Supplemental Information for:
Improved prediction of PFAS partitioning with PPLFERs and QSPRs

Trevor N. Brown, James M. Armitage, Alessandro Sangion, and Jon A. Arnot

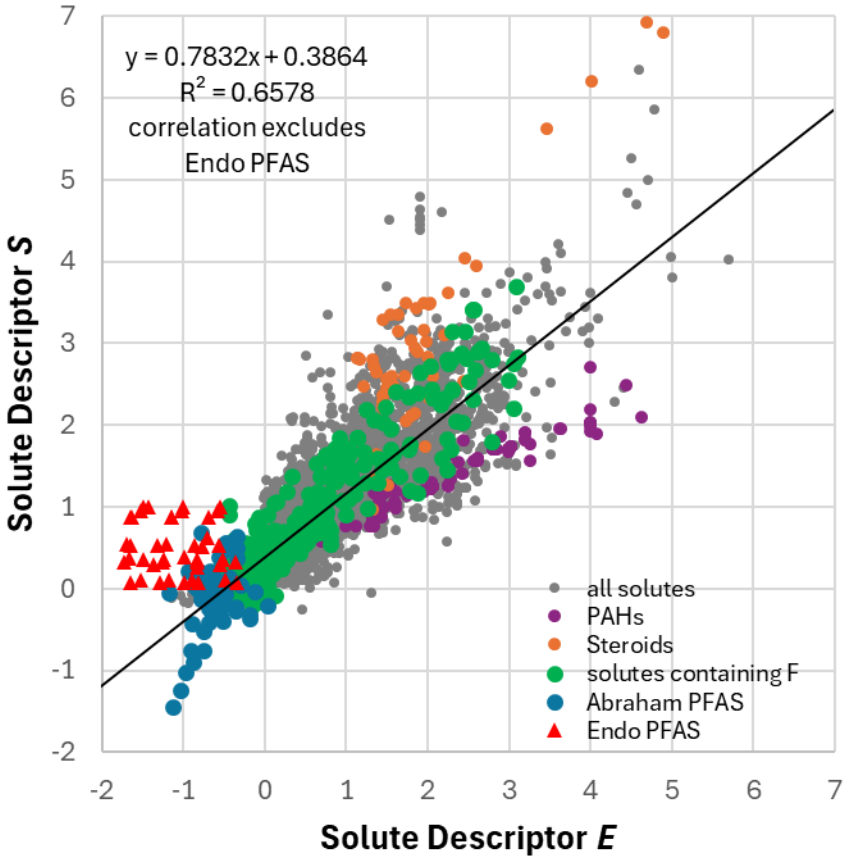
Contents

SI1. Differences in the two competing PFAS solute descriptor sets

Comparison of competing PFAS solute descriptors to other chemical classes

For this comparison we look at the dataset of reliable solute descriptors curated in previous work [1, 2]. Each set of competing descriptors has merits and limitations in its mechanistic interpretation. The descriptors from the Abraham group generally preserve the correlation between *E* and *S*, as shown in **Figure S1**, whereas the Endo data show no correlation, though the Endo *E* values have mostly been calculated from refractive index predicted by ACD Labs. As more CF_2 groups are added to the molecules the *S* value becomes more negative, which is consistent with the high electronegativity and low polarizability of fluorine atoms. The dipolarity of the OH functional group is apparently overwhelmed by the effects of the perfluorinated tail. In contrast, the Goss/Endo descriptors assume that the *S* value is constant within each chemical class of PFAS. This is true for chemical classes characterized by a polar functional group and n-alkyl tails of varying length [3], and assuming this behavior for perfluorinated tails is reasonable, and is supported by experimental data for FTOHs and FTOs [4]. In this interpretation the dipolarity of the OH functional group is dominant and adding more CF_2 groups has little further effect on the dipole of the distant OH group. **Table 1** in the main text shows the solute descriptors are similar to those of their alkyl analogues, but with lower polarizability/dipolarity, a higher hydrogen bonding donor property, and weaker van der Waals interactions. The large discrepancy between the *E* and *S* values in the Goss/Endo PFAS descriptors does not necessarily recommend the Abraham descriptors, there are other chemical classes with even larger discrepancies such as polycyclic aromatic hydrocarbons (PAHs) and steroids with descriptors calibrated by Abraham, see **Figure S1**. These chemical classes are at the opposite end of the *E*/*S* scale, but it could be that there were not enough data in the PFAS chemical space to detect the discrepancy before these new measurements.

Figure S1. Correlation of *S* vs. *E*



Abraham's use of *V* instead of $MV_{[I]}$

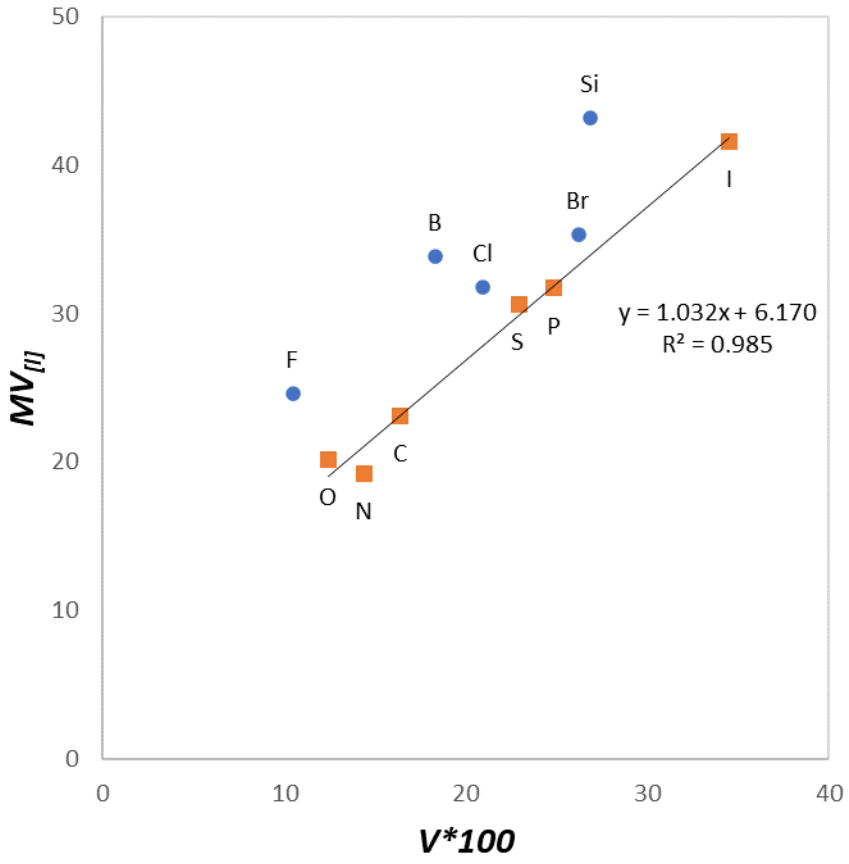
McGowan volume (*V,* named V_X in earlier work) was originally derived from parachors, which are a measure of how atomic and molecular volumes vary with temperature and surface tension of a liquid [5]. McGowan's goal was to make a model that was consistent with the parachor data, and unlike earlier work, also could be related to measures of van der Waals volume (intrinsic molecular volume) such as those obtained by X-ray crystallography. Intrinsic molar volume is independent of the molecular interactions a solute may experience in its pure phase or any other system. The atomic volumes of McGowan were not fitted individually, rather the differences between the atomic volumes of all elements in each row of the periodic table were assumed to differ by a fixed increment [5]. Therefore, changing the atomic volume of just one element within a row in the periodic table (as done specifically for fluorine atoms for PFAS PPLFER predictions by Goss and colleagues, i.e., V_F [4]) is a major departure from how *V* was originally derived. McGowan also converted the calculated atomic volumes to atomic radii and found a good correlation with previous measures of van der Waals atomic radii [5]. Decades later, research from the Abraham group found that *V* correlates strongly with other methods of calculating the van der Waals volume [6].

In early versions of the PPLFER equations liquid molar volume ($MV_{[I]}$) was used instead of *V*. Abraham decided to change from $MV_{[I]}$ to V_X for several reasons [7]; for example, getting $MV_{[I]}$ values for solids was a problem, the calculation of V_X was much simpler, and correction factors had to be applied to $MV_{[I]}$

anyway for some chemical classes resulting in an adjusted value MV_{adj}. The main difference between using *V* and $MV_{[I]}$ is that *V* is a measure of intrinsic molar volume, while $MV_{[I]}$ is a measure of bulk molar volume. Deviations between *V* and $MV_{[I]}$ should be expected because they measure different things. *V*, like other intrinsic molar volume metrics, measures only the volume occupied by the electron clouds of the molecules. $MV_{[I]}$ measures the electron clouds and the empty space around the molecules in the liquid phase. This empty space is due to the packing efficiency which depends on the shape of the molecules and on the strength of the attraction between molecules, such as hydrogen bonding and polar interactions. Because $MV_{[I]}$ varies with hydrogen bonding and polarity while *V* does not, using $MV_{[I]}$ instead of *V* produces different values of *S*, *A*, and *B* when calibrating solute descriptors.

In recent work we developed a method to predict $MV_{[I]}$ using an increment method loosely based on the calculation of *V [8]*. In this method the increment between atoms in a row of the periodic table was not fixed, but instead they were fitted by multiple linear regression against experimental $MV_{[I]}$ derived from liquid density and *MW*. This allowed the average effects of each atom on molecular packing in the bulk liquid phase to be captured. As shown in **Figure S2** the core organic subset (C, N, O, P, S) plus iodine have a strong correlation with *V* and the discrepancies amount to simply adding about 6 cm^3/mol to each increment. Note that Abraham divided McGowan volume by 100 in the original derivation of *V[7]*, so it must be multiplied by 100 to be on the same scale as $MV_{[I]}$. It is interesting to note that fluorine has only the third largest discrepancy between the $MV_{[I]}$ and *V* increments in the expanded set of organic atoms (B, C, N, O, F, Si, P, S, Cl, Br, I); boron has a larger discrepancy and silicon has the largest. The 6 cm^3/mol discrepancy must correspond to the average amount of empty space around each atom in the liquid phase. The larger deviations for halogens, boron, and silicon imply that these elements experience weak intermolecular interactions which increases the empty space around them, which is consistent with what we understand about PFAS. Organosilicon compounds have been noted to have PPLFER descriptor similar to PFAS, e.g., they have negative *S* values and their partitioning is better explained by **Equation 3** of the main text [9]. This may be because of the discrepancy between $MV_{[I]}$ and *V*, but the observed effect is not as large because many PFAS, i.e., perfluorinated, typically contain many more fluorine atoms than there are silicon atoms in most organosilicon compounds. When Goss et al. compared $MV_{[I]}$ and *V* their dataset appeared to contain no boron or silicon atoms, nor any solutes with a high degree of chlorination or bromination [4], so the discrepancies shown in **Figure S2** are not apparent.

Figure S2. Atom Increments for $MV_{[I]}$ vs. *V*100*



Recreating the PFAS solute descriptor calibrations

The Abraham set was fitted by using solver in MS Excel to minimize the errors between predictions and the experimental data for FTOHs summarized in Abraham et al [10]. In the current study, we successfully repeated the calibrations of the Abraham group, reproducing the solute descriptors with a discrepancy of less than 0.01, confirming our understanding of their data and methods. We then made modifications to explore why the Abraham solute descriptor set might be different from the Endo set. First, we repeated the Abraham group calibration using V_F instead of V, but this failed to produce agreement between the two sets of values. Second, we tested updating the $S{\cdot}S$ interaction term in the Abraham PPLFER equation for VP [11]. A term describing the interactions between solute molecules in their pure phase should be proportional to the strength of the interactions, but this is not the case for $S{\cdot}S$, because for negative values of S the term increases. To test this, we recalibrated the PPLFER for liquid *VP* using the original data but setting the term $S{\cdot}S$ to a value of 0 when S was negative. We then recalibrated the Abraham group solute descriptors using this updated equation, but again this failed to produce agreement between the two sets of descriptors. Finally, we tested the calculations further by using different forms of PPLFER equations in the calibration, i.e. **Equation 3** from the main text. To test this for the Abraham set we replaced the PPLFER equations with versions fitted in our previous work [1, 8]. This reduced the number of equations available for the fitting because there is only one version of the log

K_{AW} and VP equations. However, these equations are fitted using the original data from the Abraham group, so this is largely consistent with the Abraham group's approach, using the refitted equations from Endo would have been a major departure. We also constrained the three FTOHs to have the same *S*, *A*, and *B* values as done by Endo [12]. The solute descriptors for FTOHs recalibrated using this alteration are more consistent with the values of Endo, and when we also replaced *V* with V_F we then obtained values for *S*, *A*, and *B* of 0.20, 0.45, and 0.53. These calibrated values compare reasonably well with the values from Endo which are 0.35, 0.60, and 0.31, even considering the differences in the data used for calibration.

The Endo set calibration was reproduced using a python script and the numpy module to implement the method as described in Endo [12], with an initial guess of the descriptors made with IFSQSAR, and then iterative rounds of refitting the system parameters and the solute descriptors. We used our own databases of log K_{OW}, log K_{OA}, log K_{HxdW}, log K_{AW}, [1, 8] which have more solutes than the datasets used by Endo. The *L* solute descriptor was left fixed in the recalibration rather than allowed to "float", because the effect of this should be small and the differences between the original *L* values from [13] and the refitted values in Endo [12] are minor. The agreement between the reproduced values and those presented by Endo are good, with R^2 values of 0.98 or greater and the standard deviation between original and reproduced values was about 0.03 for *S* and *B*. However, there is a clear difference in our reproduced values for *A* which are less than those of Endo, with a slope of 0.88 and an intercept of about 0 and a standard deviation of 0.08. This difference may be due to the different log *K* datasets used for the calibration.

Despite these limitations we consider the results are in reasonable agreement and provide a sound basis from which to probe differences between the two competing solute descriptor sets. As was done with the Abraham set, we adjusted the calibration method attempting to recalibrate the Endo solute descriptor set into a form resembling the Abraham set. We changed V_F to *V* and changed the form of the PPLFER equations from **Equation 3** into the form of **Equations 1** and **2** depending on whether the system consisted of two condensed phases or a condensed phase and the gas phase consistent with the approach of Abraham. The solute descriptor *E* is missing from the data of Endo [12], so we used values from other work where available [4], or predicted the refractive index with ACD Labs (2023.1.0, build 3666) and calculated *E* with V_F or *V* as described Abraham et al. [14] but substituting *V* with V_F as was done by Goss et al. [4]. We tried excluding the solute descriptors for PFAS in the iterative PPLFER equation refitting, and we tried combinations of these three alterations. We found that the solute descriptors as calibrated by Endo are quite stable and could not produce values which resemble those of the Abraham set.

Comparing the accuracy of solute descriptor sets when calculating log *K*

We compare calculated values of the four partition ratios log K_{OW}, log K_{OA}, log K_{HxdW}, log K_{AW}, for three solutes 4:2 FTOH, 6:2 FTOH, and 8:2 FTOH and quantify the deviation between the calculated and experimental values as the root mean squared error (RMSE). These experimental data were used in the calibration of both competing sets of solute descriptors, but they are the most reliable and all three of the major partition ratios are available, so there are no other good options for comparison. We applied the PPLFER equations from our previous work [1, 8] (Brown system parameters), the recalibrated equations from Endo [12], and equations from Abraham [10]. The Abraham solute descriptors have lower RMSE than the Endo set when used with the Abraham system parameters, but using the Endo

solute descriptors with the Endo system parameters gives the lowest overall RMSE for all four partition ratios by a margin of 0.07 to 0.40. Calculations made using the Endo solute descriptors with the Abraham system parameters and vice versa are generally poor, with higher RMSE than using either of the consistent solute descriptor/system parameter sets. Calculations made with the Brown system parameters, which use **Equation 3** as the form of the PPLFER equation, but use V rather than V_F, have a slightly lower RMSE when applied with the Abraham solute descriptors than applying the Endo system parameters with the Abraham solute descriptors. Applying the Brown system parameters with the Endo solute descriptors also gives a lower RMSE than applying the Abraham system parameters with the Endo solute descriptors. This means that the Brown system parameters work better with the two competing solute descriptors sets than the sets work with their competitor's system parameters. However, the RMSEs using the Brown system parameters with the Brown solute descriptors (ranging from 0.21 to 0.49 for all four partition ratios) are higher than using the Endo solute descriptors with the Endo system parameters.

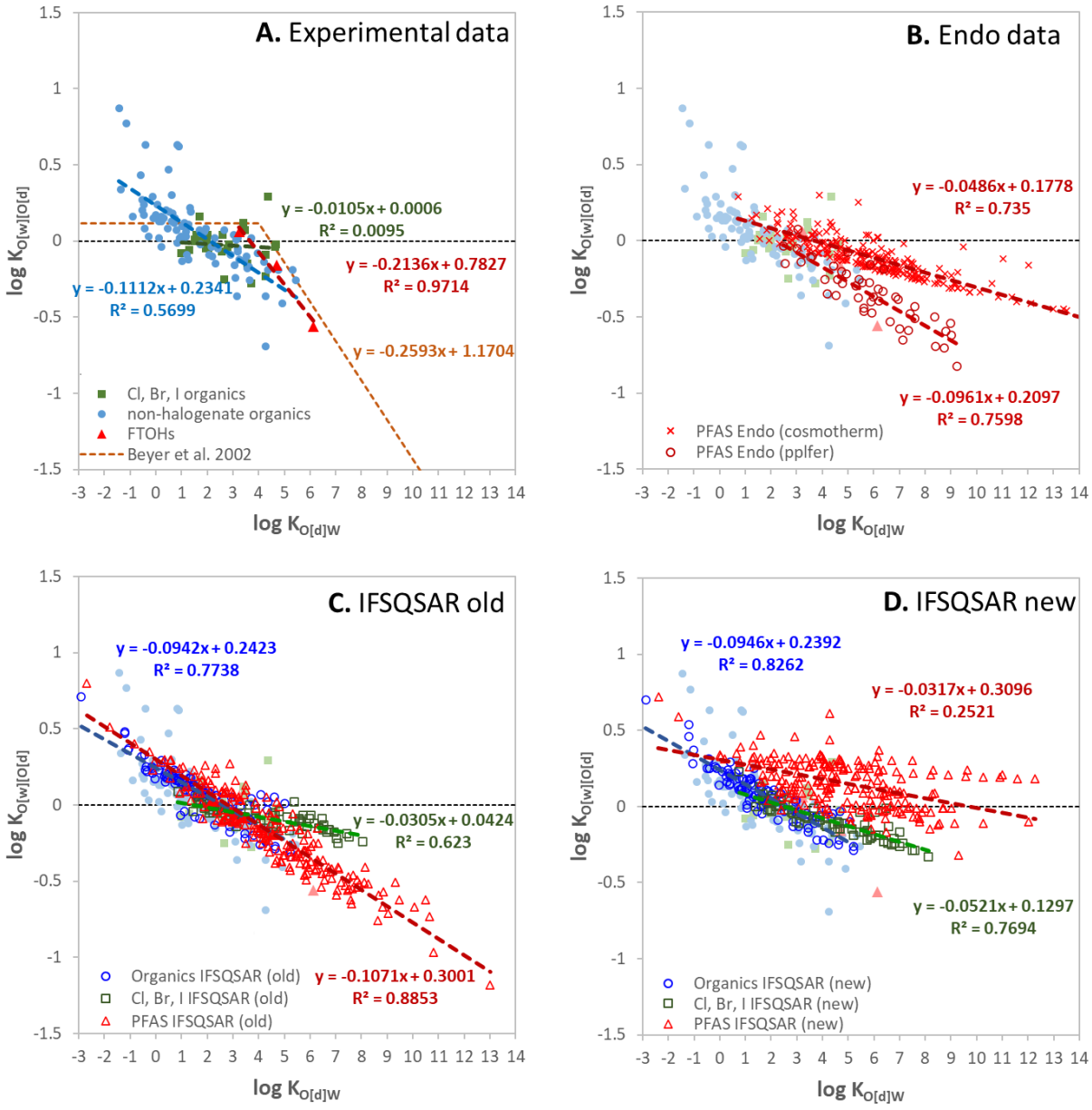
SI2. Further investigation of log K_{OA} prediction discrepancy.

An obvious question that arises from **Figures 1** and **2** in the main text and **S2** is how the predictions for log K_{OA} can have a slope so different from 1, but predictions for log K_{OW} and log K_{AW} do not, when the three properties are locked in a thermodynamic property cycle. This is possible because of the discrepancy in partitioning between "wet" octanol (octanol saturated with water) and dry octanol (neat octanol). This is a known effect, for example it is accounted for in deriving final adjusted values (FAV) [15, 16]. Measured K_{OW} considers "wet" octanol and measured K_{OA} considers "dry" octanol. Beyer et al. derived an equation relating the dry log K_{OW} (log $K_{O[d]W}$) to wet log K_{OW} (log $K_{O[w]W}$) that can be used to make the correction [15]. **Figure S3** shows plots for the hypothetical partition ratio between wet and dry octanol (log $K_{O[w]O[d]}$) vs. log $K_{O[d]W}$. When the y values in these plots are close to zero the discrepancy between wet and dry octanol measurements is negligible. In **Figure S3A** the orange line shows the equation derived by Beyer et al. based on data mostly for various chlorinated aromatics, plus some pesticides and organic acids. They proposed a two-mode model with an inflection point at log K_{OW} = 4, though inspecting the plot in the original paper the regression equation for the portion greater than 4 would have a good fit for the entire dataset [15], and Schenker et al. only use the equation for the log K_{OW} > 4 portion [16]. For comparison we show 112 solutes from the Abraham training datasets where values for all three partition ratios were available [17-19]. These are small organic solutes, mostly alkanes, with some alkenes and aromatics, and containing zero or one heteroatom functional groups. We have plotted the solutes containing halogen atoms separately. None of the chemicals in the Beyer dataset or the Abraham solutes contain any fluorine atoms. The only PFAS solutes from Endo with measured values for all three partition ratios are 4:2, 6:2 and 8:2 FTOHs [12]. It can be seen in **Figure S3A** that the regression equations for organics (y = -0.111 + 0.234) and halogenated organics (y = -0.011 + 0.001) are quite different from the Beyer et al. equation for log K_{OW} > 4 (y = -0.259x + 1.17). The regression equation for FTOHs (y = -2.14 + 0.783) is more comparable to the Beyer et al. equation; however, with only three data points from one chemical series there is not enough data to draw any conclusions with confidence. The overall trend is the same for all equations, hydrophilic chemicals favor wet octanol (log $K_{O[w]O[d]}$ values greater than 0) and hydrophobic chemicals favor dry octanol (log $K_{O[w]O[d]}$ values less than 0).

The experimental Abraham solutes are included (but faded) for comparison in the other panels of **Figure S3**. **Figure S3B** shows the data for PFAS measured and calculated by Endo [12]. Red circles are values calculated using PPLFER equations, with the solute descriptors and system parameters calibrated by Endo from experimental data. COSMO*therm* predicted values for 221 PFAS are shown as red "x" markers for additional points of comparison. The regression equation for the Endo PPLFER derived values closely matches the equation for the Abraham organics dataset, but the regression equation for COSMO*therm* data has a smaller slope indicating that the discrepancy between wet and dry octanol may be underestimated relative to the experimental data. **Figures S2C** and **S2D** show the results calculated using the old IFSQSAR (version 1.1.0) predictions and IFSQSAR updated in this work (version 1.1.1). The results for **Figure S3D** are as predicted by the final model (v.1.1.1) presented in the main text. It can be seen in **Figure S3C** that organic and halogenated organic solutes are predicted by IFSQSAR v1.1.0 to closely follow the same trends of the Abraham solute dataset, which is not surprising because these are included in the training data of IFSQSAR. IFSQSAR v.1.1.0 predicts that PFAS solutes closely match the trend of the Abraham organic solutes and the Endo PPLFER dataset. As shown in **Figure S3D** the IFSQSAR v.1.1.1 predictions for organic and halogenated organic solutes also closely follow the same trend as the

Abraham solute dataset. However, the IFSQSAR v.1.1.1 predictions for PFAS are different from all the other datasets shown in **Figure S3**, with much more scatter, a smaller slope, and a higher intercept.

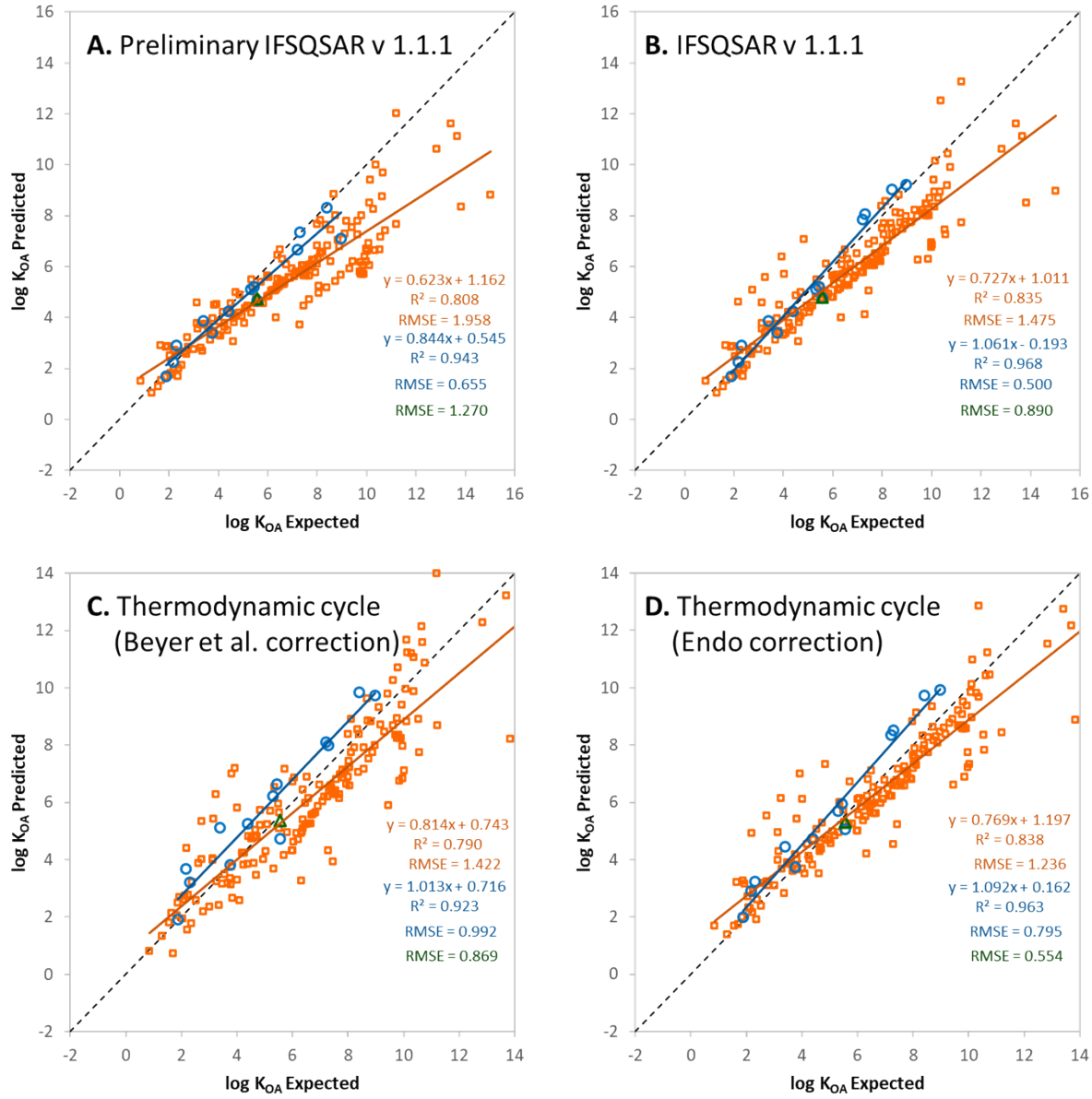
Figure S3. log $K_{O[w]O[d]}$ vs. log log $K_{O[d]W}$. Data from A are faded in B, C and D.



We undertook various analyses to better understand the discrepancy of the IFSQSAR v.1.1.1 model predictions for wet and dry octanol compared to the other predictions shown in **Figure S4**. In a preliminary version of IFSQSAR v.1.1.1 all the solute descriptors *E*, *S*, *A*, *B*, and *L* only had their regression coefficients updated to include the influence of new measured PFAS data from Endo, and to exclude suspect data points reported by Abraham. **Figure S4A** shows predictions for log K_{OA} from the preliminary model and **Figure S4B** (the same as **Figure S3B**) shows the results of the final IFSQSAR v.1.1.1. The preliminary model predictions for log $K_{O[w]O[d]}$ vs. log $K_{O[d]W}$ were very similar to the results shown in **Figure S3D** for the final model predictions, with comparable R^2, slope, and intercept. A large group of outliers is obvious in **Figure S4A**, these are a series of 28 sulfonic acids. Note that the COSMO*therm* calculations were done for the neutral forms of the PFAS, and the models in IFSQSAR likewise make predictions for the neutral forms. However, the database of reliable solute descriptors contains no sulfonic acids, because making the required measurements for the neutral form is likely impossible, so no fragments covering this functional group have been calibrated. There are sulfonamides present in the IFSQSAR training data and PFAS with this functional group are well predicted. Inspecting the predicted solute descriptors for sulfonic acids and comparing them to sulfonamides which are less acidic it was obvious that the *A* solute descriptor prediction was much too low, about 0.35 vs. 0.55 for perfluorinated sulfonamides. Increasing the *A* value for sulfonic acids by 1 or more yielded predictions in line with COSMO*therm*. When the sulfonic acids were excluded the RMSE for predictions of the preliminary model for log K_{OA} was 1.5 and the slope increased to 0.70. In the final IFSQSAR v.1.1.1 model the QSPR for *A* has been completely recalibrated to select a new set of fragments which explain the effects of including the new measured PFAS data from Endo and excluding suspect data points reported by Abraham. The completely recalibrated QSPR predicts *A* values of 0.74-1.15 for sulfonamides and 0.37-1.0 for sulfonic acids. Various minor modifications to the IFSQSAR development algorithm were tested, for example restricting or expanding the types of recursive fragments included in the fragment pool, but none of these yielded better results than the published method used to calibrate the QSPRs [2]. **Figure S4B** shows the sulfonic acids predicted by IFSQSAR v.1.1.1 are no longer outliers, and the overall slope has increased to 0.73. This updated model gives the best fit for the log K_{OA} values calculated with the Endo PPLFER calculated values, with an RMSE of 0.50. Some new outliers above the 1:1 line are sulfonyl fluorides, which are an unusual structure also not included in the Abraham training data.

Another method tested to improve the prediction of log K_{OA} was to apply a thermodynamic cycle, using the IFSQSAR v.1.1.1 predictions for log K_{OW} and log K_{AW} to estimate log K_{OA} as: log K_{OA} = log K_{OW} - log K_{AW}. Then log $K_{O[w]O[d]}$ was subtracted to correct for dry vs. wet octanol. Two methods of calculating log $K_{O[w]O[d]}$ were tested, first the regression equation from Beyer et al. for log K_{OW} > 4 (y = -0.259x + 1.17). shown in **Figure S3A**, and the regression equation based on the new Endo PPLFER data shown in **Figure S3B**; the results are shown in **Figures S3C** and **S3D**, respectively. Both methods give predictions that have a lower RMSE relative to the COSMO*therm* calculated values and appear to correct the underprediction bias; however, both thermodynamic property cycling methods yield worse predictions for the PPLFER-based log K_{OA} values, which are used to validate the QSPRs. We cannot recommend applying a thermodynamic property cycle based only on better agreement with other modelling results.

Figure S4. log K_{OA} predicted by various methods.

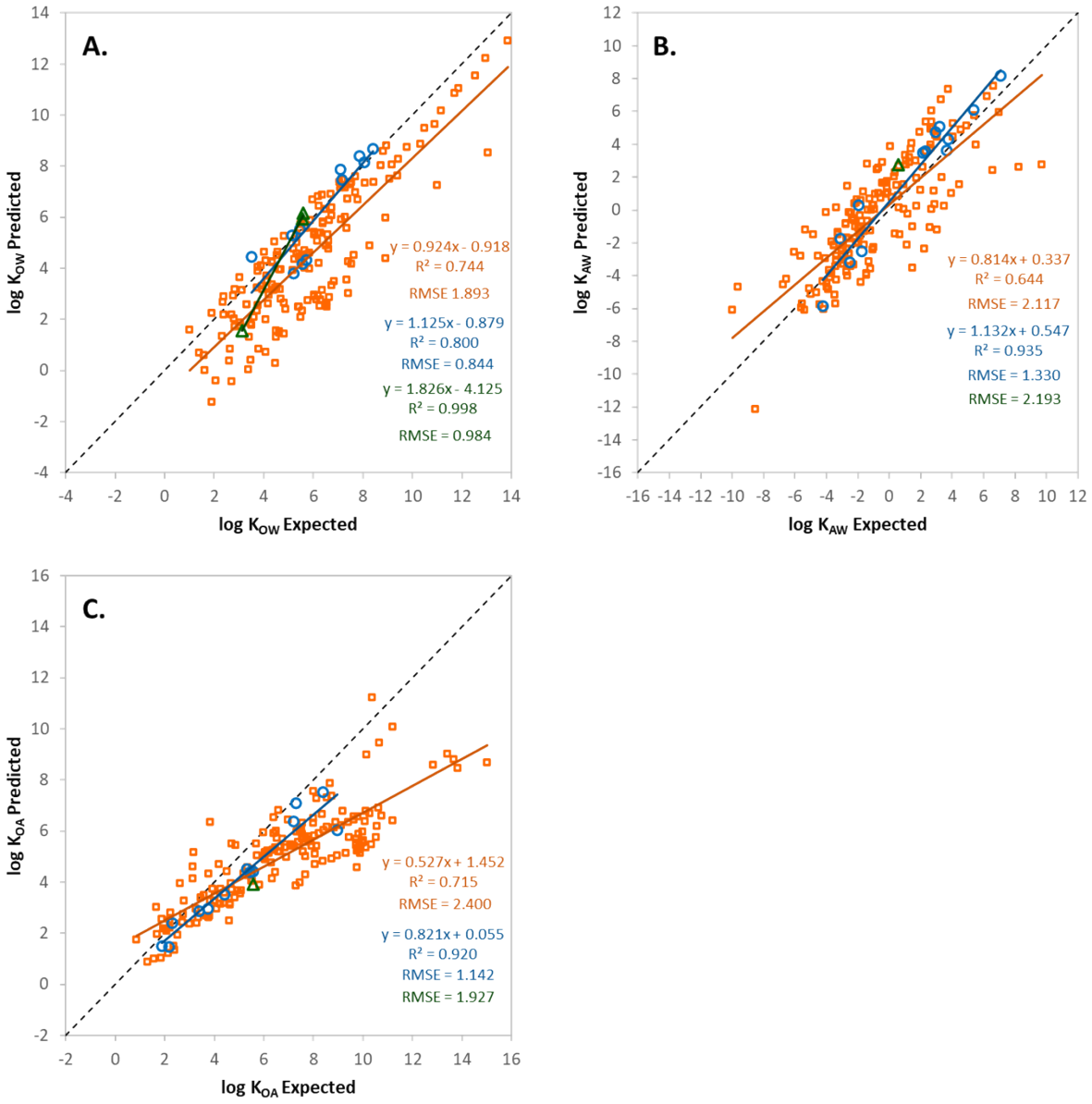


We have not assigned much weight to model agreement with the experimental log K_{OA} values for FTOHs in any of these tests, because the data are too few and not representative of the diversity of PFAS structures. In fact, the remaining underprediction of log K_{OA} may be because there were so little data available to recalibrate the system parameters. However, we tested adding more log K_{OA} values for PFAS to the system parameter recalibration from other sources [20, 21] where there was overlap with the new Endo solute descriptors, but this did not improve the results vs. the COSMO*therm* calculated values and in some cases the results were worse. It is possible that the data we added were of insufficient measurement quality, and therefore did not help to improve the results. We also tested recalibrating the system parameters for log K_{OA} based on the exact dataset used and provided by Endo, and various combinations of datasets. The recalibrated system parameters were comparable to those of Endo regardless of which PFAS and other solutes are included in the calibration dataset; therefore, we used our own dataset supplemented by the data for FTOHs used by Endo.

We have considered other possible reasons for the under-prediction of log K_{OA} at higher values and the discrepancy in the wet vs. dry octanol partitioning. We considered that the discrepancy may also arise because the recalibration of the other solute descriptor QSPRs was insufficient; however, the residuals between the IFSQSAR v1.1.1 predictions and the COSMO*therm* predictions do not show any correlation with the predicted solute descriptors, $R^2<0.06$ in all cases. A weak correlation ($R^2\approx0.35$) was apparent between the log K_{OA} residuals and the *A* solute descriptor in the preliminary IFSQSAR v 1.1.1 model but recalibrating the *A* solute descriptor QSPR eliminated this correlation without resolving the overall discrepancy. The *S* solute descriptor was subject to the largest changes in the training data, some data going from quite negative values to slightly positive values. Recalibrating the *S* solute descriptor is beyond the scope of this incremental model update, because it represents a much larger time and computing power investment. We have set a new lower boundary on the *S* solute descriptor predictions of -0.2, and this made a difference for some individual chemicals but did not affect the overall underprediction of log K_{OA}. The underprediction might be some artifact related to replacing *V* with V_F, but again the log K_{OA} residuals show no correlation with *V*, V_F, or the difference between them.

One final possibility is that the underprediction for log K_{OA} at higher values is due to some unique PFAS partitioning behavior with regards to the discrepancy in partitioning between “wet” and “dry” octanol. The fact that the Endo PPLFER results match the COSMO*therm* predictions quite well seems to exclude this as a possibility, but the underprediction is mostly observed for PFAS for which Endo did not make measurements. Of course, it is possible the COSMO*therm* predictions for PFAS are also inaccurate [13, 22] as the regression between log $K_{O[w]O[d]}$ and dry log K_{OW} shown in **Figure S3B** does not match any of the experimental data. We cannot conclude that this means COSMO*therm* is wrong however, because overall **Figure S3** shows that different chemical classes have different trends with regards to wet vs. dry octanol partitioning. As stated in the main text, Hammer and Endo [13] found that COSMO*therm* tends to over-predict solvent-air partitioning which would explain some of the observed under-prediction bias, but not all of it. The cause of the observed discrepancies remains unknown. Further modelling work might resolve this, but it is likely that further high-quality measurements of all three major partition ratios for a diverse set of PFAS will be required to determine the cause.

Figure S5. A) log K_{OW} , B) log K_{AW} , and C) log K_{OA} predicted with solute descriptors from ACD Labs, combined with PPLFER equations from [23].



SI3. Physicochemical property data (experimental, predicted) available through the US EPA CompTox dashboard for selected PFAAs

Example 1: PFOA, CAS 335-67-1
https://comptox.epa.gov/dashboard/chemical/properties/DTXSID8031865

Table S1. Water solubility data reported as "experimental" for PFOA.

Source/Reference cited	**S_W (mol/L)**	**S_W (mg/L)**	**Comments**
Ding et al. Crit. Reviews in Environ. Sci. and Tech., Volume 43, 2013 - Issue 6	8.21E-03	3400	Original citation : 3M Environmental Laboratory. (2001). Characterization study of PFOA (lot #332), primary standard: Test control reference #TCR-99030–030. Phase: Solubility Determination. St. Paul, MN: 3M Company Likely to be the water solubility of a salt (i.e., reflects water solubility of neutral and charged form)
NCCT_Physchem	8.21E-03	3400	Pseudo-replication
Danish_EPA_SCPFAS_Report_2015	8.21E-03	3400	Pseudo-replication
NCCT_Physchem	8.21E-03	3400	Pseudo-replication
3M_PFOA_Sheet	8.21E-03	3400	Pseudo-replication
Tetko et al, J Comput Aided Mol Des. 2011; 25(6):533-54	8.21E-03	3400	Pseudo-replication
NCCT_Physchem	9.00E-03	3726	
NCCT_Physchem	1.00E-02	4140	Pseudo-replication
NCCT_Physchem	1.00E-02	4140	Pseudo-replication
Tetko et al, J Comput Aided Mol Des. 2011; 25(6):533-54	1.04E-02	4300	
ATSDR_Perfluoroalkyl_Cheminfo	2.29E-02	9500	Pseudo-replication
Danish_EPA_PFOA_Report_2005	2.29E-02	9500	Pseudo-replication
Tetko et al, J Comput Aided Mol Des. 2011; 25(6):533-54	2.29E-02	9500	Pseudo-replication
NCCT_Physchem	2.30E-02	9520	Pseudo-replication
NCCT_Physchem	2.50E-02	10350	
Reported Average	**1.37E-02**	**5670**	**Average of the above 15 values, which includes multiple instances of the same value (pseudo-replication)** **Reflects water solubility of the neutral *and charged form***

NCCT_Physchem = NCCT Collected PhysChem Data manually extracted from literature and/or databases

Table S2. Water solubility data reported as "predicted" for PFOA.

Source/QSAR cited	S_W (mol/L)	S_W (mg/L)	Comments
EPISUITE	6.27e-8	0.026	Predicted values using EPISUITE WSKOW v1.42 = 0.4813 mg/L (based on log $K_{OW,N}$ = 4.81)
TEST	4.73e-6	1.96	
ACD/Labs	1.10e-3	455	
OPERA 2.6	3.31e-2	13710	Not representative of the water solubility for the neutral form alone
Calculated Average	**8.55e-3**	**3540**	Not representative of the water solubility for the neutral form alone

Table S3. Octanol-water partitioning data reported as "experimental" for PFOA.

Source/Reference cited	log K_{OW}	Comments
NCCT_Physchem	1.92	Consistent with experimental partitioning data for perfluorooctanoate (***i.e., charged form, not neutral form***) reported by [24]
NCCT_Physchem	2.80	
Danish_EPA_SCPFAS_Report 2015	3.60	
NCCT_Physchem	3.60	
Rayne et al. J. Env. Sci. And Health Part A (2009) 44(12):1145-1199	3.60	Obtained from the SI of [25] Kelly et al. report the predicted value at 25 ^{o}C (SPARC) taken from [26] minus 1 log unit
Reported Average	**3.10**	**Average of the above five values, three of which are identical (and a predicted value)**

NCCT_Physchem = NCCT Collected PhysChem Data manually extracted from literature and/or databases

Table S4. Octanol-water partitioning data reported as "predicted" for PFOA.

Source/QSAR cited	log K_{OW}	Comments
OPERA 2.6	3.11	PFOA is included in the "nearest neighbour" section of the OPERA Calculation Report with a predicted value of 4.13
ACD/Labs Consensus	5.58	Calculation details not available
EPISUITE	6.30	Calculation details not available Consistent with value generated by KOWWIN v1.67 Predicted value = 4.81 using more recent versions of KOWWIN (v1.68 and v1.69)
ACD/Labs	7.75	Calculation details not available
Calculated Average	**5.69**	

Example 2: PFOS, CAS 1763-23-1
https://comptox.epa.gov/dashboard/chemical/properties/DTXSID3031864

Table S5. Octanol-water partitioning data reported as "experimental" for PFOS.

Source/Reference cited	log K_{OW}	Comments
Rayne et al. J. Env. Sci. And Health Part A (2009) 44(12):1145-1199	4.30	Obtained from the SI of [25] Kelly et al. report the predicted value at 25 ^{o}C (SPARC) taken from [26] minus 1 log unit
NCCT_Physchem	5.50	
NCCT_Physchem	7.03	Identical to predicted value reported for ACD/Labs (see below)
Reported Average	**5.61**	**Average of the above three values, at least two of which are predicted values**

NCCT_Physchem = NCCT Collected PhysChem Data manually extracted from literature and/or databases

Table S6. Octanol-water partitioning data reported as "predicted" for PFOS.

Source/QSAR cited	log K_{OW}	Comments
ACD/Labs Consensus	4.17	Calculation details not available
OPERA 2.6	5.61	PFOS is included in the "nearest neighbour" section of the OPERA Calculation Report with a predicted value of 5.01
EPISUITE	6.28	Calculation details not available Consistent with value generated by KOWWIN v1.67 Predicted value = 4.49 using more recent versions of KOWWIN (v1.68 and v1.69)
ACD/Labs	7.03	Calculation details not available
Calculated Average	**5.77**	

References

1. Brown, T.N., *Empirical regressions between system parameters and solute descriptors of polyparameter linear free energy relationships (PPLFERs) for predicting solvent-air partitioning.* Fluid Phase Equilibria, 2021. **540**: p. 113035.
2. Brown, T.N., *QSPRs for Predicting Equilibrium Partitioning in Solvent–Air Systems from the Chemical Structures of Solutes and Solvents.* Journal of Solution Chemistry, 2022. **51**(9): p. 1101-1132.
3. Abraham, M.H., et al., *Hydrogen bonding XVI. A new solute solvation parameter, S, from gas chromatographic data.* Journal of Chromatography A, 1991. **587**(2): p. 213-228.
4. Goss, K.-U., et al., *The Partition Behavior of Fluorotelomer Alcohols and Olefins.* Environmental Science & Technology, 2006. **40**(11): p. 3572-3577.
5. McGowan, J.C., *Molecular volumes and structural chemistry.* Recueil des Travaux Chimiques des Pays-Bas, 1956. **75**(2): p. 193-208.
6. Zhao, Y.H., M.H. Abraham, and A.M. Zissimos, *Fast calculation of van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds.* Journal of Organic Chemistry, 2003. **68**(19): p. 7368-7373.
7. Abraham, M.H. and J. McGowan, *The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography.* Chromatographia, 1987. **23**(4): p. 243-246.
8. Brown, T.N., A. Sangion, and J.A. Arnot, *Identifying Uncertainty in Physical-Chemical Property Estimation with IFSQSAR.* 2024.
9. Endo, S. and K.U. Goss, *Predicting partition coefficients of Polyfluorinated and organosilicon compounds using polyparameter linear free energy relationships (PP-LFERs).* Environ Sci Technol, 2014. **48**(5): p. 2776-84.
10. Abraham, M.H. and W.E. Acree, *Descriptors for fluorotelomere alcohols. Calculation of physicochemical properties.* Physics and Chemistry of Liquids, 2021. **59**(6): p. 932-937.
11. Abraham, M.H. and W.E. Acree, *Estimation of vapor pressures of liquid and solid organic and organometallic compounds at 298.15 K.* Fluid Phase Equilibria, 2020. **519**: p. 112595.
12. Endo, S., *Intermolecular Interactions, Solute Descriptors, and Partition Properties of Neutral Per- and Polyfluoroalkyl Substances (PFAS).* Environmental Science & Technology, 2023. **57**(45): p. 17534-17541.
13. Hammer, J. and S. Endo, *Volatility and Nonspecific van der Waals Interaction Properties of Per- and Polyfluoroalkyl Substances (PFAS): Evaluation Using Hexadecane/Air Partition Coefficients.* Environmental Science & Technology, 2022. **56**(22): p. 15737-15745.
14. Abraham, M.H., A. Ibrahim, and A.M. Zissimos, *Determination of sets of solute descriptors from chromatographic measurements.* Journal of Chromatography A, 2004. **1037**(1-2): p. 29-47.
15. Beyer, A., et al., *Selecting internally consistent physicochemical properties of organic compounds.* Environmental Toxicology and Chemistry, 2002. **21**(5): p. 941-953.
16. Schenker, U., et al., *Improving data quality for environmental fate models: a least-squares adjustment procedure for harmonizing physicochemical properties of organic compounds.* Environ Sci Technol, 2005. **39**(21): p. 8434-41.
17. Abraham, M.H., et al., *Hydrogen bonding. 32. An analysis of water-octanol and water-alkane partitioning and the Δlog p parameter of seiler.* Journal of Pharmaceutical Sciences, 1994. **83**(8): p. 1085-1100.
18. Grubbs, L.M., et al., *Mathematical correlations for describing solute transfer into functionalized alkane solvents containing hydroxyl, ether, ester or ketone solvents.* Fluid Phase Equilibria, 2010. **298**(1): p. 48-53.

19. Abraham, M.H., et al., *Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapours in water at 298 K, and a new method for its determination.* Journal of the Chemical Society-Perkin Transactions 2, 1994(8): p. 1777-1791.
20. Li, L., et al., *Retrieval, Selection, and Evaluation of Chemical Property Data for Assessments of Chemical Emissions, Fate, Hazard, Exposure, and Risks.* ACS Environmental Au, 2022. **2**(5): p. 376-395.
21. Baskaran, S., Y.D. Lei, and F. Wania, *A Database of Experimentally Derived and Estimated Octanol–Air Partition Ratios (KOA).* Journal of Physical and Chemical Reference Data, 2021. **50**(4).
22. Endo, S., J. Hammer, and S. Matsuzawa, *Experimental Determination of Air/Water Partition Coefficients for 21 Per-and Polyfluoroalkyl Substances Reveals Variable Performance of Property Prediction Models.* Environmental Science & Technology, 2023.
23. Brown, T.N., A. Sangion, and J.A. Arnot, *Identifying uncertainty in physical-chemical property estimation with IFSQSAR.* J Cheminform, 2024. **16**(1): p. 65.
24. Jing, P., P.J. Rodgers, and S. Amemiya, *High Lipophilicity of Perfluoroalkyl Carboxylate and Sulfonate: Implications for Their Membrane Permeability.* Journal of the American Chemical Society, 2009. **131**(6): p. 2290-2296.
25. Kelly, B.C., et al., *Perfluoroalkyl Contaminants in an Arctic Marine Food Web: Trophic Magnification and Wildlife Exposure.* Environmental Science & Technology, 2009. **43**(11): p. 4037-4043.
26. Arp, H.P.H., C. Niederer, and K.-U. Goss, *Predicting the partitioning behavior of various highly fluorinated compounds.* Environmental science & technology, 2006. **40**(23): p. 7298-7304.