SUPPORTING INFORMATION

for

Quantum Chemically Estimated Abraham Solute Parameters Using Multiple Solvent-Water Partition Coefficients and Molecular Polarizability

Yuzhen Liang[ab], Ruichang Xiong[b], Stanley I. Sandler[b], Dominic M. Di Toro[b]*

[a] South China University of Technology, Guangzhou, Guangdong 510006, China
[b] University of Delaware, Newark, Delaware, US 19716

Corresponding author: Dr. Dominic M. Di Toro; Telephone: 1-302-831-4092; Fax: 1-302-831-3640; Email: dditoro@udel.edu

This file contains 33 pages, including 37 figures.

CONTENTS

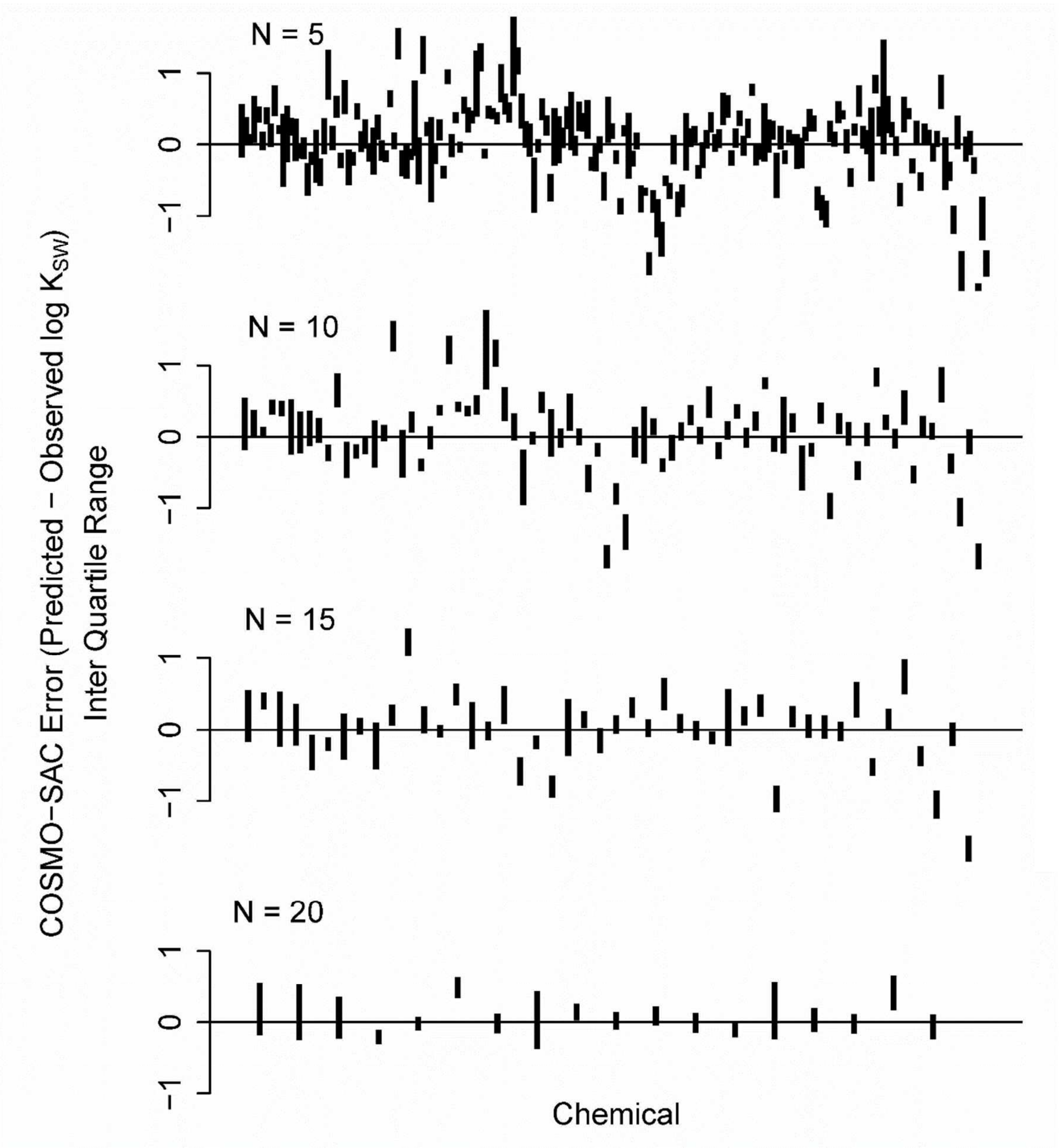

Figure S1 Plots showing the decrease of the interquartile range (IQR) of COSMO-SAC predicted solvent-water partition coefficient errors as the number of data points increases. N = minimum number of data points used to compute the IQR. Data are from Table S1. Chemicals are plotted in the solvent sequence in the table. Note that a solvent with greater than 20 observations appears in all four plots, e.g. the first IQR on the left hand side.

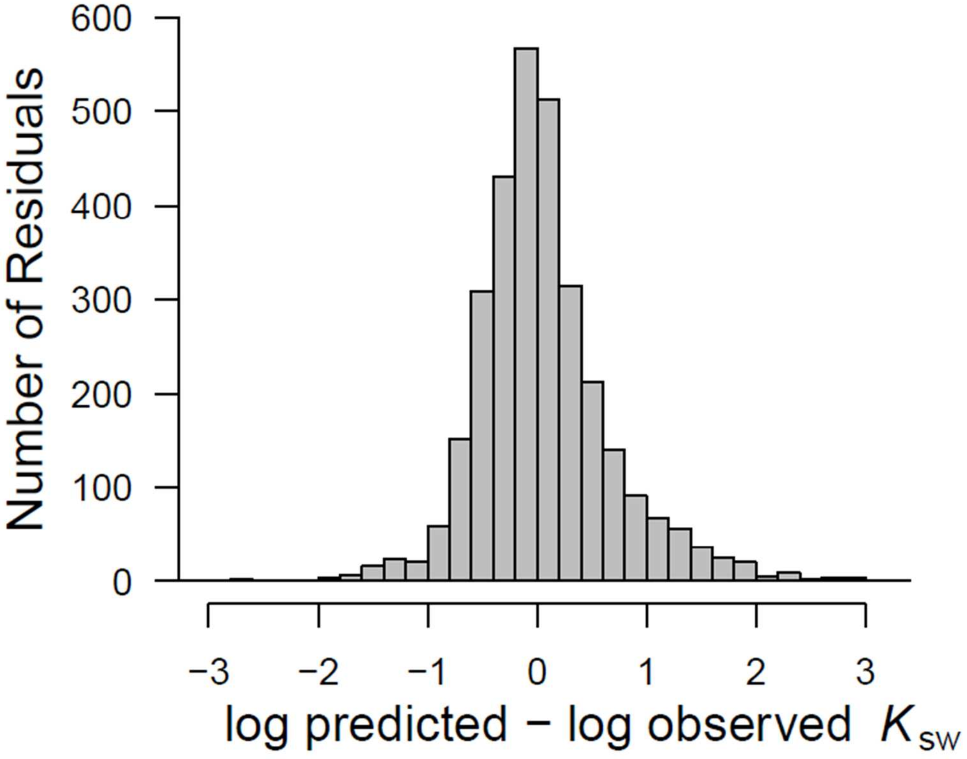

Figure S2 A histogram plot of COSMO-SAC predicted solvent-water partition coefficient (K_{sw}) errors. 24 solvent-water systems and 3095 data points are included.

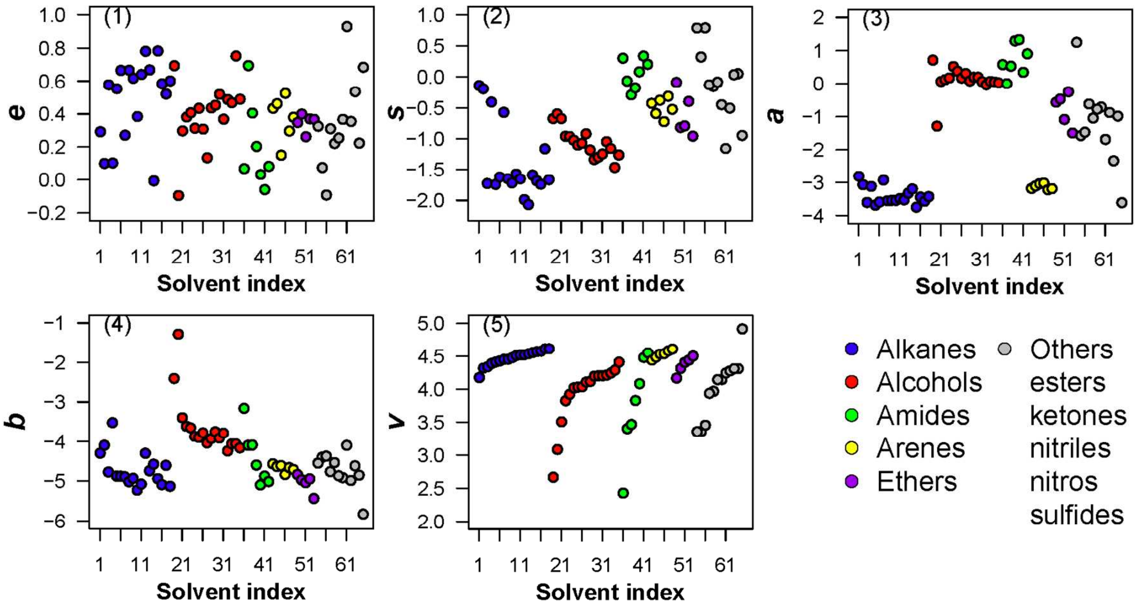

Figure S3 The solvent-water system parameters (1) *e*, (2) *s*, (3) *a*, (4) *b*, and (5) *v* for the 65 solvent-water systems used for computing QCAP. Parameters are grouped by solvent chemical classes. Within each group, the solvents are ordered from the smallest to the largest magnitude of *v*.

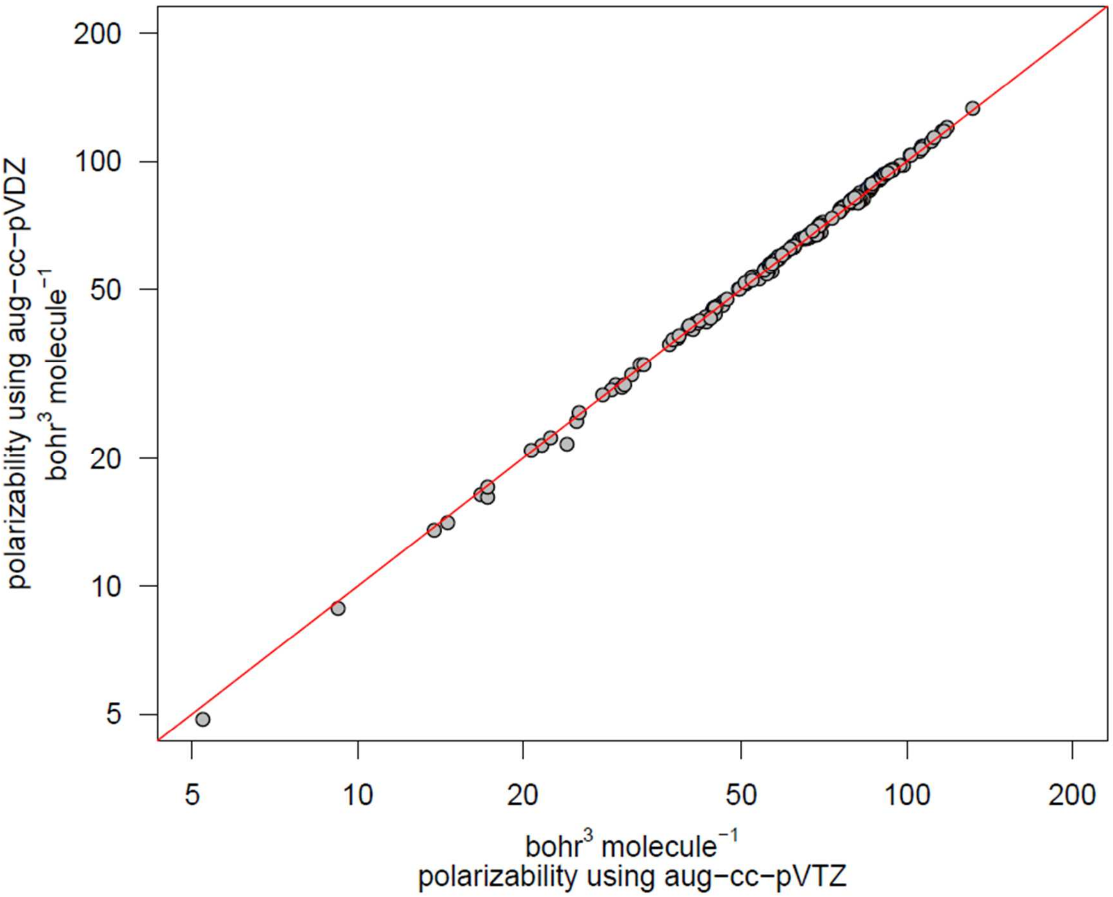

Figure S4. Molecular polarizability computed using M062X[1] with the aug-cc-pVDZ versus aug-cc-pVTZ basis set. A conversion factor of $1.48\times10^{-25}\times10^{-2}$ can be used to convert the units of borh^{3}molecule^{-1} to cm^{3}molecule^{-1}/100.

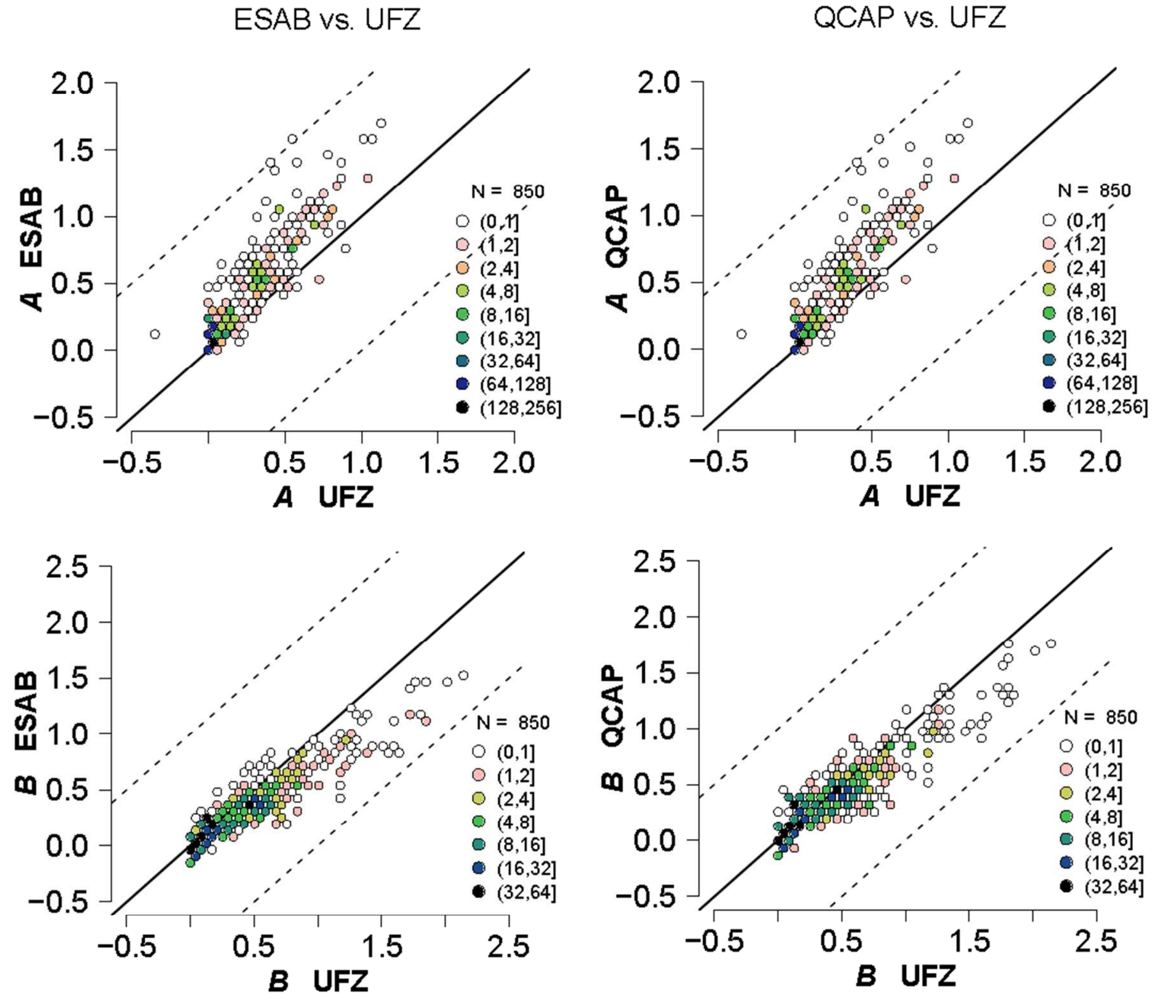

Figure S5 Hexbin two dimensional histogram plots[2] comparing the predicted versus experimentally-based solute parameters from the UFZ-LSER database (labeled UFZ) for *A*, *B*, and *V* determined by two methods listed at the top of each column: (Method I) estimate *V* independently from molecular volume and then estimate *E*, *S*, *A*, and *B* jointly with an MLR using COSMO-SAC estimated solvent-water partition coefficients, labeled ESAB; (Method II) estimate *E* from molecular polarizability, *V* from molecular volume and then *S*, *A*, and *B* from an MLR using COSMO-SAC estimated solvent-water partition coefficients, which is the QCAP method.

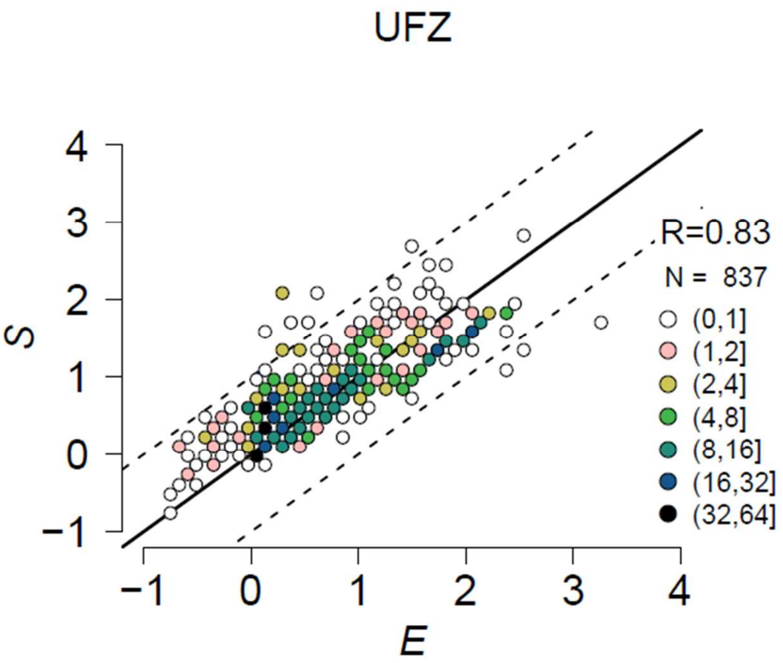

Figure S6 Hexbin two dimensional histogram plot[2] of *E* versus *S* solute parameters compiled from the UFZ-LSER database.

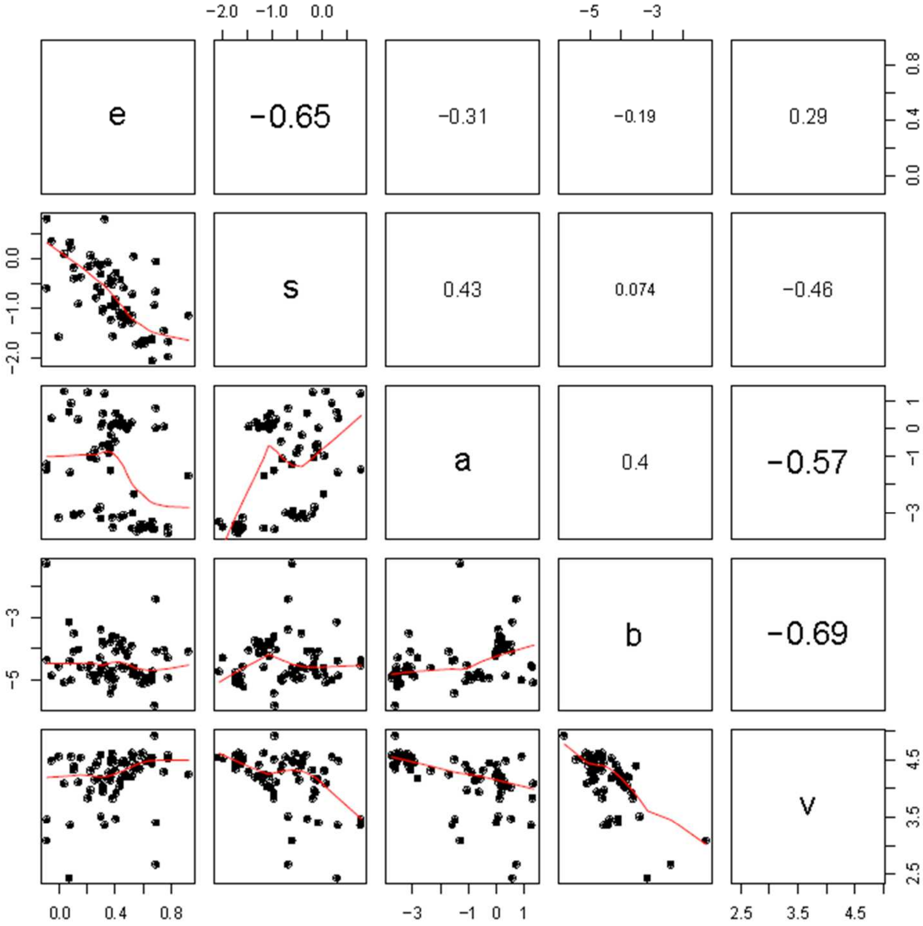

Figure S7 Pairs plot[3] of solvent parameters used to compute QCAP parameters with a loess curve plotted in the lower triangular. x axis label is the parameter above the plot, y axis label is to the right of the plot. The upper triangular panels display the Pearson coefficient (*r*) of the respective pair. For each *r* panel, x axis label is the parameter below the plot, y axis label is to the left of the plot.

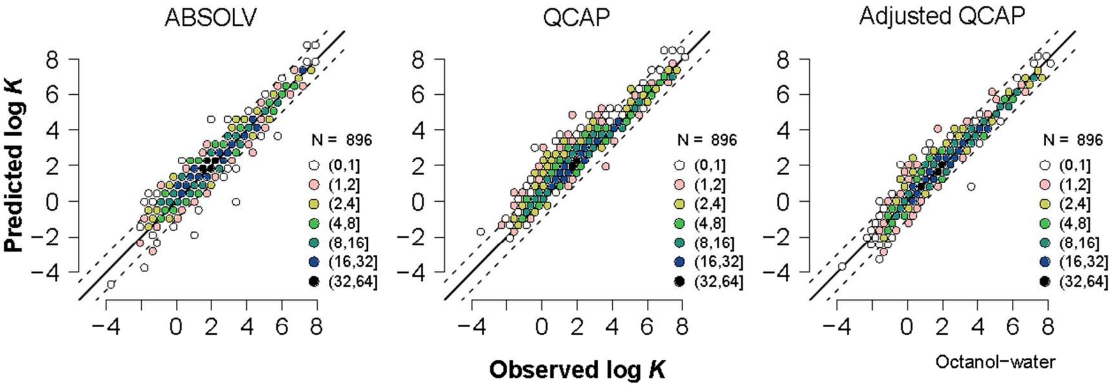

Figure S8 Hexbin two dimensional histogram plot[2] of predicted versus observed log wet octanol-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

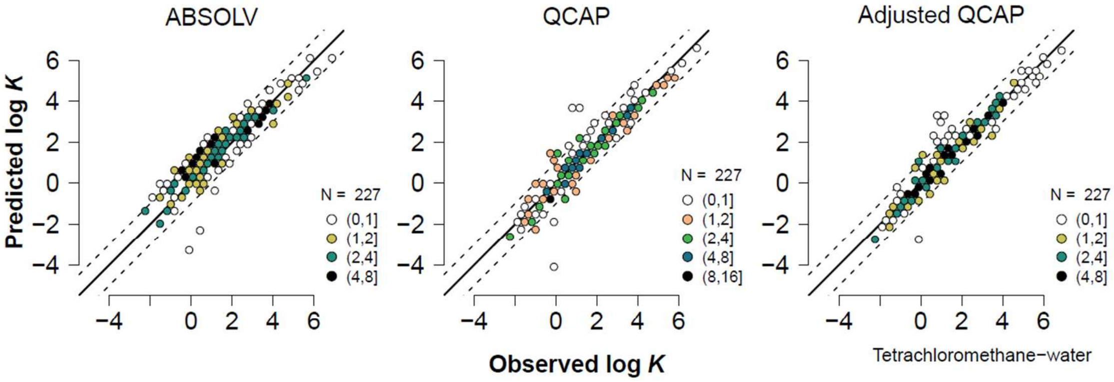

Figure S9 Hexbin two dimensional histogram plot[2] of predicted versus observed log tetrachloromethane-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

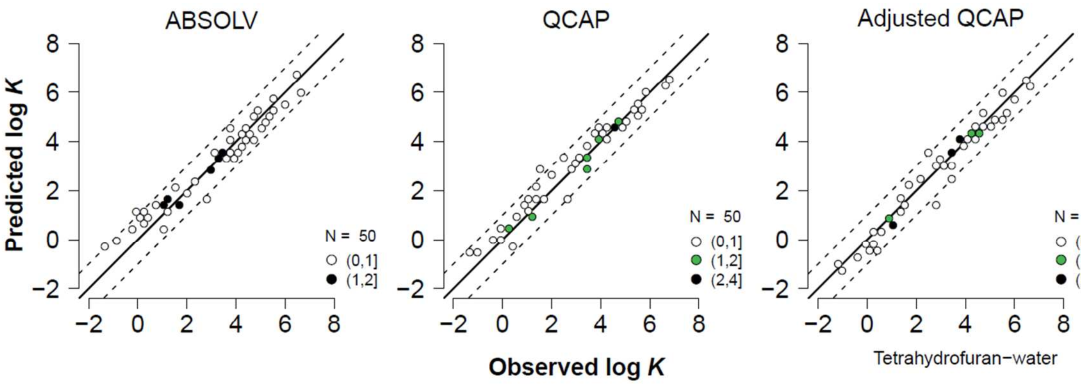

Figure S10 Hexbin two dimensional histogram plot[2] of predicted versus observed log tetrahydrofuran-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

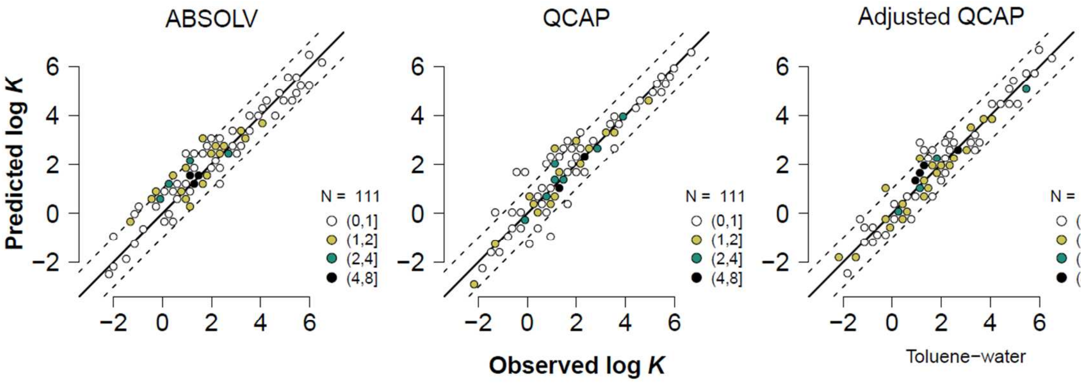

Figure S11 Hexbin two dimensional histogram plot[2] of predicted versus observed log toluene-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

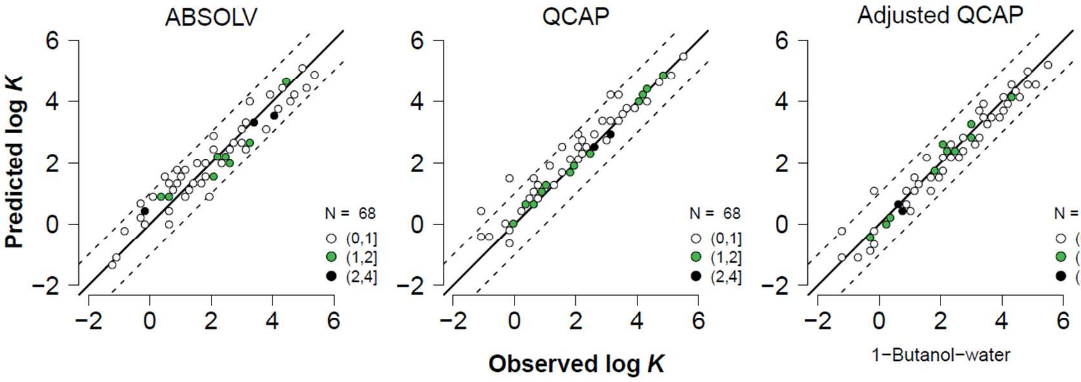

Figure S12 Hexbin two dimensional histogram plot[2] of predicted versus observed log 1-butanol-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

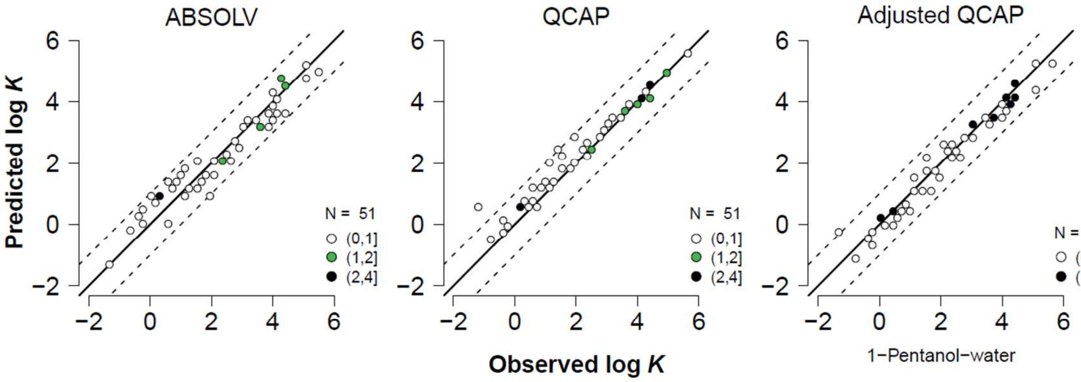

Figure S13 Hexbin two dimensional histogram plot[2] of predicted versus observed log 1-pentanol-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

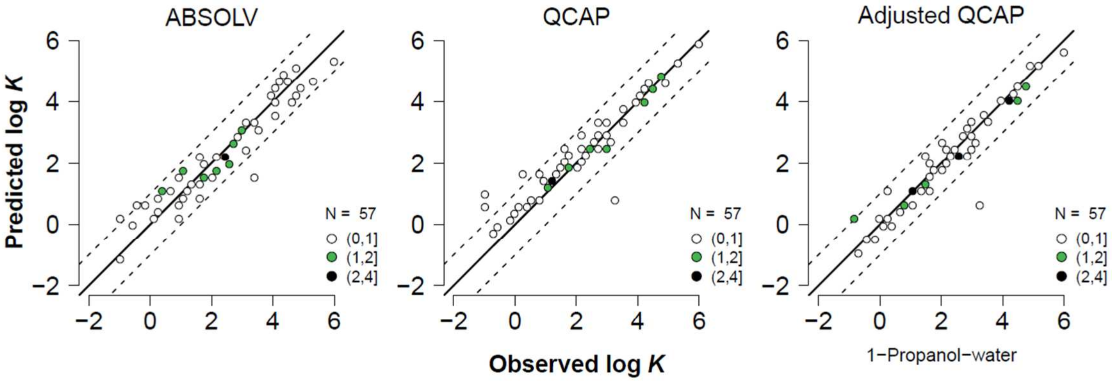

Figure S14 Hexbin two dimensional histogram plot[2] of predicted versus observed log 1-propanol-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

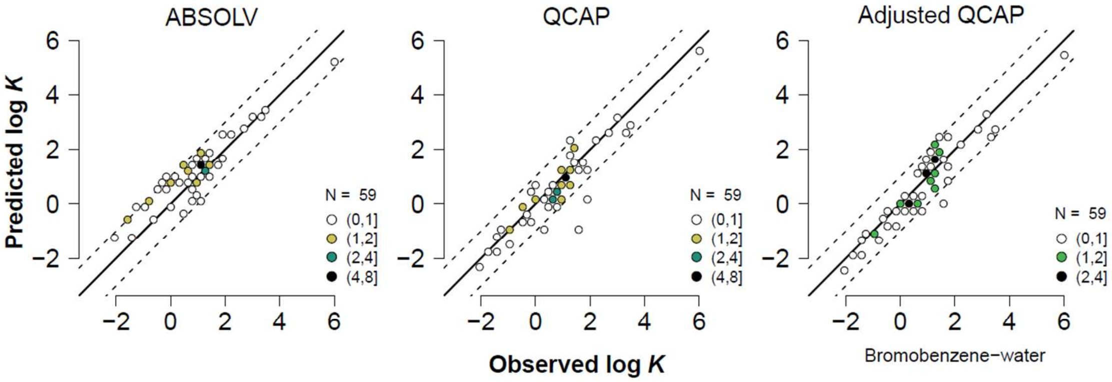

Figure S15 Hexbin two dimensional histogram plot[2] of predicted versus observed log bromobenzene-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

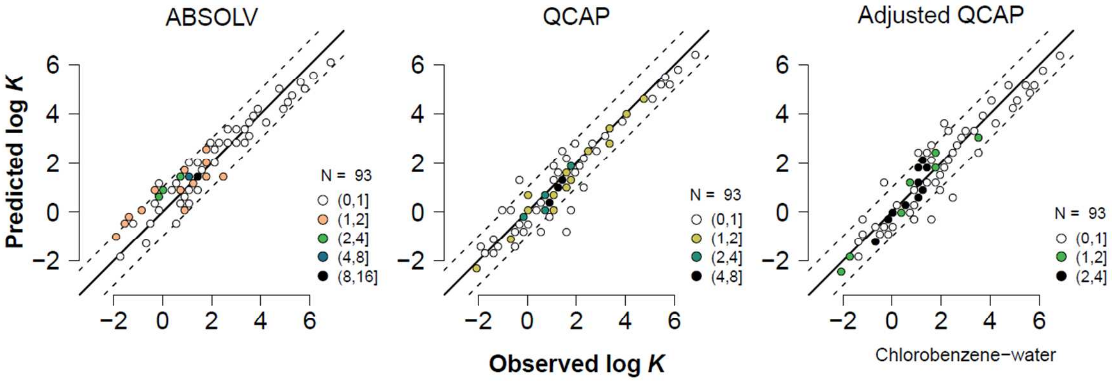

Figure S16 Hexbin two dimensional histogram plot[2] of predicted versus observed log chlorobenzene-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

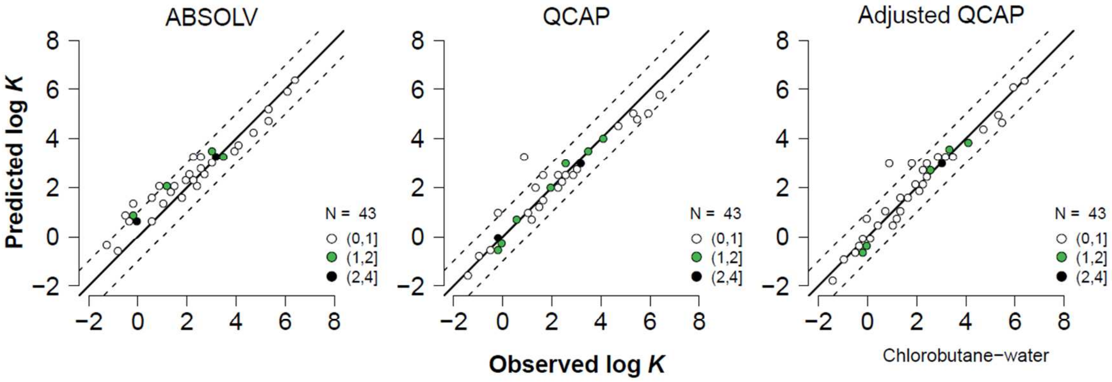

Figure S17 Hexbin two dimensional histogram plot[2] of predicted versus observed log chlorobutane-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

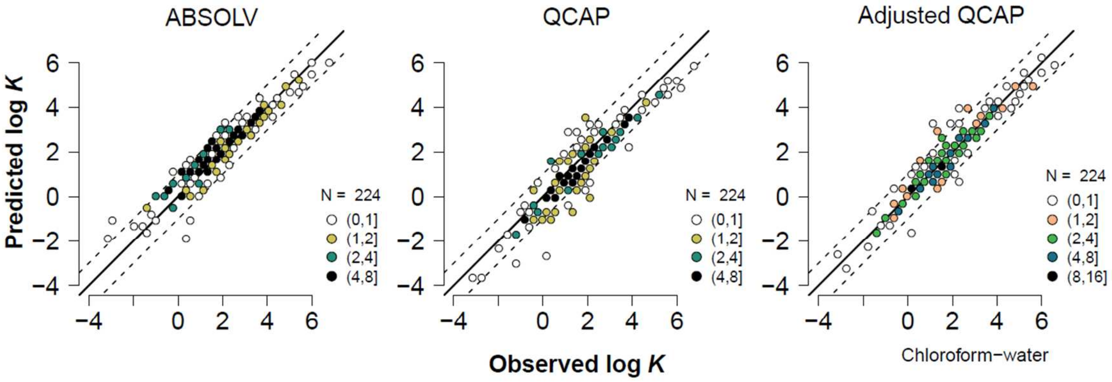

Figure S18 Hexbin two dimensional histogram plot[2] of predicted versus observed log chloroform-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

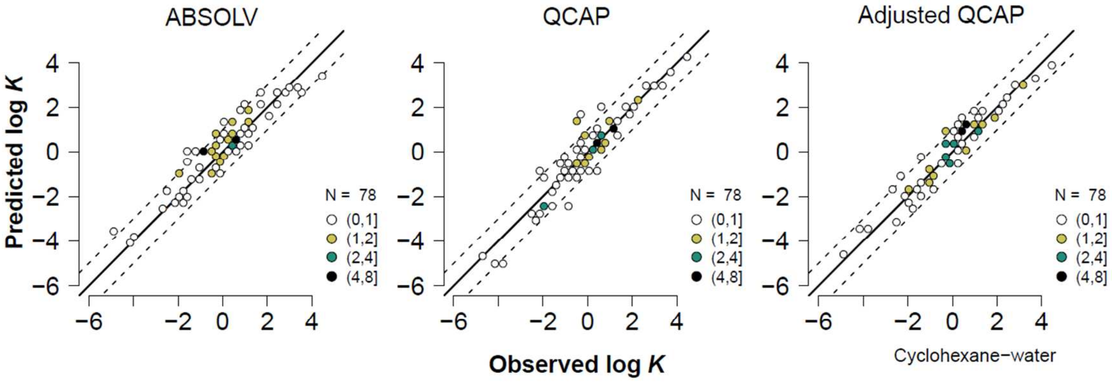

Figure S19 Hexbin two dimensional histogram plot[2] of predicted versus observed log cyclohexane-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

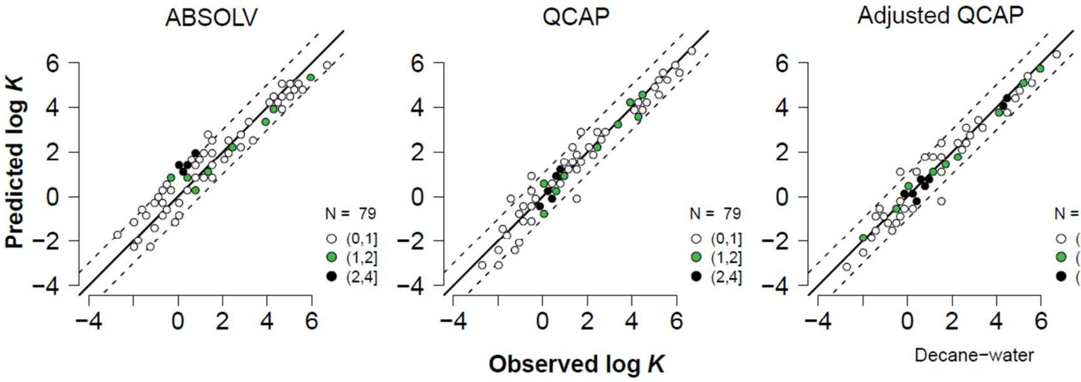

Figure S20 Hexbin two dimensional histogram plot[2] of predicted versus observed log decane-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

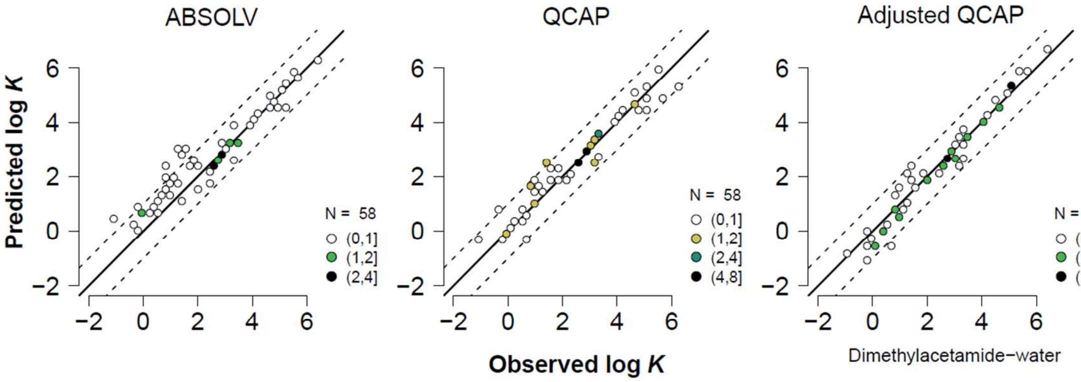

Figure S21 Hexbin two dimensional histogram plot[2] of predicted versus observed log dimethylacetamide-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

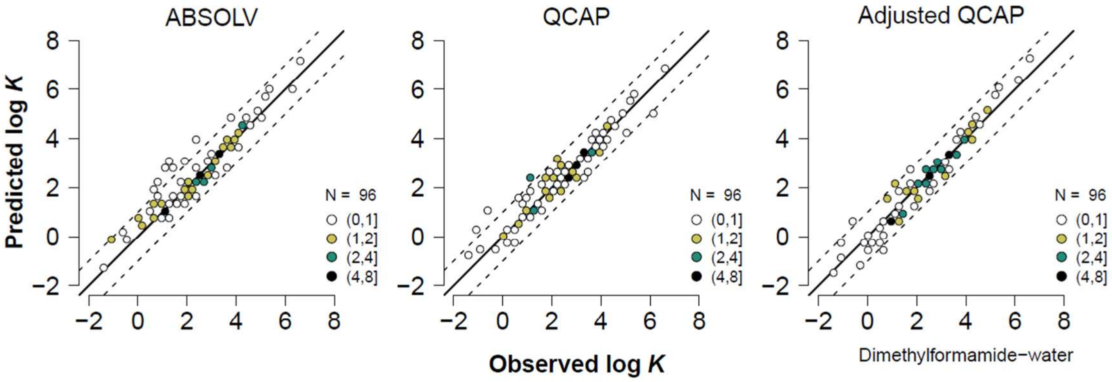

Figure S22 Hexbin two dimensional histogram plot[2] of predicted versus observed log dimethylformamide-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

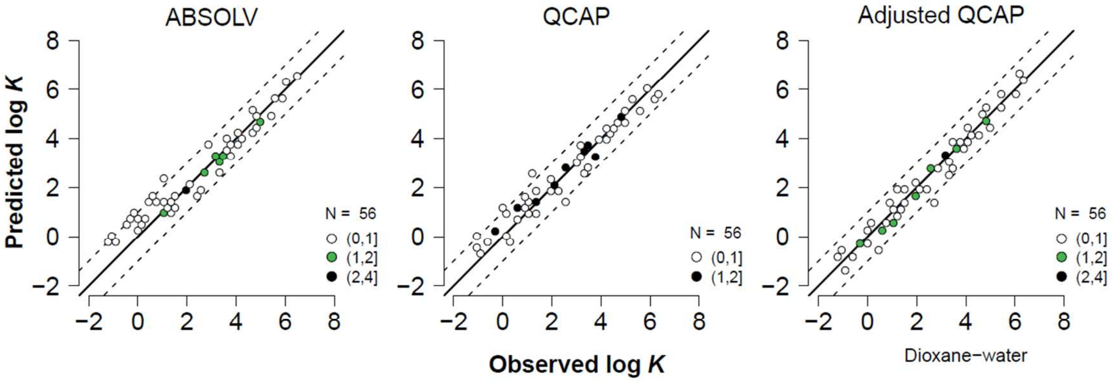

Figure S23 Hexbin two dimensional histogram plot[2] of predicted versus observed log dioxane-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

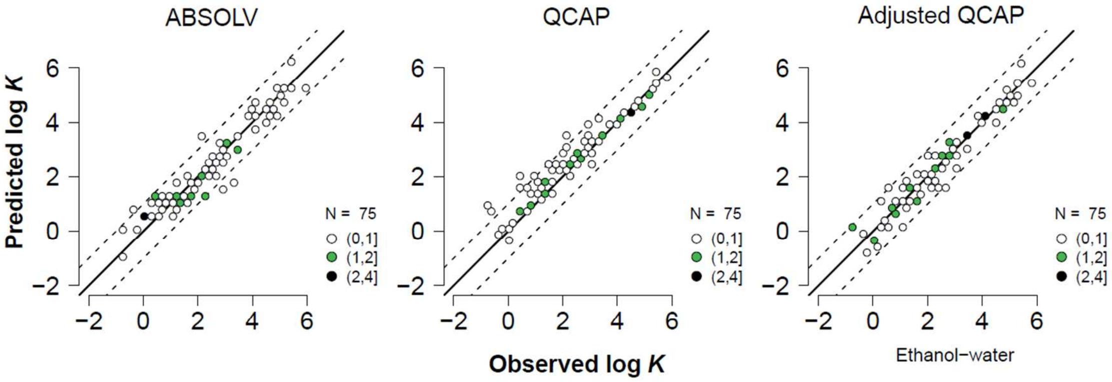

Figure S24 Hexbin two dimensional histogram plot[2] of predicted versus observed log ethanol-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

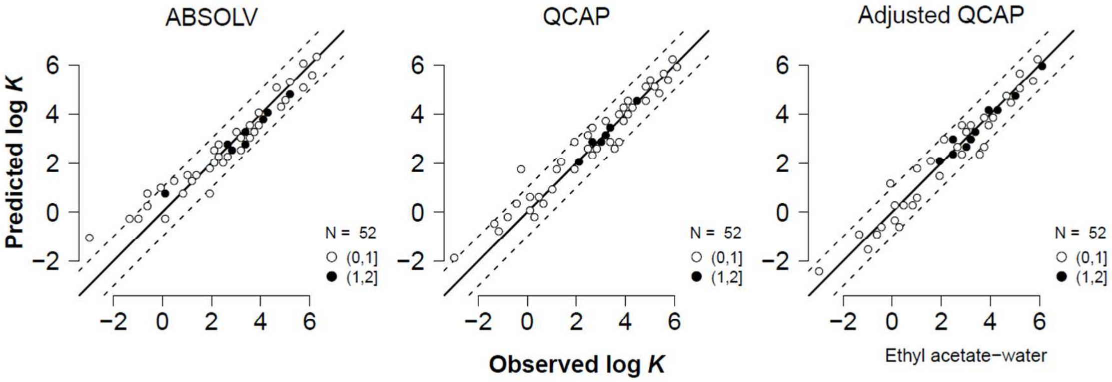

Figure S25 Hexbin two dimensional histogram plot[2] of predicted versus observed log ethyl acetate-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

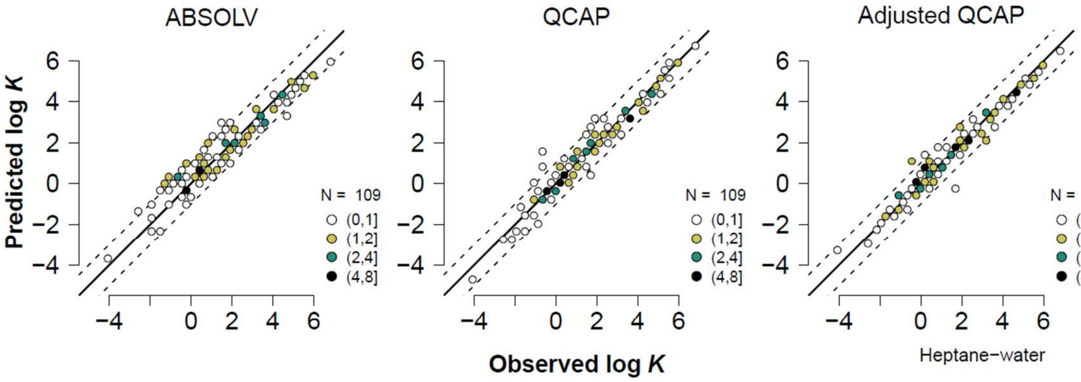

Figure S26 Hexbin two dimensional histogram plot[2] of predicted versus observed log heptane-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

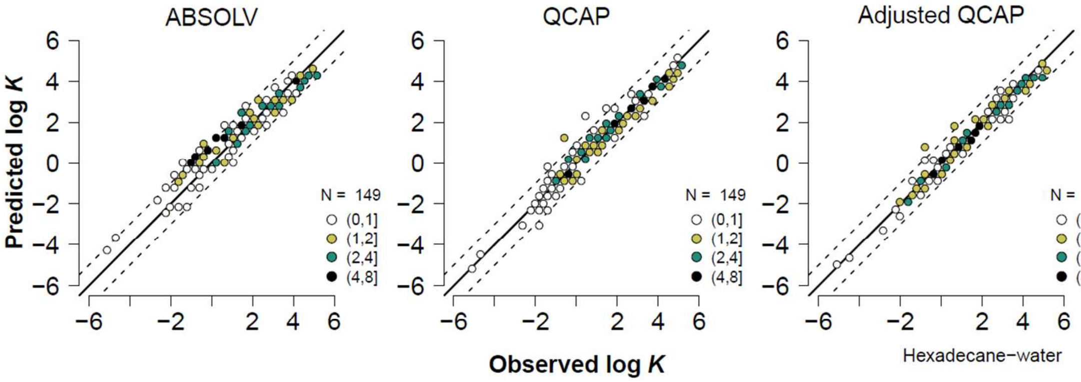

Figure S27 Hexbin two dimensional histogram plot[2] of predicted versus observed log hexadecane-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

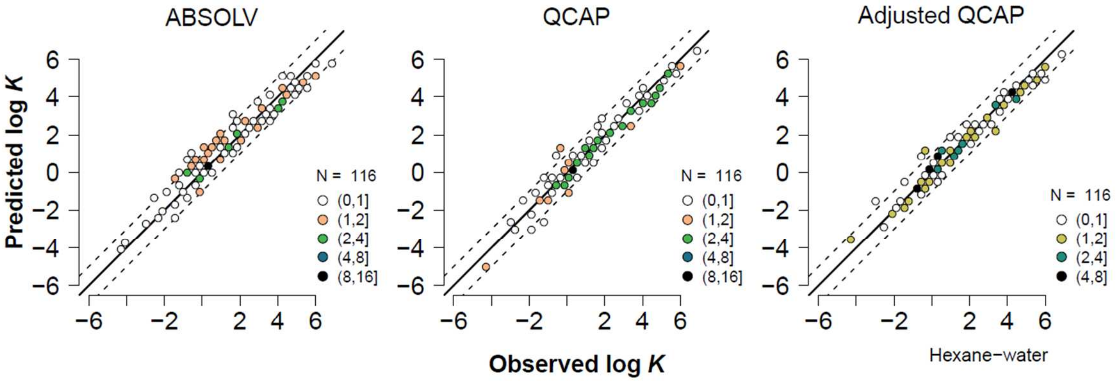

Figure S28 Hexbin two dimensional histogram plot[2] of predicted versus observed log hexane-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

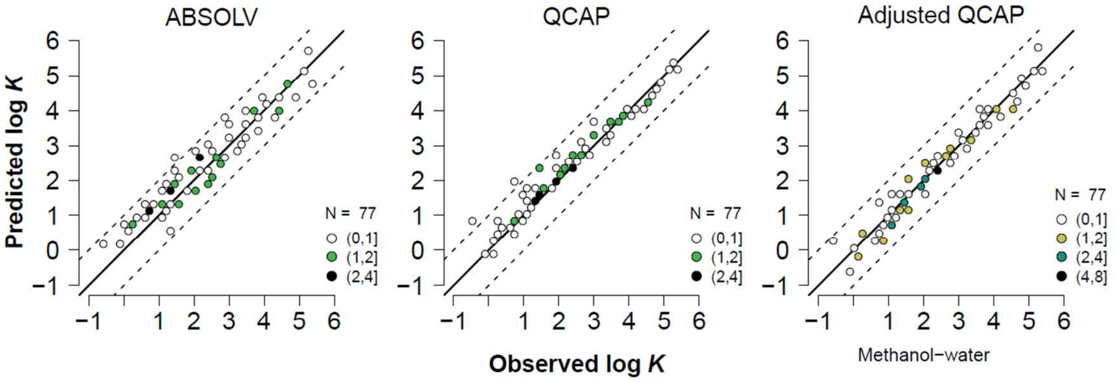

Figure S29 Hexbin two dimensional histogram plot[2] of predicted versus observed log methanol-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

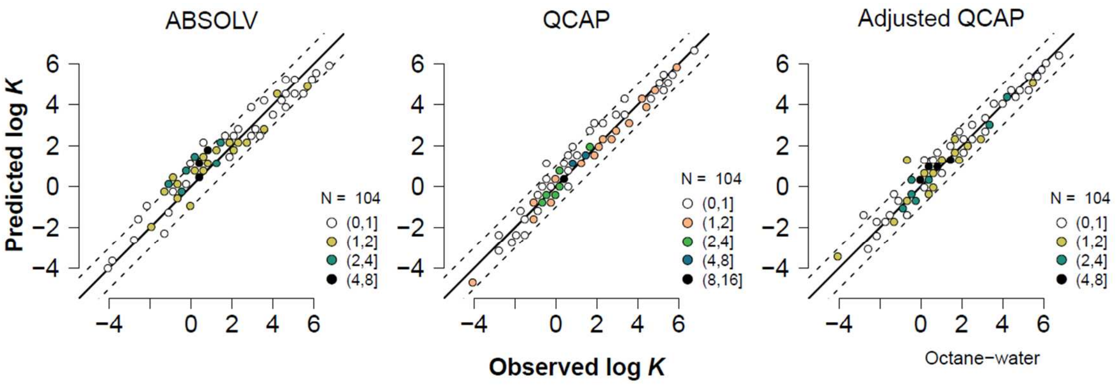

Figure S30 Hexbin two dimensional histogram plot[2] of predicted versus observed log octane-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

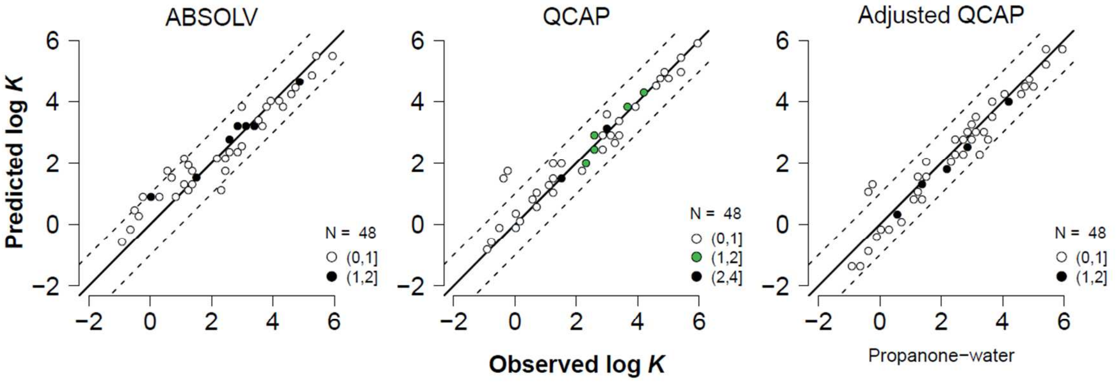

Figure S31 Hexbin two dimensional histogram plot[2] of predicted versus observed log propanone-water partition coefficients. Solid lines show perfect agreement. Dashed lines represent ±1 order of magnitude.

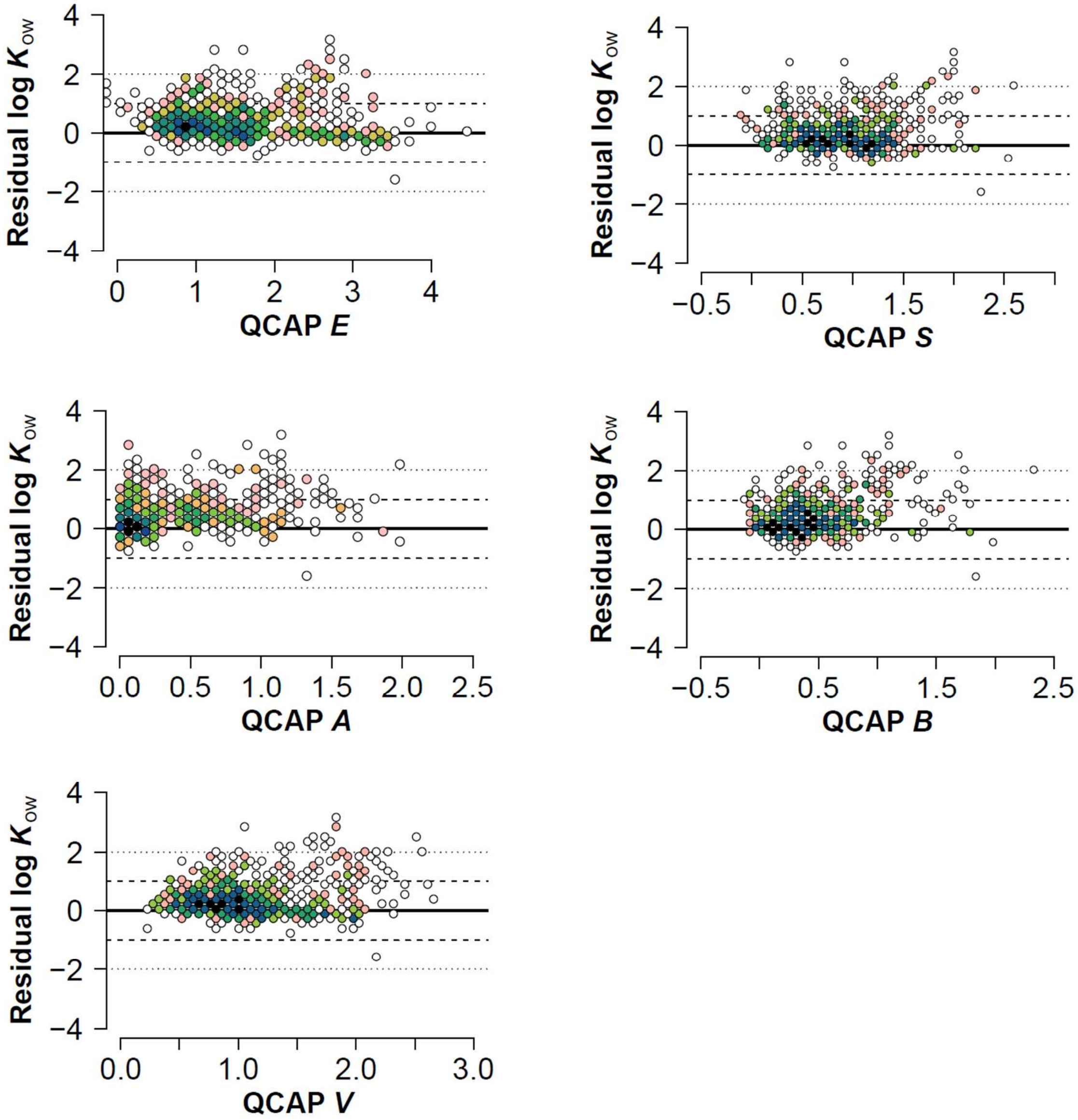

Figure S32 Hexbin two dimensional histogram plots[2] presenting residuals (= predicted log K_{ow} – observed log K_{ow}) of wet octanol-water partition coefficients versus *E*, *S*, *A*, *B*, *V* of QCAP parameters.

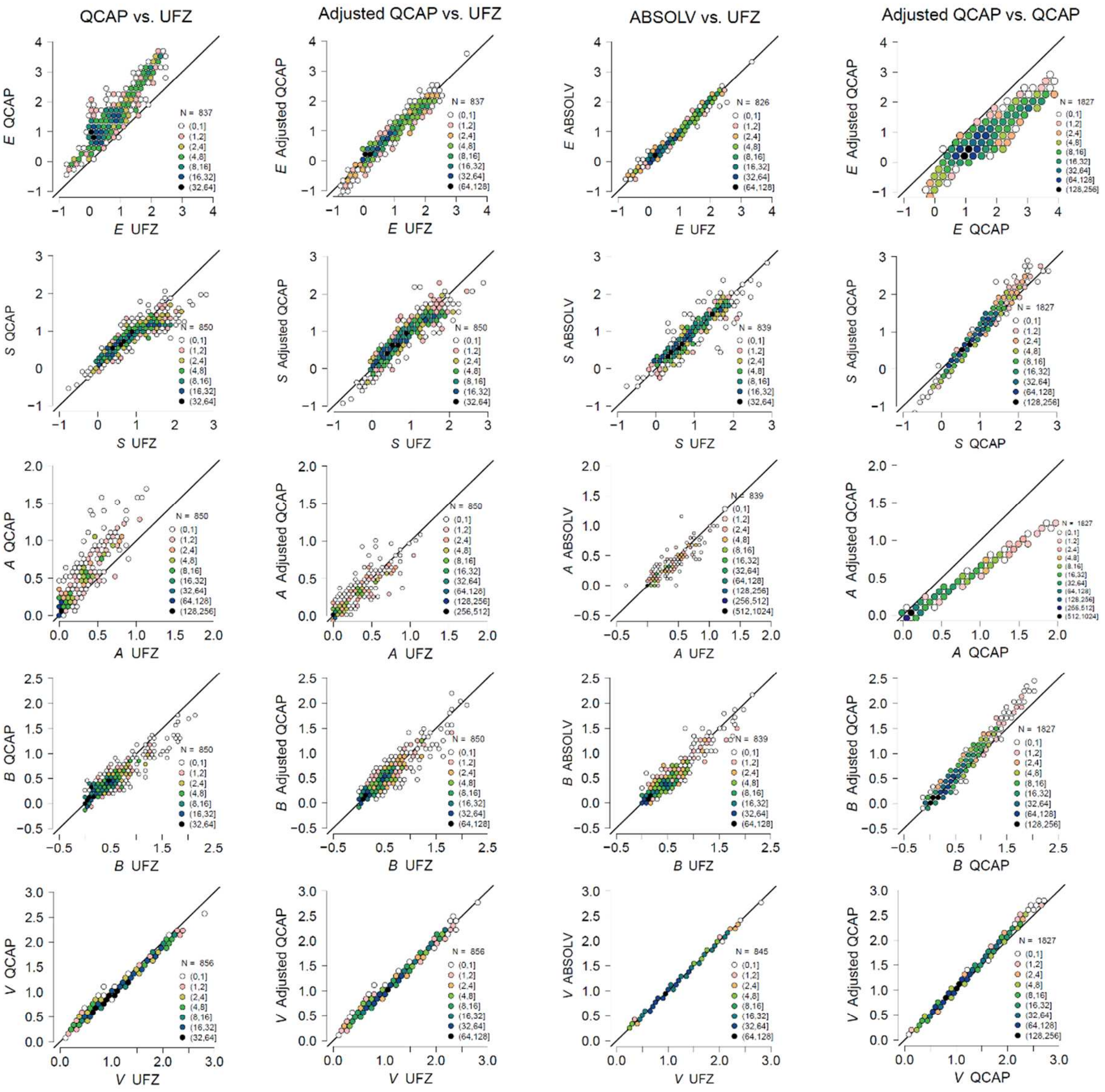

Figure S33 Hexbin two dimensional histogram plots[2] comparing the solute parameters (*E*, *S*, *A*, *B*, *V*) between QCAP and UFZ, between Adjusted QCAP and UFZ, between ABSOLV and UFZ, and between Adjusted QCAP and QCAP.

Contact information for requesting the COSMO-SAC model

The quantum chemical COSMO-SAC model is available on request. Interest groups can contact Dr. Sandler by email at sandler@udel.edu.

Discussion of the solute constant term in the QCAP model

It is possible that in addition to a constant, c_i, for each solvent, an additional constant, c_j, should be included for each solute. The following equation is used:

$$\log K_{i,j} = c_i + c_j + e_i E_j + s_i S_j + a_i A_j + b_i B_j + v_i V_j \qquad \text{(S1)}$$

where the subscript *i* denotes the solvent and *j* denotes the solute. When $c_j = 0$, Eq. (S1) is the general form of the Abraham pp-LFER model.

Based on the QCAP method, E_j is computed from the molecular polarizability and V_j is estimated from molecular volume. Since E_j and V_j are known parameters, the equations for the unknowns parameters (c_j, S_j, A_j, and B_j) are:

$$\log K_i - \ c_i - v_i V - e_i E = c_j + \ s_i S + a_i A + b_i B \qquad \text{(S2)}$$

The four unknowns (c_j, *S*, *A*, and *B*) are estimated using the MLR function (lm function) in the R programing language[3] applied to Eq. (S2).

Of the 1827 solute constants c_j, 1049 are relatively small and within ± 0.5 (Figure S34 (1)). More than half of the solute constants c_j (955 out of 1827) are statistically insignificant at the 0.05 level (Figure S34 (2)). Additionally, there was no obvious accuracy improvement for the prediction of solvent-water partition coefficients with one extra solute parameter c_j (Figure S35). Therefore, the original form of the Abraham pp-LFER without the solute constant term is preferred to be used for computing the QCAP solute parameters.

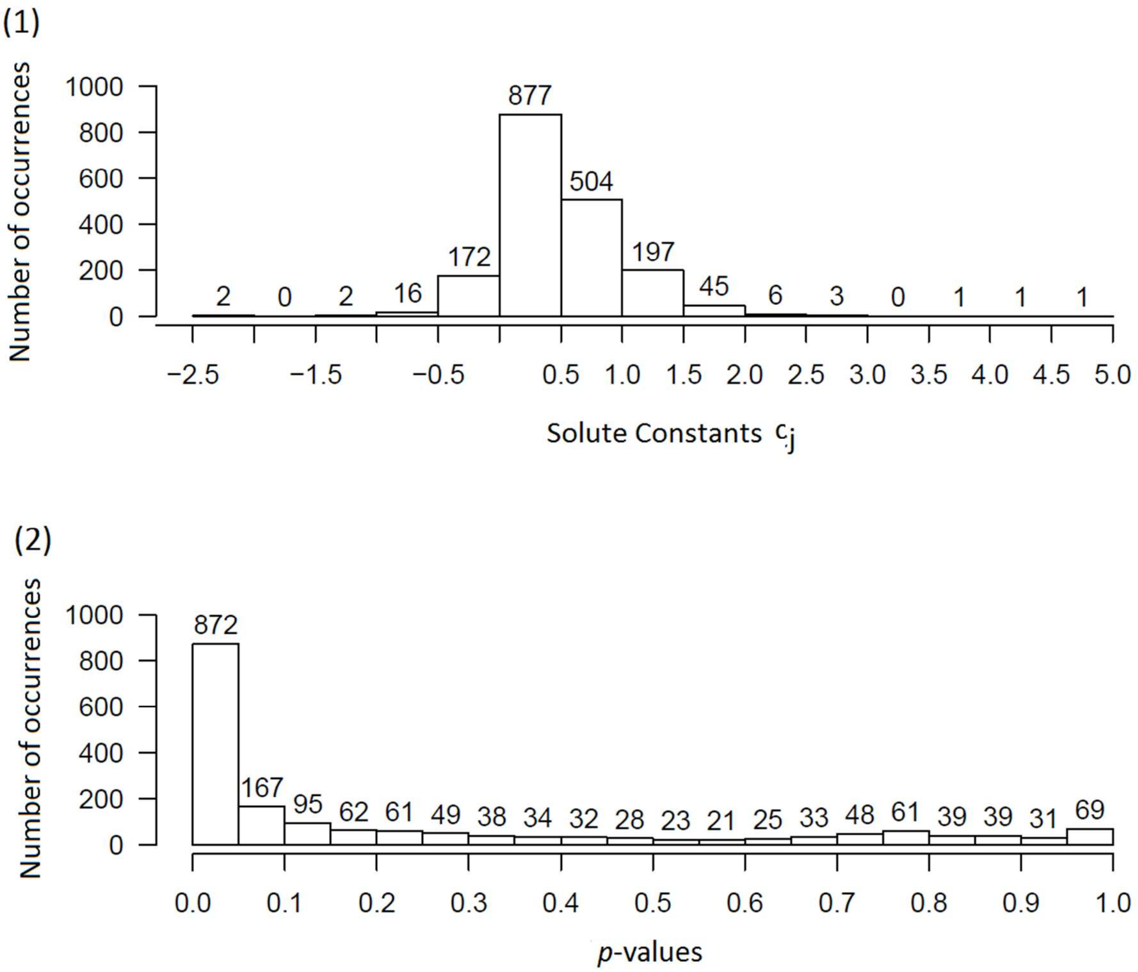

Figure S34. (1) Histogram of the solute constants; (2) Histogram of *p*-values of the solute constants.

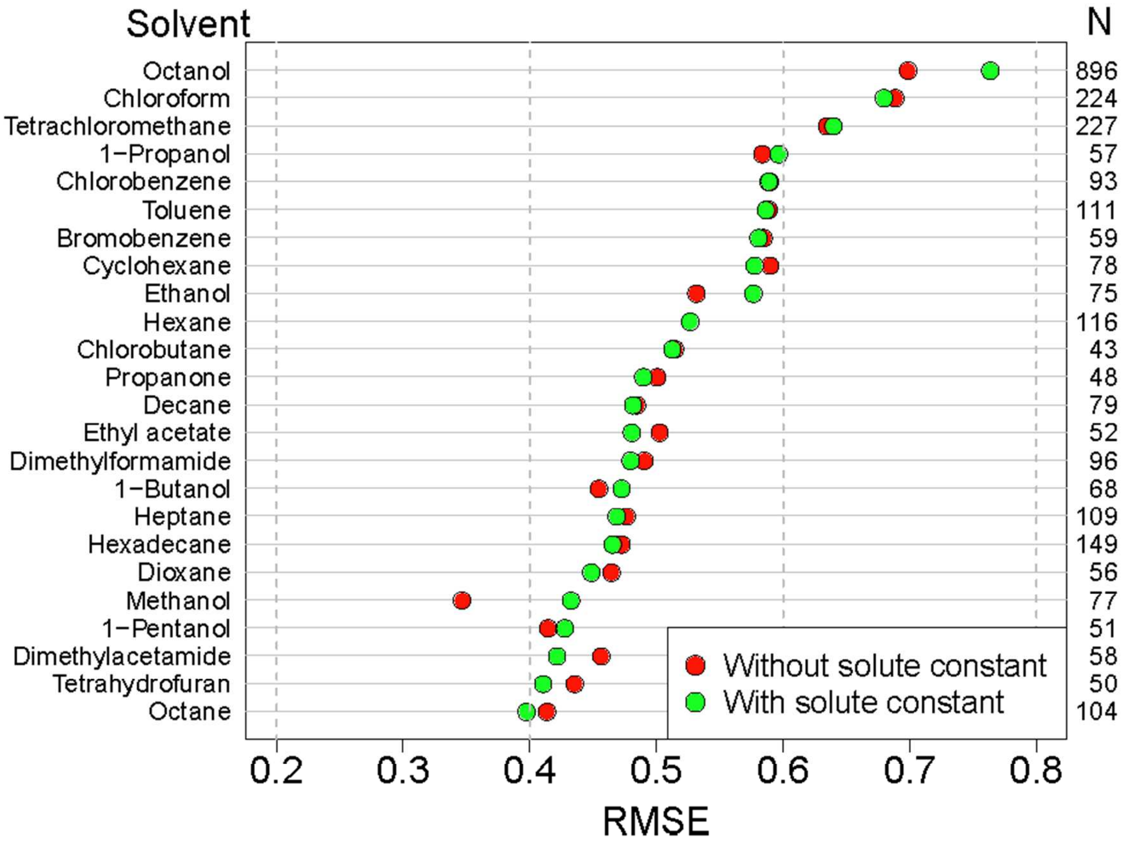

Figure S35. Dot plot presenting RMSEs of the residuals (log predicted - log observed partition coefficients) for the predictions of solvent-water partition coefficients. Predictions by two methods are compared: (1) Abraham pp-LFER model without solute constant, which is the QCAP model; and (2) Abraham pp-LFER model with a solute constant. The solvents listed on the left axis are ordered from the smallest to the largest QCAP RMSE. The right axis presents N, the number of solutes in each system.

Comparison of the measured and molecular polarizability computed refractive index

A comparison of the Abraham parameter *E* computed from the molecular polarizability is compared to those from the UFZ-LSER database in Figure 2. A possible cause for the discrepancy may be due to the estimate of the index of refraction computed using the Clausius–Mossotti of Eq. (6). It can be inverted so that the index of refraction η can be estimated using the molecular polarizability, α:

$$\eta = \sqrt{\frac{3\bar{V} + 8\pi N_{\mathrm{A}}\alpha}{3\bar{V} - 4\pi N_{\mathrm{A}}\alpha}} \tag{S3}$$

For Eq. (S3), the COSMO calculated molecular volume V (cm^{3}mol^{-1}/100) can be used to estimate the molar volume $\bar{V}$.

A comparison of the measured and molecular polarizability computed refractive index (Eq. S3) is shown in Figure S36. The measured refractive index (η_{exp}) was compiled from the ChemSpider database[4] and is listed in Table S4. An offset is observed between the predicted and measured refractive index. A regression analysis yielded a slope close to 1 (slope = 0.963) and an intercept = 0.206.

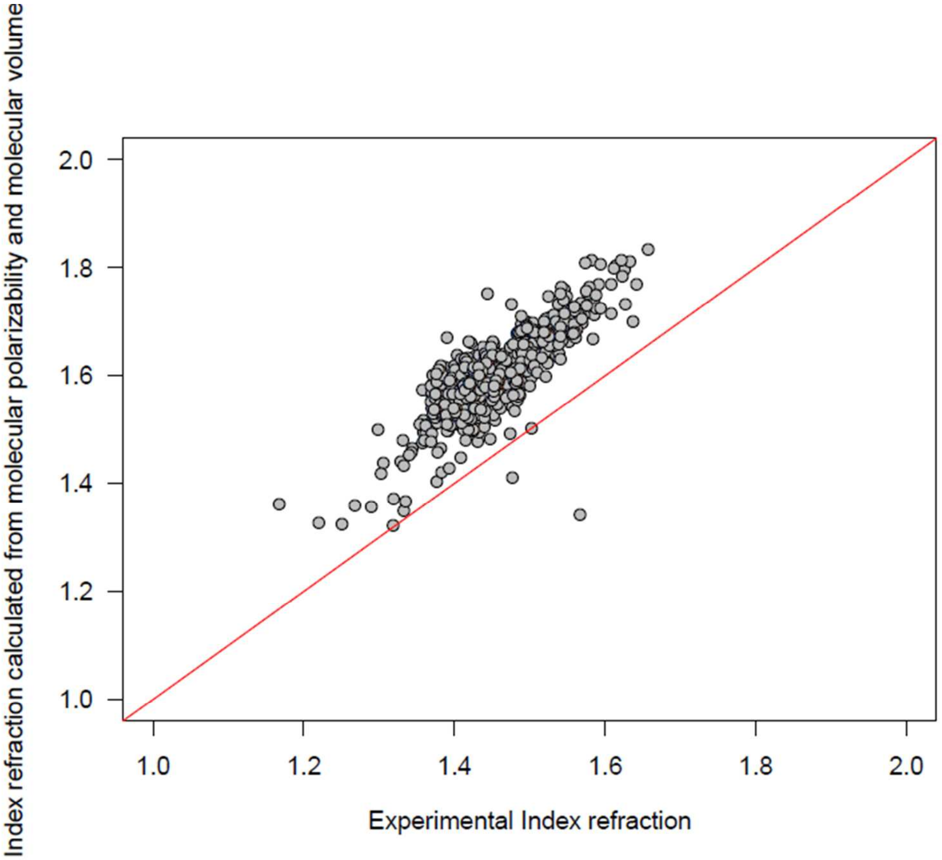

Figure S36. A comparison of the measured and molecular polarizability computed refractive index.

Discussion of the solute parameters estimated using the corrected refractive index in the QCAP model

We examined whether replacing the refractive index estimated from molecular polarizability with a corrected refractive index, using the regression described above, would affect the prediction capability of the QCAP solute parameters. A regression analysis of the dataset of Figure S36 yielded the following equation (N = 714, R^2 = 0.671):

$$\eta_{\text{exp}} = 0.698\eta_{\text{quantum}} + 0.331 \qquad \text{(S4)}$$

which can be used to relate the quantum chemically computed refractive index η_{quantum} to the experimental value. This equation can be used to adjust the index of refraction which is denoted by $\eta_{\text{corrected}}$. The resulting estimate of E, denoted by $E_{\text{corrected}}$, is computed using the corrected refractive index $\eta_{\text{corrected}}$ and molecular volume V:

$$E_{\text{corrected}} = 10\left[\frac{(\eta_{\text{corrected}}^2 - 1)}{(\eta_{\text{corrected}}^2 + 2)}\right]V - 2.832V + 0.526 \qquad \text{(S5)}$$

It is informative to compare the solute parameters estimated using the uncorrected refractive index (i.e., computed directly from molecular polarizability using Eq. (7)) and those estimated using the corrected refractive index based on the QCAP method to see if correcting the index of refraction using Eq. (S4) improves the performance of the resulting QCAP. The estimation procedure is the same. For solute parameters estimated using the corrected refractive index, $E_{\text{corrected}}$ and V are known parameters, and the unknown parameters ($S_{\text{corrected}}$, $A_{\text{corrected}}$, and $B_{\text{corrected}}$) are estimated using the MLR function (lm function) in the R programing language[3] applied to Eq. (S6):

$$\log K_i - c_i - v_i V - e_i E_{\text{corrected}} = s_i S_{\text{corrected}} + a_i A_{\text{corrected}} + b_i B_{\text{corrected}} \qquad \text{(S6)}$$

The results are compared in Figure S37 which presents the RMSEs of the predicted vs measured log solvent-water partition coefficients for the 24 solvents for which experimental data are available. This is the same comparison as made in Figure 3. There was no significant accuracy difference for the prediction of solvent-water partition coefficients between solute parameters estimated using the uncorrected refractive index and those estimated using the corrected refractive index. Therefore, solute parameters estimated based on the molecular polarizability without additional correction is preferred since the corrected index of refraction requires Eq. (S4) which may not apply to new compounds.

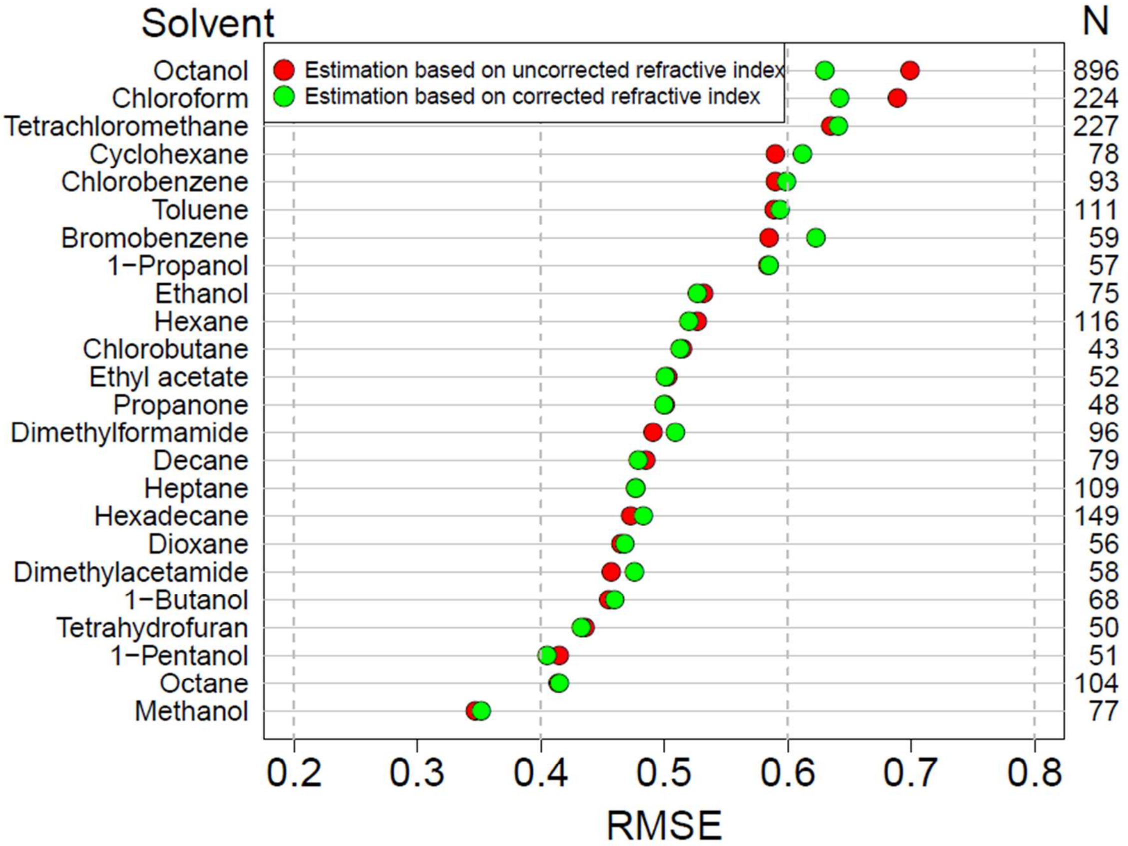

Figure S37. Dot plot presenting RMSEs of the residuals (log predicted - log observed partition coefficients) for the predictions of solvent-water partition coefficients. Predictions by two methods are compared: (1) Abraham pp-LFER model with solute parameters estimated using the uncorrected refractive index that is computed from the molecular polarizability in the QCAP model; and (2) Abraham pp-LFER model with solute parameters estimated using the corrected refractive index in the QCAP model. The solvents listed on the left axis are ordered from the smallest to the largest QCAP RMSE. The right axis presents N, the number of solutes in each system.

References

1. Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008,** *120* (1-3), 215-241.

2. Hexagonal Binning Routines, v 1.27.1. Carr, D.; Lewin-Koh, N.; Mächler, M. http://github.com/edzer/hexbin (accessed May 22, 2016).

3. R: A language and environment for statistical computing, v 3.0.2. R Core Team. R Foundation for Statistical Computing: Vienna, Austria, 2013. http://www.R-project.org/.

4. ChemSpider. Royal Society of Chemistry: 2015. http://www.chemspider.com/ (accessed June, 2017).