower=0> lambda; // likelihood scale
}

parameters {
  vector[K] u[I]; // solute feature vectors
  vector[K] v[J]; // solvent feature vectors
}

```

```
15 model {
16   // prior: draw feature vectors for all solutes and
         solvents:
17   for (i in 1:I)
18     u[i] ~ normal(0,sigma_0);
19   for (j in 1:J)
20     v[j] ~ normal(0,sigma_0);
21   // likelihood: model the probability of ln_D as a
        normal distribution
22   // around the dot product of the feature vectors:
23   for (i in 1:I) {
24     for (j in 1:J) {
25       if (ln_D[i,j] != -99) { // train to available
            data only
26         ln_D[i,j] ~ normal(u[i]' * v[j], lambda);
27       }
28     }
29   }
30 }
```

S.7.2 MCM-Boosting

```
 1 data {
 2   int<lower=0> I; // number of solutes
 3   int<lower=0> J; // number of solvents
 4   int<lower=0> K; // number of latent dimensions
 5   real R[I,J]; // matrix of residuals of logarithmic
         diffusion coefficients
 6   real<lower=0> sigma_0; // prior standard deviation
 7   real<lower=0> lambda; // likelihood scale
 8 }
 9
10 parameters {
11   vector[K] u[I]; // solute feature vectors
12   vector[K] v[J]; // solvent feature vectors
13 }
14
15 model {
16   // prior: draw feature vectors for all solutes and
         solvents:
17   for (i in 1:I)
18     u[i] ~ normal(0,sigma_0);
19   for (j in 1:J)
20     v[j] ~ normal(0,sigma_0);
```

```
21   // likelihood: model the probability of R as a
        normal distribution around the dot product of
        the feature vectors:
22   for (i in 1:I) {
23     for (j in 1:J) {
24       if (R[i,j] != -99) { // train to available
            data only
25         R[i,j] ~ normal(u[i]' * v[j], lambda);
26       }
27     }
28   }
29 }
```

S.7.3 MCM-Whisky: Distillation

```
 1 data {
 2   int<lower=0> I; // number of solutes
 3   int<lower=0> J; // number of solvents
 4   int<lower=0> K; // number of latent dimensions
 5   real ln_D[I,J]; // matrix of logarithmic diffusion
        coefficients
 6   real<lower=0> sigma_0; // prior standard deviation
 7   real<lower=0> lambda; // likelihood scale
 8 }
 9
10 parameters {
11   vector[K] u[I]; // solute feature vectors
12   vector[K] v[J]; // solvent feature vectors
13 }
14
15 model {
16   // prior: draw feature vectors for all solutes and
        solvents:
17   for (i in 1:I)
18     u[i] ~ normal(0,sigma_0);
19   for (j in 1:J)
20     v[j] ~ normal(0,sigma_0);
21   // likelihood: model the probability of ln_D as a
        normal distribution around the dot product of
        the feature vectors:
22   for (i in 1:I) {
23     for (j in 1:J) {
24       if (ln_D[i,j] != -99) { // train to available
            data only
25         ln_D[i,j] ~ cauchy(u[i]' * v[j], lambda);
```

```
      }
    }
  }
}
```

S.7.4 MCM-Whisky: Maturation

```
data {
  int<lower=0> I; // number of solutes
  int<lower=0> J; // number of solvents
  int<lower=0> K; // number of latent dimensions
  real ln_D[I,J]; // matrix of logarithmic diffusion
      coefficients
  real<lower=0> lambda; // likelihood scale
  vector<lower=0>[K] sigma_0_u[I]; // Prior standard
      deviation (Solutes)
  vector<lower=0>[K] sigma_0_v[J]; // Prior standard
      deviation (Solvents)
  vector[K] mu_0_u[I]; // prior mean (Solutes)
  vector[K] mu_0_v[J]; // prior mean (Solvents)
}

parameters {
  vector[K] u[I]; // solute feature vectors
  vector[K] v[J]; // solvent feature vectors
}

model {
  // prior: draw feature vectors for all solutes and
      solvents:
  for (i in 1:I)
    u[i] ~ normal(mu_0_u[i],sigma_0_u[i]);
  for (j in 1:J)
    v[j] ~ normal(mu_0_v[j],sigma_0_v[j]);
  // likelihood: model the probability of ln_D as a
     normal distribution around the dot product of
     the feature vectors:
  for (i in 1:I) {
    for (j in 1:J) {
      if (ln_D[i,j] != -99) {//available data only
        ln_D[i,j] ~ normal(u[i]' * v[j], lambda);
      }
    }
  }
}
```

References

[1] R. L. Rowley, W. V. Wilding, J. L. Oscarson, Y. Yang, N. A. Zundel, T. E. Daubert and R. P. Danner, *DIPPR Data Compilation of Pure Chemical Properties*, Design Institute for Physical Properties, AIChE, 2003, `https://www.aiche.org/dippr`, Database date: 2018, retrieved via The DIPPR Information and Data Evaluation Manager for the Design Institute for Physical Properties - Version 12.3.0 (May 2018 Public).

[2] O. R. Quayle, *Chem. Rev.*, 1953, **53**, 439–589.

[3] J. R. Partington, *An Advanced Treatise on Physical Chemistry, Vol. I, Fundamental Principles: The Properties of Gases*, Longmans, London, 1949.

[4] C. R. Wilke and P. Chang, *AIChE J.*, 1955, **1**, 264–270.

[5] K. A. Reddy and L. K. Doraiswamy, *Ind. Eng. Chem. Fundam.*, 1967, **6**, 77–79.

[6] E. G. Scheibel, *Ind. Eng. Chem.*, 1954, **46**, 2007–2008.

[7] M. T. Tyn and W. F. Calus, *J. Chem. Eng. Data*, 1975, **20**, 106–109.

[8] B. E. Poling, J. M. Prausnitz and J. P. O'Connell, *The Properties of Gases and Liquids*, McGraw-Hill, New York, 2001.

[9] A. Einstein, *Ann. Phys.*, 1905, **322**, 549–560.

[10] A. Gierer and K. Wirtz, *Z. Naturforsch.*, 1953, **8**, 522–532.

[11] R. Evans, G. Dal Poggetto, M. Nilsson and G. A. Morris, *Anal. Chem.*, 2018, **90**, 3987–3994.

[12] D. R. Olander, *AIChE J.*, 1961, **7**, 175–176.

[13] *Dortmund Data Bank*, 2019, `www.ddbst.com`.