Journal of Solution Chemistry (2023) 52:51–69
https://doi.org/10.1007/s10953-022-01214-7

Predicting the Temperature Dependence of the Octanol–Air Partition Ratio: A New Model for Estimating ΔU°_{OA}

Sivani Baskaran[1,2] · Akshay Podagatlapalli[1] · Alessandro Sangion[1,3] · Frank Wania[1,2]

Received: 11 May 2022 / Accepted: 11 September 2022 / Published online: 8 November 2022
© The Author(s) 2022

Abstract

The octanol–air partition ratio (K_{OA}) describes the partitioning of a chemical between air and octanol and is often used to approximate other partitioning phenomena in environmental chemistry (e.g., blood–air, atmospheric particulate matter–air, polyurethane foam-air). Such partitioning processes often occur at environmental temperatures other than 25 °C. Enthalpies ΔH°_{OA} or internal energies ΔU°_{OA} of phase transfer are used to express the temperature dependence of the K_{OA}. Existing poly-parameter linear free energy relationships (ppLFERs) for predicting ΔH°_{OA} were developed using a relatively small dataset. In this work we utilize a recently developed comprehensive K_{OA} database to create and curate a ΔU°_{OA} dataset containing 195 chemicals and use this dataset in the development of new predictive equations. Using the QSAR development platform QSARINS we evaluate the use of Abraham descriptors, other molecular descriptors, and the $\log_{10} K_{OA}$ at 25 °C as variables in different multilinear regression equations for ΔU°_{OA}. The ΔU°_{OA} of neutral organic chemicals can be reliably predicted using only the $\log_{10} K_{OA}$ ($\mathrm{RMSE_{EXT}} = 6.86\ \mathrm{kJ{\cdot}mol^{-1}}$, $\mathrm{R^2_{adj}} = 0.94$), only the solute's hydrogen acidity A and the logarithm of the hexadecane–air partition ratio L ($\mathrm{RMSE_{EXT}} = 7.23\ \mathrm{kJ{\cdot}mol^{-1}}$, $\mathrm{R^2_{adj}} = 0.93$), or A and $\log_{10} K_{OA}$ ($\mathrm{RMSE_{EXT}} = 6.76\ \mathrm{kJ{\cdot}mol^{-1}}$, $\mathrm{R^2_{adj}} = 0.95$).

Keywords Phys–chem properties · Property estimation · Octanol–air partitioning · Partition ratios · Temperature dependence · ppLFERs

Sivani Baskaran and Akshay Podagatlapalli are co-first authors.

Frank Wania
frank.wania@utoronto.ca

[1] Department of Physical and Environmental Sciences, University of Toronto Scarborough, Toronto, Canada

[2] Department of Chemistry, University of Toronto, Toronto, Canada

[3] Arnot Research and Consulting (ARC), Toronto, Canada

1 Introduction

The octanol–air equilibrium partitioning ratio (K_{OA}) describes the partitioning of a chemical between octanol and air and is commonly used to approximate the partitioning between various organic phases and the gas phase, including soil organic matter [1, 2] plant foliage [3, 4], atmospheric particulate matter [5, 6], and milk and blood [7]. Previous work on contaminants in outdoor environments (e.g., [8–10]) and in various terrestrial organisms (e.g., [11]) has shown the importance of K_{OA} in understanding the distribution, fate, and bioaccumulation potential of volatile and semi-volatile organic chemicals in the environment. However, these processes often occur at temperatures other than standard temperatures (i.e., 25 °C) and accurate understanding of the partitioning of a chemical in these systems may require a temperature correction.

The temperature dependence of $\log_{10} K_{\mathrm{OA}}$ can be described by the internal energy of phase transfer from octanol to air ($\Delta U^{\circ}_{\mathrm{OA}}$, kJ·mol^{-1}) using the van't Hoff equation:

$$\log_{10}\frac{K_{\mathrm{OA}}\mathrm{at}T_2}{K_{\mathrm{OA}}\mathrm{at}T_1} = -\frac{U^{\circ}_{\mathrm{OA}}}{R}\cdot\left(\frac{1}{273.15+T_2}-\frac{1}{273.15+T_1}\right)\cdot\log_{10}e \tag{1}$$

where R is the ideal gas constant (8.314 ·10^{-3} kJ·K^{-1}·mol^{-1}), T_1 and T_2 are two temperatures (°C), and $\log_{10} e$ (i.e., 0.43) is applied for the logarithm base change. While $\Delta U^{\circ}_{\mathrm{OA}}$ is itself temperature dependent, the derivation of Eq. 1 requires it to be constant over small temperature ranges [12] and therefore a $\Delta U^{\circ}_{\mathrm{OA}}$ can be derived by regressing $\log_{10} K_{\mathrm{OA}}$ against reciprocal temperature $(273.15+T)^{-1}$.

The concentration of a solute in the gas phase can be expressed volumetrically (e.g., in units of mol·m^{-3}) or as a partial pressure (in Pa). If the K_{OA} is calculated using volumetric gas phase concentrations, a regression of $\log_{10} K_{\mathrm{OA}}$ against reciprocal temperature yields $\Delta U^{\circ}_{\mathrm{OA}}$. However, when K_{OA} is defined in terms of partial pressure, the enthalpy of phase transfer ($\Delta H^{\circ}_{\mathrm{OA}}$, kJ·mol^{-1}) is obtained. $\Delta U^{\circ}_{\mathrm{OA}}$ or $\Delta H^{\circ}_{\mathrm{OA}}$ can be converted into each other using [12, 13]:

$$\Delta U^{\circ}_{\mathrm{OA}} = \Delta H^{\circ}_{\mathrm{OA}} + RT \tag{2}$$

The direct determination of the K_{OA} and its temperature dependence using various experimental approaches can be challenging and time consuming. Recent work by Baskaran et al. [14] found that, compared to other K_{OA} prediction techniques, polyparameter linear free energy relationships (ppLFERs) using Abraham descriptors are a reliable, fast and easy-to-use method for estimating K_{OA}. Traditionally ppLFERs using Abraham descriptors combine six system constants and five solute descriptors to describe how a chemical interacts and partitions between two phases [15]. The system constants, represented by lower-case letters, are determined using a multiple linear regression of the property against the solute descriptors, expressed with upper-case letters. These solute descriptors include: *E* (excess molar refraction), *S* (polarizability/dipolarity), *A* (hydrogen bond acidity), *B* (hydrogen bond basicity), *V* (McGowan molar volume), and *L* ($\log_{10}$ of the hexadecane–air partition ratio). The product of a solute descriptor and a system constant, such as *sS* or *bB*, describes the energetic contribution of one particular type of intermolecular interaction to the property [15].

Properties describing the interaction of a solute between the gas phase and a condensed phase can be calculated using two different ppLFERs which differ in the use of either the *V* or *E* parameter.

$$\log_{10} K = c + eE + sS + aA + bB + lL \tag{3}$$

$$\log_{10} K = c + sS + aA + bB + vV + lL \tag{4}$$

Equation 3 is intended to exclusively describe partitioning between the gas phase and a condensed phase, while Eq. 4 can be applied to any two phases, including two condensed phases [15].

In previous work [14] we presented a ppLFER without either V or E to estimate the K_{OA} at 25 °C, that performed as well as the 5-parameter equations by Abraham and Acree [16] and Endo and Goss [17]. Jin et al. [18] also developed a ppLFER equation for $\log_{10} K_{\mathrm{OA}}$ which directly incorporates temperature in the multiple linear regression but used a small and limited data set [14]. In order to predict K_{OA} values at temperatures other than 25 °C, Baskaran et al. [14] recommend using the 4-parameter ppLFER equation for K_{OA} and a ppLFER equation for $\Delta H^{\circ}_{\mathrm{OA}}$ by Mintz et al. [19]. The $\Delta H^{\circ}_{\mathrm{OA}}$ and the $\log_{10} K_{\mathrm{OA}}$ at 25 °C estimated this way were found to be highly correlated ($R^2 > 0.98$) [14], in analogy to previous work that has shown enthalpies of absorption ($\Delta H^{\circ}_{\mathrm{ads}}$) and the logarithm of the adsorption constants ($\log_{10} K_{\mathrm{ads}}$) to be linearly correlated ($R^2 > 0.91$) [20]. Likewise, strong correlations between $\Delta H^{\circ}_{\mathrm{vap}}$ and P_{L} [20, 21] have been observed and Goss and Schwarzenbach [20] note that previous work (e.g., [22]) indicated strong relationships between enthalpies and partitioning ratios. The high correlation between $\log_{10} K_{\mathrm{OA}}$ and $\Delta H^{\circ}_{\mathrm{OA}}$ suggests that the temperature dependence of K_{OA} can be estimated from K_{OA} directly.

Other quantitative structure property relationships (QSPRs) for $\log_{10} K_{\mathrm{OA}}$ use different molecular, topographic, geometric, and quantum-chemical descriptors [12], which require commercial software or intensive computational power. The Open (Quantitative) Structure-activity/property Relationship App (OPERA) model by Mansouri et al. [23], uses two molecular descriptors computed with the PaDEL descriptor software [24] to estimate $\log_{10} K_{\mathrm{OA}}$ at 25 °C. PaDEL calculates 1, 2, and 3 dimensional (1D, 2D, and 3D) molecular descriptors, whereby an increase in the dimensionality corresponds to the complexity of the encoded information [24]; we collectively refer to these descriptors as PaDEL descriptors. The OPERA model utilizes the PaDEL-predicted $\log_{10}$ hexadecane-air partition ratio (L_{PaDEL}) and the number of hydrogen-bond donor atoms (nHBDon), both 2D descriptors [25]. The L_{PaDEL} is identical to the L Abraham solute descriptor, while the nHBDon is somewhat similar to the A Abraham solute descriptor, because they both describe the capacity of the chemical to donate protons.

Both the Abraham and PaDEL descriptors are easily acquired. Two types of Abraham descriptors, experimental and estimated, are available from the UFZ-LSER website [26]. Experimental descriptors are measured directly or through chromatographic retention time techniques [15]. Solute descriptors can also be predicted from a chemical's Simplified Molecular Input Line Entry System (SMILES) notation using the IFS-QSAR models [27] integrated into the UFZ-LSER website and the EAS-E Suite platform [28] or directly from the standalone python package available on GitHub [29]. All PaDEL descriptors can be estimated from the SMILES strings using either a stand-alone software [24] or through QSARINS [30].

Since the development of the $\Delta H^{\circ}_{\mathrm{OA}}$ ppLFER by Mintz et al. [19, 31], a large database of $\log_{10} K_{\mathrm{OA}}$ has been assembled [12]. This database can be used to create a comprehensive dataset of $\Delta U^{\circ}_{\mathrm{OA}}$ for chemicals with measured $\log_{10} K_{\mathrm{OA}}$ data at multiple temperatures. In this work we utilize a newly assembled and curated $\Delta U^{\circ}_{\mathrm{OA}}$ database to develop linear regression models for $\Delta U^{\circ}_{\text{OA}}$ using Abraham solute descriptors, PaDEL descriptors, and/or the $\log_{10} K_{\text{OA}}$ at 25 °C.

2 Methods

2.1 Data Curation

The development of a reliable model for $\Delta U^{\circ}_{\text{OA}}$ relies on acquiring empirical data and processing the data through several data curation steps to reduce errors. We use measured $\log_{10} K_{\text{OA}}$ values from the K_{OA} database [12] to derive experimental $\Delta U^{\circ}_{\text{OA}}$ values. We also searched for directly measured $\Delta H^{\circ}_{\text{OA}}$ and $\Delta U^{\circ}_{\text{OA}}$ values, obtained using calorimetric techniques, to build our training and external validation datasets.

2.1.1 Data from the K_{OA} Database

All measured values of $\log_{10} K_{\text{OA}}$ were extracted from the K_{OA} database and filtered to remove any measurements made (i) using indirect techniques (data from gas chromatographic retention times in octanol-filled columns were considered directly obtained), (ii) using water-saturated octanol, (iii) for a mixture or a chemical with an ambiguous structure, (iv) for inorganic and labelled compounds. We also removed duplicate values and values considered unreliable within the database. Finally, chemicals with measured K_{OA} values at less than four different temperatures were removed, which left 149 chemicals. $\text{Log}_{10} K_{\text{OA}}$ values were linearly regressed against inverse absolute temperature to obtain $\Delta U^{\circ}_{\text{OA}}$ from the slope. We then considered the strength of the correlation based on R^2, the standard error of $\Delta U^{\circ}_{\text{OA}}$ as derived from the error of the slope, and how the $\Delta U^{\circ}_{\text{OA}}$ calculated from this regression compared with published data.

We eliminated 10 chemicals from the dataset because the $\log_{10} K_{\text{OA}}$ values used to calculate $\Delta U^{\circ}_{\text{OA}}$ values were obtained from a single reference and the $\Delta U^{\circ}_{\text{OA}}$ values calculated in our regression disagreed with the published value from that reference. In many instances, the published $\Delta U^{\circ}_{\text{OA}}$ was calculated from a regression of three $\log_{10} K_{\text{OA}}$ values at different temperatures. Including a fourth $\log_{10} K_{\text{OA}}$ measurement (obtained via personal communication with the authors, see [12]) caused the calculated $\Delta U^{\circ}_{\text{OA}}$ to deviate from the reported value. We assume that the authors of these original measurement did not hold strong confidence in these unpublished $\log_{10} K_{\text{OA}}$ measurements. Thus, data for chlorinated dibenzo-*p*-dioxin (CDD) 48, 50, 54, 66 and 73, and brominated diphenyl ether (BDE) 183 were removed from the $\Delta U^{\circ}_{\text{OA}}$ dataset. For *o,p′*-dichlorodiphenyltrichloroethane, *p,p′*-dichlorodiphenyldichloroethane, *cis*-nonachlor and endrin, removing individual $\log_{10} K_{\text{OA}}$ values did not lead to an agreement between previously published and calculated $\Delta U^{\circ}_{\text{OA}}$ values, which suggests that there may be some disparity in the $\log_{10} K_{\text{OA}}$ values used in the regressions (also obtained via direct communication with the authors) and the published data. Four chemicals (isopropyl ether, methane, BDE 99, and BDE 153) in the dataset were removed because the R^2 was below 0.95 and removing any outlier left only three datapoints in the regression. The regressions of the $\log_{10} K_{\text{OA}}$ against inverse temperature for these chemicals are shown in Figs. SI 1–SI 3.

Although there have been multiple measurements of $\log_{10} K_{\text{OA}}$ for perfluoroalkyl substances (PFAS), agreement between data from different studies is poor. As such, all PFAS compounds (perfluorooctane sulfonamido ethanols, perfluorooctane sulfonamide, fluorotelomer alcohols, and fluorotelomer acrylates) were excluded from the $\Delta U^{\circ}_{\mathrm{OA}}$ dataset. The discrepancies between these values is further discussed in Sect. 3.1.1.

By assessing the temperature regression plots for a few chemicals, it was clear that some measured $\log_{10} K_{\mathrm{OA}}$ data deviated from others. In many cases these outliers were from the same paper and often used older and/or objectively less precise measurement techniques (e.g., [32–40]). In these instances, we removed the outliers and recalculated the $\Delta U^{\circ}_{\mathrm{OA}}$ using the remaining $\log_{10} K_{\mathrm{OA}}$ data. In the case of propanol, after removing the obvious outliers from Eger et al. [34], we took the average $\Delta U^{\circ}_{\mathrm{OA}}$ value calculated from regressing temperature dependent $\log_{10} K_{\mathrm{OA}}$ data from Lei et al. [41] and Gruber et al. [42], which subsequently caused the elimination of a single datapoint published by Abraham et al. [43]. Plots of these regressions, including the outliers, are available in Figs. SI 5 and SI 6 in the supporting information.

For some compounds the R^2 for individual sets of temperature dependent $\log_{10} K_{\mathrm{OA}}$ values was high, however, individual $\log_{10} K_{\mathrm{OA}}$ values at specific temperatures deviated by 0.3 to 1 $\log_{10}$ units between papers. Therefore, we took the average $\Delta U^{\circ}_{\mathrm{OA}}$ value calculated from $\log_{10} K_{\mathrm{OA}}$ values from each paper, in other words, we essentially took the average of the two slopes. This applied to hexachlorobenzene (HCB), and the polychlorinated biphenyls (PCBs) 153 and 180 (see Fig. SI 7).

After this step, we were left with 123 chemicals with reliable $\Delta U^{\circ}_{\mathrm{OA}}$ values from the K_{OA} database.

2.1.2 Direct Measurements of $\Delta H^{\circ}_{\mathrm{OA}}$ and $\Delta U^{\circ}_{\mathrm{OA}}$

Mintz et al. [19, 31] used $\Delta H^{\circ}_{\mathrm{OA}}$ values for 138 chemicals to develop ppLFER equations for the enthalpy of phase change between wet and dry octanol and air. $\Delta H^{\circ}_{\mathrm{OA}}$ values used to calibrate those models were cross-referenced against the original sources cited by Mintz et al. [31]. During this process, we found more directly measured $\Delta H^{\circ}_{\mathrm{OA}}$ values in the literature. The $\Delta H^{\circ}_{\mathrm{OA}}$ values were subsequently converted to $\Delta U^{\circ}_{\mathrm{OA}}$ using Eq. 2. We took the average of the $\Delta U^{\circ}_{\mathrm{OA}}$ value when more than one literature value was available.

Any chemical from the Mintz et al. dataset of $\Delta H^{\circ}_{\mathrm{OA}}$ values that already had a $\Delta U^{\circ}_{\mathrm{OA}}$ value calculated from the K_{OA} database was excluded, because these $\Delta H^{\circ}_{\mathrm{OA}}$ values were direct measurements made using calorimetric measurement or were based on $\log_{10} K_{\mathrm{OA}}$ data already included in the K_{OA} database [12]. This way preference was given to values obtained from multiple direct measurements of $\log_{10} K_{\mathrm{OA}}$ over individual calorimetric measurements. 72 chemicals with directly measured $\Delta H^{\circ}_{\mathrm{OA}}$ values (converted to $\Delta U^{\circ}_{\mathrm{OA}}$) were included in our dataset.

2.2 Chemical Identifiers, Descriptors, and Data Splitting

The chemical names, acronym, CAS number and SMILES notation for all 195 chemicals in the $\Delta U^{\circ}_{\mathrm{OA}}$ dataset were obtained from either the K_{OA} database or from the CompTox Dashboard. We considered three types of descriptors for estimating $\Delta U^{\circ}_{\mathrm{OA}}$: (i) the $\log_{10} K_{\mathrm{OA}}$ at 25 °C, (ii) Abraham descriptors, and (iii) PaDEL descriptors.

Abraham solute descriptors (*E*, *S*, *A*, *B*, *V*, and *L*) were obtained from the UFZ-LSER database [26]. For experimental solute descriptors, we gave preference to UFZ pre-selected values over ABSOLV values. Estimated solute descriptors for all chemicals were calculated from their SMILES using the IFS-QSARs built into EAS-E Suite [28, 29]. 178

chemicals had a full set of experimental solute descriptors, 4 chemicals were only missing experimental E values, and 13 chemicals had only estimated solute descriptors.

1444 1D and 2D PaDEL descriptors and 881 PubChem fingerprints (v2.21) were obtained from the chemicals' SMILES notation using QSARINS [30]. 3D molecular descriptors were not included, because while the chemicals in the $\Delta U^{\circ}_{\mathrm{OA}}$ dataset are structurally simple, requiring 3D geometrically optimized chemical structures (e.g., SDF or MOL) and 3D optimization cannot be efficiently scaled up for high throughput applications and as such is not ideal for property estimations in the context of chemical risk assessment. Filters were applied to the PaDEL descriptors, wherein descriptors that had pair-wise correlation greater than 95% or where more than 80% of the descriptor was constant for the whole dataset were excluded. We also removed all descriptors that had missing values. In the end 531 PaDEL descriptors and fingerprints were considered, which included PaDEL calculated Abraham descriptors and 32 PubChem fingerprints. PubChem Fingerprints were excluded from model development and only used in a cluster analysis for data splitting as described later in this section.

Directly measured $\log_{10} K_{\mathrm{OA}}$ values were extracted from the K_{OA} database [12] and the average used when more than one experimental value existed. Estimated $\log_{10} K_{\mathrm{OA}}$ values were calculated using the experimental and calculated Abraham solute descriptors and the 4-parameter ppLFER for K_{OA} [14].

The curated $\Delta U^{\circ}_{\mathrm{OA}}$ dataset was split into a model development and an external validation dataset based on the availability of descriptors for each chemical. 54 chemicals that had either some missing experimental Abraham solute descriptors or an estimated $\log_{10} K_{\mathrm{OA}}$ values were set aside to be used for external validation. The remaining 141 chemicals were used to develop multiple linear regression (MLR) models using the descriptors.

In order to reduce bias in descriptor selection, we split the development dataset using four different splitting techniques, with a ratio of 3:1 (i.e., 75% of chemicals were used in the training set and 25% in the validation set) [44]. In the first and second splits, chemicals were ordered by $\Delta U^{\circ}_{\mathrm{OA}}$ and $\log_{10} K_{\mathrm{OA}}$ at 25 °C, respectively, and the chemicals with the highest and lowest value were included in the training dataset. In the random split, 75% of chemicals were randomly selected to be a part of the training dataset. We used cluster analysis (Ward's method and Tanimoto distance) using PubChem fingerprints and Principal Component Analysis (PCA) of experimental Abraham solute descriptors to group chemicals in the model development dataset into 4 structural clusters (see Figs. SI 11 and SI 12). From each cluster we randomly selected 75% of the chemicals to be used as a training dataset. Figure 1 shows how the chemical datasets were split and used to develop $\Delta U^{\circ}_{\mathrm{OA}}$ models. Table SI 2 contains $\Delta U^{\circ}_{\mathrm{OA}}$ values, descriptors, and splits for all chemicals in the $\Delta U^{\circ}_{\mathrm{OA}}$ dataset.

2.3 Model Calculation and Selection

Model calculation and selection were completed using QSARINS [30] on the four different splits of the model development dataset. For each split we developed different models based on the different kind of molecular descriptors. First, we explored the traditional ppLFER equations assessing all combinations (all-subsets with 1 to 5 descriptors) using

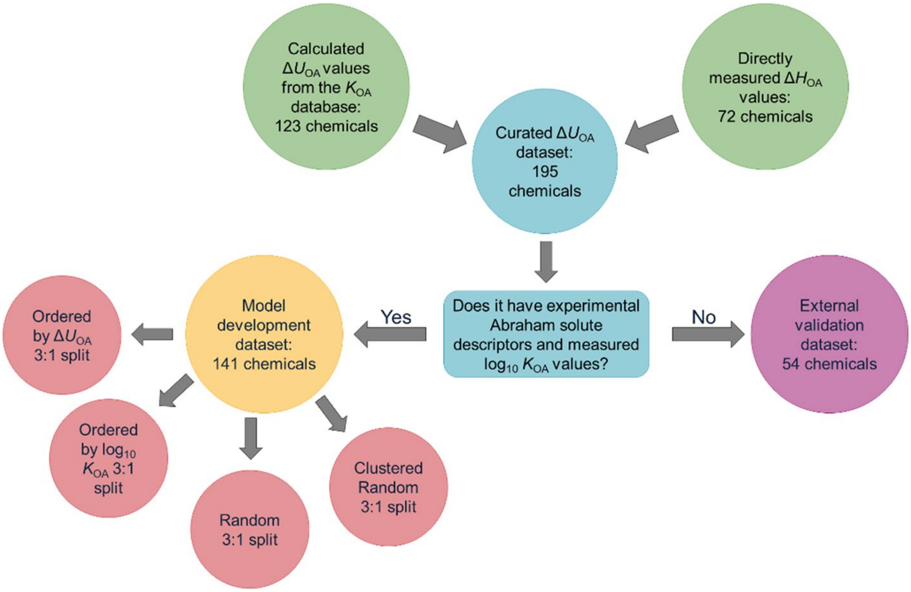
Calculated ΔU_{OA} values from the K_{OA} database: 123 chemicals
Directly measured ΔH_{OA} values: 72 chemicals
Curated ΔU_{OA} dataset: 195 chemicals
Does it have experimental Abraham solute descriptors and measured $\log_{10} K_{OA}$ values?
Yes
No
External validation dataset: 54 chemicals
Model development dataset: 141 chemicals
Ordered by ΔU_{OA} 3:1 split
Ordered by $\log_{10}$ K_{OA} 3:1 split
Random 3:1 split
Clustered Random 3:1 split

Fig. 1 A scheme of the source of ΔU°_{OA} data and how this dataset is split and used for model development and validation. 3:1 split indicates that 75% of the chemicals in the dataset were used for the training data and the remaining 25% of chemicals were included in an internal validation process

experimental Abraham descriptors. Then, we substituted the Abraham L with the experimental $\log_{10}$ K_{OA} at 25 °C, as both variables describe partitioning between the gas phase and a hydrophobic condensed phase. Given the high number of PaDEL descriptors, we used the all-subset (up to 2 variables) and the Genetic Algorithm (up to 4 variables, parameters are included in the SI) built into QSARINS to identify the best PaDEL descriptors combination. Finally, we assessed the performance of the model using the two PaDEL descriptors used by the OPERA model for estimating $\log_{10}$ K_{OA}, namely the number of hydrogen bond donors (*n*HBDon) and $\log_{10}$ hexadecane-air partition ratio (L) [25].

During model development and selection, we filtered out models with p-values for the regression coefficients higher than 0.05, as we could not be confident that the values of these coefficients were different from 0. We restricted the maximum number of variables (i.e., descriptors) to be included in a regression equation to 4 to ensure models remained relatively simple and to avoid overfitting issues. Model performance across all four splits was assessed using different variations of the determination coefficient (R^2_{adj}, Q^2_{Loo}, Q^2_{LMO}, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}), the concordance correlation coefficients (CCC_{EXT}), root mean squared errors ($RMSE_{TR}$, $RMSE_{EXT}$) and mean absolute errors (MAE_{TR}, MAE_{EXT}) on the training set and on the prediction/external set [45–48]. Table SI 1 summarises these statistics for the models within each split.

After identifying the best models for different types of descriptors, we used all 141 ΔU°_{OA} values as the training set to develop new equations using the same descriptors (i.e., the so-called full model). These models were assessed using the 54 chemicals in the external validation dataset. It is important to note that in the external validation set, not all chemicals have empirical descriptors namely for A, L, and $\log_{10}$ K_{OA} at 25 °C. When experimental A and L values were not available, IFS-QSAR estimated solute descriptors were used. Log_{10} K_{OA} values at 25 °C were estimated using the

4-parameter ppLFER equation for K_{OA} [14] using experimental solute descriptors. All PaDEL descriptors are calculated from the SMILES notation of a chemical and so all descriptors are considered estimates.

3 Results and Discussion

3.1 Data Availability and the General Applicability Domain

Analysis of the $\log_{10} K_{\mathrm{OA}}$ database showed a bimodal distribution in the experimental data [12]. Given the high correlation between $\log_{10} K_{\mathrm{OA}}$ and $\Delta U^{\circ}_{\mathrm{OA}}$ [14], it is not surprising that the same trend is observed for the curated $\Delta U^{\circ}_{\mathrm{OA}}$ dataset (Fig. 2). This pattern arises because most $\Delta U^{\circ}_{\mathrm{OA}}$ values in this dataset are derived from temperature dependent measurements of the $\log_{10} K_{\mathrm{OA}}$ and thus limitations on the experimental techniques for $\log_{10} K_{\mathrm{OA}}$ are extended to the available $\Delta U^{\circ}_{\mathrm{OA}}$ measurements. Direct measurements of $\Delta U^{\circ}_{\mathrm{OA}}$ using the calorimetric techniques are applied to small volatile compounds and the $\Delta U^{\circ}_{\mathrm{OA}}$ of these compounds are less negative and fall within a limited range. The most common class of chemicals in the $\Delta U^{\circ}_{\mathrm{OA}}$ dataset are PCBs (Fig. SI 13), followed by polychlorinated naphthalenes (PCNs), alcohols, and ketones.

Most chemicals in the $\Delta U^{\circ}_{\mathrm{OA}}$ dataset have A values of 0 and B values less than 0.6 (Fig. SI 14), however only 36 (~ 18%) chemicals have both A and B values equal to 0. The polarity of a chemical (defined as $aA + bB$, where the system descriptors are from the four parameter ppLFER equation for K_{OA} [12, 14]), shows there is some variability in the hydrogen bonding ability with octanol of the chemicals in the $\Delta U^{\circ}_{\mathrm{OA}}$ dataset. However, there are few very polar compounds. Descriptor values for E and S are more widely distributed, while the distributions of the L and V descriptors resembles the distribution of $\Delta U^{\circ}_{\mathrm{OA}}$ values.

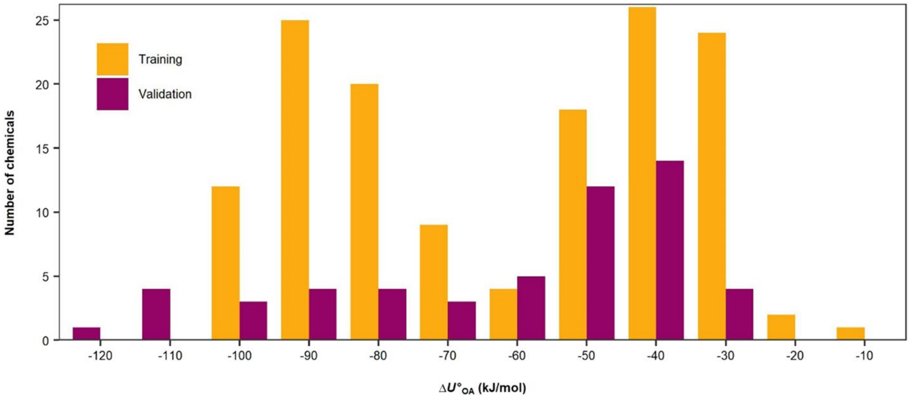

Fig. 2 Distribution of $\Delta U^{\circ}_{\mathrm{OA}}$ data used to train and validate the models. Chemicals were included in the training data when all experimental Abraham solute descriptors and an experimental $\log_{10} K_{\mathrm{OA}}$ value at 25 °C were available

3.1.1 Perfluoroalkyl Substances (PFAS)

The partitioning properties of PFAS compounds are extremely difficult to measure because they can act as surfactants and likely have a very low solubility in octanol. Data from different studies regularly display divergent results and it is not possible to establish which ones are correct.

Most fluorotelomer alcohols (FTOHs) have $\log_{10} K_{\mathrm{OA}}$ values around 5, which is at the lower limits of the generator column technique. They are also too polar to apply the gas-chromatographic retention time technique [49]. In combination, this means obtaining reliable measurements for the K_{OA} of FTOHs is quite difficult. Figure SI 15 compares the $\Delta U^{\circ}_{\mathrm{OA}}$ values of FTOHs calculated from K_{OA} measurements reported by Goss et al. [50] and Thuens et al. [51] (temperature regression plots in Fig. SI 10). We would expect that $\Delta U^{\circ}_{\mathrm{OA}}$ values for FTOHs become more negative with increasing chain length, because the K_{OA} values increase with chain length. While this occurs for $\Delta U^{\circ}_{\mathrm{OA}}$ values from Goss et al. [50], $\Delta U^{\circ}_{\mathrm{OA}}$ declines with increasing chain length when using the data from Thuens et al. [51]. As the $\Delta U^{\circ}_{\mathrm{OA}}$ values from Goss et al. [50] have been calculated from measurements at only two (4:2 FTOH) and three (6:2 FTOH and 8:2 FTOH) temperatures, there is insufficient data to include $\Delta U^{\circ}_{\mathrm{OA}}$ values for FTOHs.

The temperature dependence of the K_{OA} for methyl and ethyl perfluorooctane sulfonamido ethanols (Me-FOSE and Et-FOSE) reported by two studies [52, 53] and therefore also the $\Delta U^{\circ}_{\mathrm{OA}}$ derived from those data are quite different (Fig. SI 8). Dreyer et al. [52] also published K_{OA} data on methyl and ethyl perfluorooctane sulfonamides (Me-FOSA and Et-FOSA), whereby the plot suggests that the values obtained at very low and high temperatures may be outliers. Given the uncertainty of the K_{OA} values for Me- FOSA, Et-FOSA, and Et-FOSE, we have also excluded those for Me-FOSE as there is no way for us to tell whether they are valid. Similar to Et-FOSE, the $\log_{10} K_{\mathrm{OA}}$ reported by Dreyer et al. [52] for 6:2 and 8:2 fluorotelomer acrylates (6:2 FTAc and 8:2 FTAc) also display poor linearity with reciprocal temperature ($R^2 < 0.9$) and therefore we chose to exclude also the K_{OA} data for 10:2 FTAc from the $\Delta U^{\circ}_{\mathrm{OA}}$ dataset (Fig. SI 9). Ultimately, we decided to exclude all PFAS compounds from the data set.

3.2 Model Selection

3.2.1 Model Prioritization

As not all chemicals with $\Delta U^{\circ}_{\mathrm{OA}}$ values had all available descriptors, a subset of 141 chemicals was used to test and identify the best performing models within each split. We selected five models to examine in detail. Those are the regressions using (i) A and L, (ii) only the $\log_{10} K_{\mathrm{OA}}$ at 25 °C, (iii) A and $\log_{10} K_{\mathrm{OA}}$ at 25 °C, (iv) nHBDon and L_{PaDEL}, and (v) nHBDon, number of carbons, nC, and number of halogens, nX. As initially hypothesized, the $\log_{10} K_{\mathrm{OA}}$ at 25 °C alone proved to be a very good predictor of $\Delta U^{\circ}_{\mathrm{OA}}$ (Table SI 1). Meanwhile the model using the Abraham solute descriptors A and L performed almost as well. The $\log_{10}$ hexadecane-air partition ratio L can describe the non-polar interactions between the compound and the solvent while the hydrogen bond acidity A describes the potential for a chemical to donate a hydrogen. These two parameters describe similar chemical interactions as the two PaDEL descriptors L_{PaDEL} and nHBDon, used to predict $\log_{10} K_{\mathrm{OA}}$ in the OPERA model.

Table 1 Proposed equations for calculating the $\Delta U^{\circ}_{\mathrm{OA}}$ (kJ·mol^{-1}) of an organic compound

No	Equation
1	$\Delta U^{\circ}_{\mathrm{OA}} = -56.40(\pm1.32)\cdot A - 8.58(\pm0.20)\cdot L - 5.28(\pm1.32)$
2	$\Delta U^{\circ}_{\mathrm{OA}} = -8.75(\pm0.18)\cdot \log_{10} K_{\mathrm{OA}} - 5.07(\pm1.19)$
3	$\Delta U^{\circ}_{\mathrm{OA}} = -8.92(\pm0.18)\cdot \log_{10} K_{\mathrm{OA}} - 18.34(\pm4.72)\cdot A - 3.37(\pm1.21)$
4	$\Delta U^{\circ}_{\mathrm{OA}} = -16.63(\pm2.28)\cdot n\mathrm{HBDon} - 7.55(\pm0.21)\cdot L_{\mathrm{PaDEL}} - 9.26(\pm1.52)$
5	$\Delta U^{\circ}_{\mathrm{OA}} = -20.64(\pm2.47)\cdot n\mathrm{HBDon} - 3.78(\pm0.22)\cdot nC - 6.12(\pm0.31)\cdot nX - 9.82(\pm1.88)$

Table 2 Internal validation results for the five selected models with all 141 chemicals in the training dataset

No	Variables	R^2	R^2_{adj}	$\mathrm{RMSE_{TR}}$	$\mathrm{MAE_{TR}}$	$\mathrm{RMSE_{CV}}$	$\mathrm{MAE_{CV}}$	Q^2_{LOO}	Q^2_{LMO}
1	A, L	0.94	0.93	6.37	4.27	6.48	4.35	0.93	0.93
2	$\log_{10} K_{\mathrm{OA}}$	0.94	0.94	6.02	4.58	6.11	4.65	0.94	0.94
3	$\log_{10} K_{\mathrm{OA}}, A$	0.95	0.95	5.72	4.15	5.82	4.24	0.95	0.94
4	$n\mathrm{HBDon}, L_{\mathrm{PaDEL}}$	0.90	0.90	7.73	5.41	7.87	5.52	0.90	0.90
5	nHBDon, nC, nX	0.89	0.89	8.15	6.01	8.40	6.19	0.89	0.89

The model selected with the Genetic Algorithm using only PaDEL descriptors uses *n*HBDon, *nC*, and *nX*. These descriptors are not able to describe chemicals to the same extent Abraham descriptors do and their selection is a consequence of the data used to train the models. This is discussed in more detail in the following section.

All five models performed well during internal and external validation processes across all four splits (Table SI 1). The R^2_{adj} describes the R^2 while correcting for the number of descriptors in a model. The $\mathrm{RMSE_{TR}}$ and $\mathrm{MAE_{TR}}$ was in almost all cases slightly smaller than the $\mathrm{RMSE_{EXT}}$ and $\mathrm{MAE_{EXT}}$, however given the range of the $\Delta U^{\circ}_{\mathrm{OA}}$ values used in model development, the difference is likely unremarkable. Internal cross validation using leave-one-out (Q^2_{LOO}) and leave-more-out (Q^2_{LMO}) was very similar to the R^2 and R^2_{adj} for all models (> 0.90). However, the external predictive ability of the models (Q^2_{F1}, Q^2_{F2}, Q^2_{F3}) were slightly lower than the R^2. The $\mathrm{CCC_{EXT}}$, describing both the precision and accuracy of all models was almost always greater than 0.90.

3.2.2 External Validation

Of the 54 chemicals in the external validation dataset, 17 chemicals had experimental $\log_{10} K_{\mathrm{OA}}$ values and 41 chemicals had experimental *A* and *L* solute descriptors that were retrieved from the UFZ-LSER database. All chemicals without experimental $\log_{10} K_{\mathrm{OA}}$ values had all the experimental solute descriptors necessary to estimate $\log_{10} K_{\mathrm{OA}}$ using the four parameter ppLFER equation [14]. As the PaDEL descriptors for all chemicals were estimated, we did not differentiate between the source of *A*, *L*, and $\log_{10} K_{\mathrm{OA}}$ at 25 °C used to externally validate the model, however wherever possible we used experimental $\log_{10} K_{\mathrm{OA}}$ values and solute descriptors. Full model equations are listed in Table 1.

By using all 141 chemicals to train the models and 54 chemicals to externally validate the models we see that Model 3 using $\log_{10} K_{\mathrm{OA}}$ and *A* had the best overall performance

Table 3 External validation results for the five selected models with all 141 chemicals in the training dataset and 54 chemicals used for external validation

No	Variables	RMSE_{EXT}	MAE_{EXT}	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC_{EXT}	$h_{ii} > h^*$	$\lvert sr_i \rvert > 2.5\sigma$
1	A, L	7.23	5.40	0.92	0.92	0.92	0.96	7	9
2	$\log_{10} K_{\text{OA}}$	6.86	5.20	0.93	0.93	0.92	0.96	0	7
3	$\log_{10} K_{\text{OA}}, A$	6.76	5.20	0.93	0.93	0.93	0.96	7	9
4	nHBDon, L_{PaDEL}	6.50	4.85	0.93	0.93	0.93	0.97	15	4
5	nHBDon, nC, nX	10.96	7.62	0.81	0.81	0.81	0.89	2	11

Included is the number of chemicals in the training dataset with very high leverage values ($h_{ii} > h^*$) and the number of chemicals with large standardized residuals ($\lvert \text{sr}_i \rvert > 2.5\sigma$)

(Tables 2, 3). Model 3 has the highest overall R^2, R^2_{adj}, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, and CCC_{EXT}. Model 2, using $\log_{10} K_{\text{OA}}$, performs almost as well as Model 3, followed by Model 1. Models 4 and 5 had the poorest performance.

In Fig. 3, we can see that all models generally perform better at higher $\Delta U^{\circ}_{\text{OA}}$ (i.e., less negative) values. Models 4 and 5 appear to have a larger number of outliers (> 10 kJ·mol^{-1}) at lower $\Delta U^{\circ}_{\text{OA}}$ values relative to Models 1, 2, and 3; however, it is important to recognize that this relative difference between predicted and experimental values is smaller for increasingly negative $\Delta U^{\circ}_{\text{OA}}$ values. An error of − 10 kJ·mol^{-1} for a $\Delta U^{\circ}_{\text{OA}}$ of − 90 kJ·mol^{-1} corresponds to ~ 11% whereas the same error for a $\Delta U^{\circ}_{\text{OA}}$ of − 30 kJ·mol^{-1} is ~ 33%.

In Fig. SI 16 we present Williams plots of the standardized residual plotted against the leverage. The standardized residual (sr_i) of a chemical i is obtained by correcting residuals (r_i) using the standard deviation (sd) of the model and the leverage of a prediction (h_{ii}):

$$sr_i = \frac{r_i}{\text{sd} \cdot \sqrt{1 - h_{ii}}} \tag{5}$$

Standardized residual values are considered high when the absolute value is greater than 2.5 [44]. The leverage of a training chemical indicates how much influence it has on the regression; chemicals with high leverages have a large influence on the model [44]. A chemical has a high leverage value if the leverage is greater than the warning leverage (h^*) which is a function of the number of chemicals in the training dataset (n) and the number of parameters in the model (k) [44]:

$$h^* = 3 \cdot \frac{k + 1}{n} \tag{6}$$

Standardized residuals and leverage values were obtained from QSARINS.

Considering only the number of chemicals with high leverage and high standardized residuals (Table 3), Model 2 has no chemical in the training set with a large leverage value and it has only 7 chemicals with high standardized residuals. The largest standardized residual was for 1,2,3,4-tetramethylbenzene in the external validation dataset; however, the $\log_{10} K_{\text{OA}}$ (4.4) measured using the variable headspace ratio technique is known to be at the upper limit of the technique [41]. Thus, it likely that this high standardized residual value is due to the high uncertainty of the reported K_{OA}, $\Delta U^{\circ}_{\text{OA}}$ or both, rather than an error in the model.

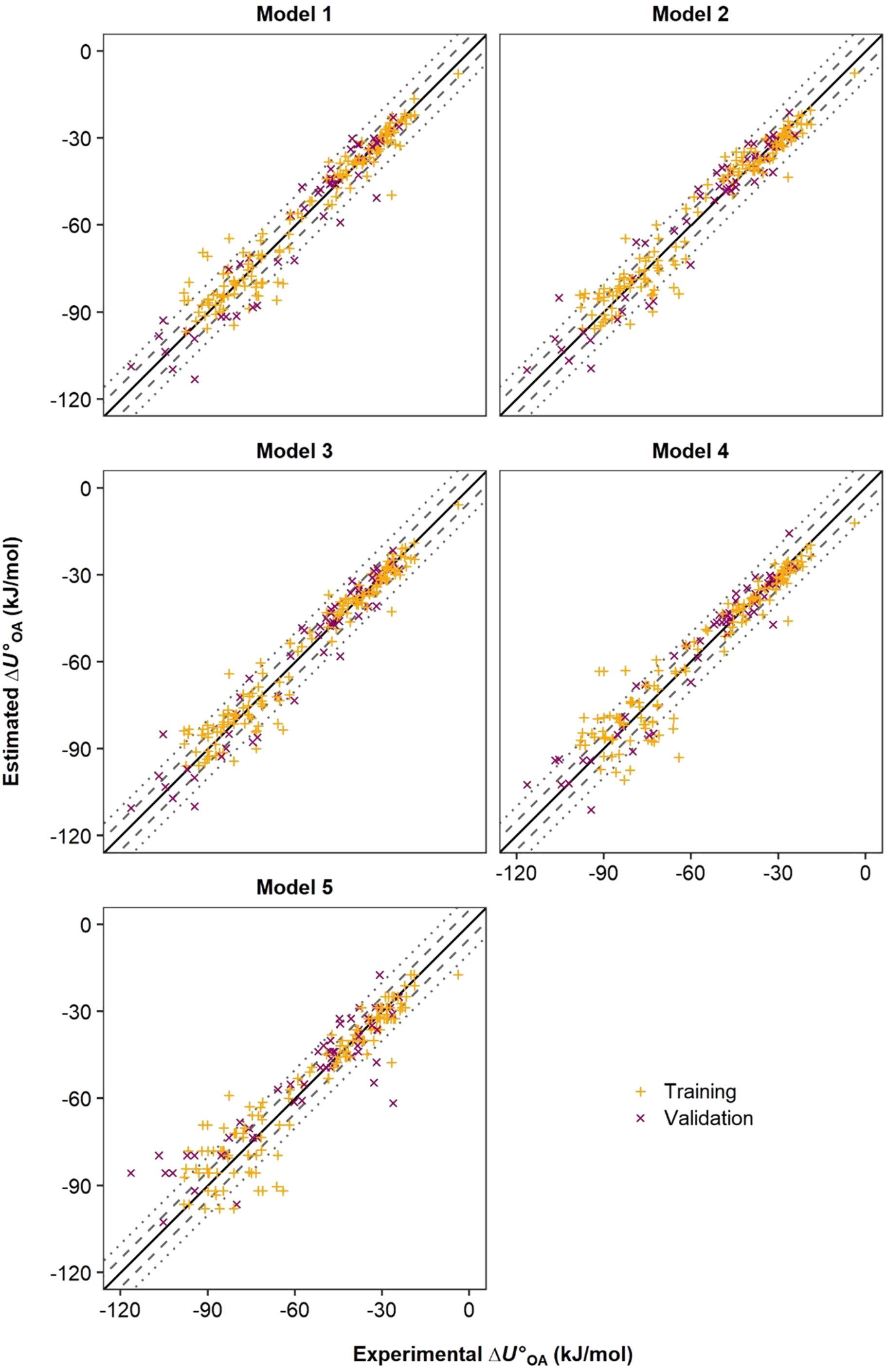

Fig. 3 Estimated $\Delta U^{\circ}_{\mathrm{OA}}$ values plotted against experimental values for the training and external validation compounds. 1:1 lines are shown in black. Grey dashed and dotted lines indicate ±5 and ±10 kJ·mol^{-1}, respectively

Models 1 and 3 had the same chemicals (external validation and training chemicals) with high leverage values. This high leverage is driven entirely by the *A* parameter used in both models. Chemicals with *A* greater than 0.35 have high leverage values. This is because most chemicals in the $\Delta U^{\circ}_{\mathrm{OA}}$ dataset have *A* values equal to 0. For example, methanoic acid (in the external validation dataset) has both a high leverage values and high standardized residuals in Models 1 and 3 and has the highest *A* value of 0.76.

Many of the same chemicals with high leverage values in Models 1 and 3 also had high leverage values in Model 4, driven by the *n*HBDon values. While *n*HBDon is a countable parameter, the $\Delta U^{\circ}_{\mathrm{OA}}$ dataset only includes chemicals with either *n*HBDon values equal to 0 or 1. Thus, all chemicals with an *n*HBDon value equal to 1 had a high leverage value. Nonetheless, these chemicals had low standardized residuals. The three chemicals with high leverage in Model 5 (2 training and 1 external validation compounds) all have long carbon chains. However, they are not the only chemicals of that size.

There was also an overlap in chemicals with high standardized residuals between models. Fluorene had a high standardized residual in all models. PCB 155 and 1,2,3,4-tetramethylbenzene had high standardized residuals in all models except Model 1 and 4, respectively. The standardized residuals for BDE 154 and *trans*-nonachlor was high in Models 1, 2, and 3, while β-HCH and δ-HCH had high standardized residuals in Models 1, 4, and 5. Heptachlor had high standardized residuals in model 1 and 5, and PCBs 61 and 77 had high standardized residuals in models 2 and 3.

The internal validation results presented in Table 2 show Models 1, 2, and 3 perform better than Models 4 and 5 based on R^2, R^2_{adj}, $\mathrm{RMSE_{TR}}$, $\mathrm{MAE_{TR}}$, $\mathrm{RMSE_{CV}}$, $\mathrm{MAE_{CV}}$, Q^2_{LOO}, and Q^2_{LMO}. Model 5 has the highest $\mathrm{RMSE_{EXT}}$ followed by Model 1. Model 5 also has the highest $\mathrm{MAE_{EXT}}$ and lowest $Q^2_{\mathrm{F1\text{-}F3}}$ and $\mathrm{CCC_{EXT}}$. Despite Model 4's comparatively poor internal validation results, the external validation results ($\mathrm{RMSE_{EXT}}$, $\mathrm{MAE_{EXT}}$, $Q^2_{\mathrm{F1\text{-}F3}}$, $\mathrm{CCC_{EXT}}$) for this model are similar if not better than results for Models 1, 2, and 3.

While Models 1 and 3 have excellent internal and external validation results, the applicability domain of these models is limited by the *A* descriptor. Model 4's internal validation results are poor compared to the other models, and are limited to chemicals with *n*HBDon values of 0. Model 5 was the worst performing model with high internal and external errors (RMSE and MAE values). Model 2 is consistently one of the best models when considering the internal and external validations, the number of outliers (high standardized residuals) and the number of chemicals with high leverage. Thus, the $\log_{10} K_{\mathrm{OA}}$ at 25 °C (Model 2) is generally sufficient to make a quick and reliable estimation of the $\Delta U^{\circ}_{\mathrm{OA}}$.

3.3 Comparison with Other $\Delta U^{\circ}_{\mathrm{OA}}$ Models

The $\Delta H^{\circ}_{\mathrm{OA}}$ ppLFER model by Mintz et al. [19] has been shown to reliably predict the temperature dependence of K_{OA} [14]. We compared the performance of this ppLFER for $\Delta H^{\circ}_{\mathrm{OA}}$ [19], converted to $\Delta U^{\circ}_{\mathrm{OA}}$, with Model 2. Since both models are built on similar datasets and an evaluation with chemicals that had been used to train and develop the models would bias the results, we rely only on chemicals not used in the training dataset for either model. This limits the comparison to 28 chemicals including several BDEs and CDDs. The training data for the Mintz et al. [19] model includes 138 chemicals and experimental solute descriptors and experimental $\Delta H^{\circ}_{\mathrm{OA}}$ values for all chemicals are provided in a previous publication [31]. During our data curation process we found small errors in the reported $\Delta H^{\circ}_{\mathrm{OA}}$ and misidentification of a few chemicals which means some of the descriptors used are erroneous. The Mintz et al. [19] model also includes inorganic chemicals (e.g., nitrogen

Table 4 Statistics on the residual (kJ·mol^{-1}) for ΔU°_{OA} predictions made using Model 2 ($\log_{10} K_{OA}$) and the Mintz et al. [19] model for chemicals not included in the training set of either model ($n=28$) and all chemicals in the ΔU°_{OA} dataset ($n=195$)

Model	Dataset	n	Bias	MAE	SD	RMSE	R^2	|residual|> 10	|residual|> 5
Mintz et al	all chemicals	195	0.75	4.86	6.98	7.00	0.96	30	69
Model 2	all chemicals	195	-0.13	4.75	6.28	6.27	0.97	19	70
Mintz et al	external only	28	5.07	7.99	8.94	10.14	0.95	10	17
Model 2	external only	28	1.32	5.96	7.89	7.86	0.96	5	14

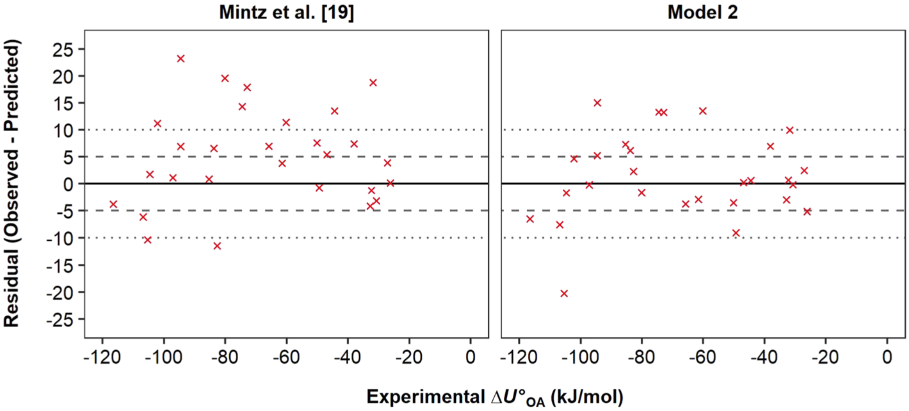

Fig. 4 Comparing the performance of Model 2 ($\log_{10} K_{OA}$) with the Mintz et al. model [19] using chemicals not included in the training set of both models. Grey dashed lines indicate errors of ± 5 kJ·mol^{-1} and grey dotted lines indicate errors of ± 10 kJ·mol^{-1}

and xenon) which were excluded from our model development. Due to these differences, we compare the statistics of the model residuals (Table 4) and the residual of each prediction (Fig. 4).

For both models we use the best available descriptors (i.e., experimental values wherever possible and otherwise IFS-QSAR estimated values). In this combined external validation process (Table 4), Model 2 has a smaller bias (average residual), MAE, standard deviation (SD), and RMSE. The Mintz et al. [19] model has a larger number of chemicals with errors exceeding 10 kJ·mol^{-1} and Model 2 has more ΔU°_{OA} estimates with absolute errors less than 5 kJ·mol^{-1} (Fig. 4).

If we compare the performance of the Mintz et al. [19] model and Model 2 when using all 195 chemicals in the ΔU°_{OA} dataset, we see that Model 2 performs better than the Mintz et al. [19] model (Table 4). While there is little difference between the statistics on the residuals, fewer chemicals have residuals greater than 10 kJ·mol^{-1} when using Model 2. Both models overestimated the ΔU°_{OA} by at least 10 kJ·mol^{-1} for fluorene, PCN 57, *cis*-chlordane, *trans*-chlordane, and *trans*-nonachlor—all chemicals within the training data set of Model 2, apart from *trans*-nonachlor. The models also underestimated the ΔU°_{OA} for CDD 1, 1,2,3,4-tetramethylbenzene, BDE 17, BDE 28, BDE 154, PCB 61, PCB 155, of which CDD 1 and BDEs 17, 28, and 154 were in the external validation dataset for Model 2.

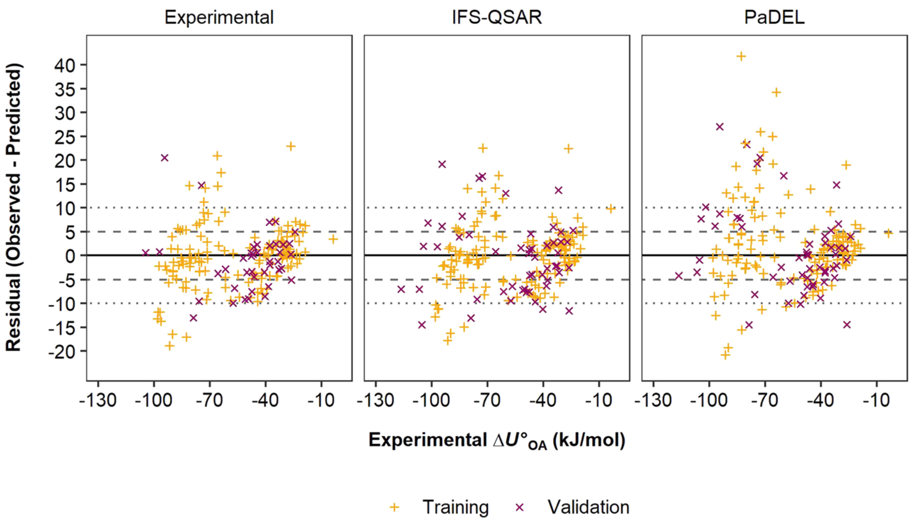

Fig. 5 Residuals between observed $\Delta U^{\circ}_{\mathrm{OA}}$ and the $\Delta U^{\circ}_{\mathrm{OA}}$ predicted by Model 2 depending on whether experimental, IFS-QSAR-predicted, or PaDEL-predicted Abraham descriptors were used. Grey dashed lines indicate errors of ± 5 kJ·mol^{-1} and grey dotted lines indicate errors of ± 10 kJ·mol^{-1}

Table 5 Statistics on the residuals for the performance of Model 2 using estimated K_{OA} values using experimental, IFS-QSAR-predicted, and PaDEL-predicted Abraham descriptors

Descriptors	*n*	Mean	MAE	SD	RMSE	|residual|> 10	|residual|> 5
Experimental	182	−0.26	4.65	6.49	6.48	17	64
IFS-QSAR	195	−0.20	5.31	7.21	7.19	25	78
PaDEL	195	2.09	6.32	8.99	9.21	37	81

In addition to the model by Mintz et al. [19], there is a ppLFER for wet-octanol air ΔH° values [31] and models for $\Delta H^{\circ}_{\mathrm{OA}}$ using a support vector machine method, artificial neural networks, and MLRs based on various molecular descriptors [54]. The wet-octanol air $\Delta H^{\circ}_{\mathrm{OA}}$ model is based on the same dataset used to develop the Mintz et al. [19] model above [31]. The latter three models also use the same dataset as a basis for model development but randomly split their dataset of 127 chemicals into a training dataset of 89 chemicals and validation dataset of 38 chemicals [54]. The reported correlation coefficients for these models [31, 54] are similar to, or greater than what we report for Model 2 and the standard errors of the models developed using a support vector machine method and artificial neural networks is smaller than the RMSE of Model 2. However, these models cannot be used easily in high throughput applications because acquiring the molecular descriptors requires multiple software applications [54], which can be time-consuming and lends itself to increased chances of human error.

3.4 Sources for Descriptors

While we used only experimentally derived solute descriptors to develop and train the models in this work, the source of solute descriptors may impact model performance.

Sources for the descriptors include experimental values from the UFZ-LSER database [26], IFS-QSAR estimates from the standalone IFS-QSAR software [27, 29] or EAS-E Suite [28], or PaDEL estimates made using QSARINS [30]. The first two are easily obtained from non-commercial web-based software. In Fig. 5 we compare Model 2's estimated $\Delta U^{\circ}_{\mathrm{OA}}$ if only experimental, IFS-QSAR predicted, or PaDEL predicted solute descriptors are used. The $\log_{10} K_{\mathrm{OA}}$ at 25 °C used by Model 2 is calculated using the ppLFER equation with the 4 parameters S, A, B, and L [14]. Because not all chemicals have experimental solute descriptors, fewer estimations are available for comparison. Table 5 shows that the IFS-QSAR-estimated solute descriptors perform almost as well as experimental solute descriptors. Regardless of which set of descriptors is used, $\Delta U^{\circ}_{\mathrm{OA}}$ values of 1-decanol, β-HCH, δ-HCH, fluorene, dieldrin, BDE 29, PCB 77, and PCN 57 are overestimated and $\Delta U^{\circ}_{\mathrm{OA}}$ values of 1,2,3,4-tetramethylbenzene, BDE 154, PCB 61, PCN 180, and PCB 155 are underestimated. Clearly, the IFS-QSARs are preferable over PaDEL for predicting solute descriptors, considering the smaller associated MAE, standard deviation, and RMSE for the former.

3.5 The Relationship Between $\log_{10}$ *K* and ΔU°

The strong relationship between $\log_{10} K_{\mathrm{OA}}$ and $\Delta U^{\circ}_{\mathrm{OA}}$ in Model 2 mirrors similar relationships between vapour pressure ($\log_{10} P_{\mathrm{L}}$) and the enthalpy of vaporization ($\Delta H^{\circ}_{\mathrm{vap}}$) and between adsorption constants ($\log_{10} K_{\mathrm{ads}}$) and the enthalpies of adsorption ($\Delta H^{\circ}_{\mathrm{ads}}$) [20]. Goss and Schwarzenbach cite other such relationships and provide a thermodynamic explanation for their occurrence [20]. The equation relating $\Delta H^{\circ}_{\mathrm{vap}}$ and P_{L} [20] can be combined with a linear equation relating a chemical's P_{L} with its $\log_{10} K_{\mathrm{OA}}$ [55] and a conversion to $\Delta U^{\circ}_{\mathrm{vap}}$ ($\Delta U^{\circ}_{\mathrm{vap}} = \Delta H^{\circ}_{\mathrm{vap}} - \mathrm{RT}$) [13]:

$$\Delta U^{\circ}_{\mathrm{vap}} = 8.90 \cdot \log_{10} K_{\mathrm{OA}} + 10.42 - RT \tag{7}$$

The slope in this equation is very similar to the absolute value of the slope in Model 2 (− 8.75). This is apparent in a vertical shift when estimated $\Delta U^{\circ}_{\mathrm{OA}}$ and $-\Delta U^{\circ}_{\mathrm{vap}}$ values are plotted against experimental $\Delta U^{\circ}_{\mathrm{OA}}$ values (Fig. SI 18). The difference in Eq. 2 in Table 1 and Eq. 7 corresponds to the internal energy of dissolution of the chemical in octanol (i.e., $\Delta U^{\circ}_{\mathrm{dissol}} = \Delta U^{\circ}_{\mathrm{OA}} + \Delta U^{\circ}_{\mathrm{vap}}$). This analysis suggests that the $\Delta U^{\circ}_{\mathrm{dissol}}$ is very small and is relatively constant for compounds with a wide range in K_{OA}. The observed relationship between K_{OA} and $\Delta U^{\circ}_{\mathrm{OA}}$ can perhaps be extrapolated to other partitioning systems such as K_{OW} and $\Delta U^{\circ}_{\mathrm{OW}}$. Indeed, if $\Delta U^{\circ}_{\mathrm{dissol}}$ is a small and constant for a wide range of compounds it may be possible to estimate $\Delta U^{\circ}_{\mathrm{OW}}$ from the enthalpy of dissolution of a chemical in water or even the water solubility of a compound. However large datasets with experimental values of each property are necessary before such a relationship can be verified.

4 Conclusion

We developed new models for predicting the $\Delta U^{\circ}_{\mathrm{OA}}$ which use fewer descriptors than earlier models. The best performing model (Model 2) relies only on the $\log_{10} K_{\mathrm{OA}}$ at 25 °C and additional parameters do not notably improve model performance. This model has similar or slightly improved performance relative to previous estimation techniques for $\Delta H^{\circ}_{\mathrm{OA}}$ and $\Delta U^{\circ}_{\mathrm{OA}}$, which relied on more or more complex descriptors. Our work parallels previous findings that that $\Delta H^{\circ}_{\mathrm{vap}}$ can be predicted quite well just from P_{L} [20]. Further measurements of $\log_{10} K_{\mathrm{OA}}$ and $\Delta U^{\circ}_{\mathrm{OA}}$ values could improve the applicability domain of these models particularly for chemicals with $\log_{10} K_{\mathrm{OA}}$ values between 4 and 6 ($\Delta U^{\circ}_{\mathrm{OA}}$ values between − 70 and − 50 kJ·mol^{-1}), and for more polar compounds with more complex hydrogen bonding abilities.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10953-022-01214-7.

Author Contributions *Conceptualization:* SB, FW; *Data Curation:* SB, AP; *Formal analysis and investigation:* SB, AP, AS; *Visualization:* SB; *Writing—original draft preparation:* SB, AP; *Writing—review and editing:* SB, AS, FW; *Funding acquisition:* FW; *Supervision:* FW.

Funding This work was funded by the European Chemical Industry Council (CEFIC) through project ECO 41 of the Long-range Research Initiative (LRI).

Data Availability All data generated or analysed during this study are included in this published article, its supplementary information files, and in previously published cited works.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical Approval This work was funded by the European Chemical Industry Council (CEFIC) through project ECO 41 of the Long-range Research Initiative (LRI).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Harner, T., Green, N.J.L., Jones, K.C.: Measurements of octanol−air partition coefficients for PCDD/Fs: a tool in assessing air−soil equilibrium status. Environ. Sci. Technol. **34**, 3109–3114 (2000)
2. Hippelein, M., McLachlan, M.S.: Soil/air partitioning of semivolatile organic compounds. 2. Influence of temperature and relative humidity. Environ. Sci. Technol. **34**, 3521–3526 (2000)
3. Hiatt, M.H.: Leaves as an indicator of exposure to airborne volatile organic compounds. Environ. Sci. Technol. **33**, 4126–4133 (1999)
4. Paterson, S., Mackay, D., Bacci, E., Calamari, D.: Correlation of the equilibrium and kinetics of leaf-air exchange of hydrophobic organic chemicals. Environ. Sci. Technol. **25**, 866–871 (1991)
5. Harner, T., Bidleman, T.F.: Octanol−air partition coefficient for describing particle/gas partitioning of aromatic compounds in urban air. Environ. Sci. Technol. **32**, 1494–1502 (1998)
6. Kaupp, H., McLachlan, M.S.: Gas/particle partitioning of PCDD/Fs, PCBs, PCNs and PAHs. Chemosphere **38**, 3411–3421 (1999)
7. Batterman, S., Zhang, L., Wang, S., Franzblau, A.: Partition coefficients for the trihalomethanes among blood, urine, water, milk and air. Sci. Total Environ. **284**, 237–247 (2002)

8. Davie-Martin, C.L., Hageman, K.J., Chin, Y.-P., Rougé, V., Fujita, Y.: Influence of temperature, relative humidity, and soil properties on the soil–air partitioning of semivolatile pesticides: laboratory measurements and predictive models. Environ. Sci. Technol. **49**, 10431–10439 (2015)
9. Shoeib, M., Harner, T.: Using measured octanol-air partition coefficients to explain environmental partitioning of organochlorine pesticides. Environ. Toxicol. Chem. **21**, 984–990 (2002)
10. Qiao, L.-N., Hu, P.-T., Macdonald, R., Kannan, K., Nikolaev, A., Li, Y.-F.: Modeling gas/particle partitioning of polybrominated diphenyl ethers (PBDEs) in the atmosphere: a review. Sci. Total Environ. **729**, 138962 (2020)
11. Kelly, B.C., Gobas, F.A.P.C.: An arctic terrestrial food-chain bioaccumulation model for persistent organic pollutants. Environ. Sci. Technol. **37**, 2966–2974 (2003)
12. Baskaran, S., Lei, Y.D., Wania, F.: A database of experimentally derived and estimated octanol–air partition ratios (K_{OA}). J. Phys. Chem. Ref. Data. **50**, 043101 (2021)
13. Goss, K.-U., Eisenreich, S.J.: Adsorption of VOCs from the gas phase to different minerals and a mineral mixture. Environ. Sci. Technol. **30**, 2135–2142 (1996)
14. Baskaran, S., Lei, Y.D., Wania, F.: Reliable prediction of the octanol–air partition ratio. Environ. Toxicol. Chem. **40**, 3166–3180 (2021)
15. Endo, S., Goss, K.-U.: Applications of polyparameter linear free energy relationships in environmental chemistry. Environ. Sci. Technol. **48**, 12477–12491 (2014)
16. Abraham, M.H., Acree, W.E.: Comparison of solubility of gases and vapours in wet and dry alcohols, especially octan-1-ol. J. Phys. Org. Chem. **21**, 823–832 (2008)
17. Endo, S., Goss, K.-U.: Predicting partition coefficients of polyfluorinated and organosilicon compounds using polyparameter linear free energy relationships (PP-LFERs). Environ. Sci. Technol. **48**, 2776–2784 (2014)
18. Jin, X., Fu, Z., Li, X., Chen, J.: Development of polyparameter linear free energy relationship models for octanol–air partition coefficients of diverse chemicals. Environ. Sci. Process. Impacts **19**, 300–306 (2017)
19. Mintz, C., Burton, K., Ladlie, T., Clark, M., Acree, W.E., Abraham, M.H.: Enthalpy of solvation correlations for gaseous solutes dissolved in dibutyl ether and ethyl acetate. Thermochim. Acta **470**, 67–76 (2008)
20. Goss, K.-U., Schwarzenbach, R.P.: Empirical prediction of heats of vaporization and heats of adsorption of organic compounds. Environ. Sci. Technol. **33**, 3390–3393 (1999)
21. MacLeod, M., Scheringer, M., Hungerbühler, K.: Estimating enthalpy of vaporization from vapor pressure using Trouton's rule. Environ. Sci. Technol. **41**, 2827–2832 (2007)
22. Abraham, M.H., Duce, P.P., Morris, J.J., Taylor, P.J.: Hydrogen bonding. Part 2.–Equilibrium constants and enthalpies of complexation for 72 monomeric hydrogen-bond acids with *N*-methylpyrrolidinone in 1,1,1-trichloroethane. J. Chem. Soc. Faraday Trans. I Phys. Chem. Condens. Phases **83**, 2867 (1987)
23. Mansouri, K., Grulke, C.M., Judson, R.S., Williams, A.J.: OPERA models for predicting physicochemical properties and environmental fate endpoints. J. Cheminform. **10**, 1–19 (2018)
24. Yap, C.W.: PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. **32**, 1466–1474 (2011)
25. Mansouri, K., Williams, A.: QMRF - Title: K_{OA} model for the octanol/air partition coefficient prediction from OPERA models (2017)
26. Ulrich, N., Endo, S., Brown, T., Bronner, G., Abraham, M.H., Goss, K.-U.: UFZ-LSER database v 3.2. http://www.ufz.de/lserd (2017)
27. Brown, T.N.: QSPRs for predicting equilibrium partitioning in solvent–air systems from the chemical structures of solutes and solvents. J. Solution. Chem. **51**, 1101–1132 (2022)
28. ARC Arnot Research and Consulting Inc.: Exposure and Safety Estimation (EAS-E) Suite. ARC Arnot Research and Consulting Inc, Toronto (2022)
29. Brown, T.N.: IFSQSAR - A python package for applying QSARs. ver. 1.0. https://github.com/tnbrowncontam/ifsqsar (2022)
30. Gramatica, P., Chirico, N., Papa, E., Cassani, S., Kovarich, S.: QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. J. Comput. Chem. **34**, 2121–2132 (2013)
31. Mintz, C., Clark, M., Acree, W.E., Abraham, M.H.: Enthalpy of solvation correlations for gaseous solutes dissolved in water and in 1-octanol based on the Abraham model. J. Chem. Inf. Model. **47**, 115–121 (2007)
32. Berti, P., Cabani, S., Conti, G., Mollica, V.: Thermodynamic study of organic compounds in octan-1-ol. Processes of transfer from gas and from dilute aqueous solution. J. Chem. Soc. Faraday Trans. I Phys. Chem. Condens. Phases **82**, 2547–2556 (1986)
33. Cheong, W.J.: Measurements of limiting activity coefficients of homologous series of solutes and their application to the study of retention mechanism in reversed phase liquid chromatography. University of Minnesota, PhD Thesis (1989)

34. Eger, E.I.I., Ionescu, P., Laster, M.J., Gong, D., Hudlicky, T., Kendig, J.J., Harris, R.A., Trudell, J.R., Pohorille, A.: Minimum alveolar anesthetic concentration of fluorinated alkanols in rats: relevance to theories of narcosis. Anesth. Analg. **88**, 867–876 (1999)
35. Fang, Z., Ionescu, P., Chortkoff, B.S., Kandel, L., Sonner, J., Laster, M.J., Eger, E.I.I.: Anesthetic potencies of n-alkanols: results of additivity and solubility studies suggest a mechanism of action similar to that for conventional inhaled anesthetics. Anesth. Analg. **84**, 1042–1048 (1997)
36. Fang, Z., Sonner, J., Laster, M.J., Ionescu, P., Kandel, L., Koblin, D.D., Eger, E.I.I., Halsey, M.J.: Anesthetic and convulsant properties of aromatic compounds and cycloalkanes: implications for mechanisms of narcosis. Anesth. Analg. **83**, 1097–1104 (1996)
37. Hiatt, M.H.: Analyses of fish tissue by vacuum distillation/gas chromatography/mass spectrometry. Anal. Chem. **69**, 1127–1134 (1997)
38. Rohrschneider, L.: Solvent characterization by gas-liquid partition coefficients of selected solutes. Anal. Chem. **45**, 1241–1247 (1973)
39. Taheri, S., Laster, M.J., Liu, J., Eger, E.I.I., Halsey, M.J., Koblin, D.D.: Anesthesia by *n*-alkanes not consistent with the Meyer–Overton hypothesis: determinations of the solubilities of alkanes in saline and various lipids. Anesth. Analg. **77**, 7–11 (1993)
40. Treves, K., Shragina, L., Rudich, Y.: Measurement of octanol–air partition coefficients using solid-phase microextraction (SPME)–application to hydroxy alkyl nitrates. Atmos. Environ. **35**, 5843–5854 (2001)
41. Lei, Y.D., Baskaran, S., Wania, F.: Measuring the octan-1-ol air partition coefficient of volatile organic chemicals with the variable phase ratio headspace technique. J. Chem. Eng. Data **64**, 4793–4800 (2019)
42. Gruber, D., Langenheim, D., Gmehling, J., Moollan, W.: Measurement of activity coefficients at infinite dilution using gas−liquid chromatography. 6. Results for systems exhibiting gas−liquid interface adsorption with 1-octanol. J. Chem. Eng. Data **42**, 882–885 (1997)
43. Abraham, M.H., Le, J., Acree, W.E., Carr, P.W., Dallas, A.J.: The solubility of gases and vapours in dry octan-1-ol at 298 K. Chemosphere **44**, 855–863 (2001)
44. Gramatica, P.: Principles of QSAR modeling: comments and suggestions from personal experience. Int. J. Quant. Struct.-Prop. Relatsh. **5**, 61–97 (2020)
45. Chirico, N., Gramatica, P.: Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. J. Chem. Inf. Model. **52**, 2044–2058 (2012)
46. Chirico, N., Gramatica, P.: Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J. Chem. Inf. Model. **51**, 2320–2335 (2011)
47. Gramatica, P., Sangion, A.: A historical excursus on the statistical validation parameters for QSAR models: a clarification concerning metrics and terminology. J. Chem. Inf. Model. **56**, 1127–1131 (2016)
48. Lin, L.I.-K.: A concordance correlation coefficient to evaluate reproducibility. Biometrics **45**, 255–268 (1989)
49. Baskaran, S., Lei, Y.D., Wania, F.: Response to comment on "A database of experimentally derived and estimated octanol–air partition ratios (K_{OA})". J. Phys. Chem. Ref. Data **51**, 026102 (2022)
50. Goss, K.-U., Bronner, G., Harner, T., Hertel, M., Schmidt, T.C.: The partition behavior of fluorotelomer alcohols and olefins. Environ. Sci. Technol. **40**, 3572–3577 (2006)
51. Thuens, S., Dreyer, A., Sturm, R., Temme, C., Ebinghaus, R.: Determination of the octanol−air partition coefficients (K_{OA}) of fluorotelomer alcohols. J. Chem. Eng. Data **53**, 223–227 (2008)
52. Dreyer, A., Langer, V., Ebinghaus, R.: Determination of octanol−air partition coefficients (K_{OA}) of fluorotelomer acrylates, perfluoroalkyl sulfonamids, and perfluoroalkylsulfonamido ethanols. J. Chem. Eng. Data **54**, 3022–3025 (2009)
53. Shoeib, M., Harner, T., Ikonomou, M., Kannan, K.: Indoor and outdoor air concentrations and phase partitioning of perfluoroalkyl sulfonamides and polybrominated diphenyl ethers. Environ. Sci. Technol. **38**, 1313–1320 (2004)
54. Golmohammadi, H., Dashtbozorgi, Z., Jr.: Application of QSPR for the prediction of gas to 1-octanol solvation enthalpy using support vector regression. Phys. Chem. Liq. **51**, 182–202 (2013)
55. Xiao, H., Wania, F.: Is vapor pressure or the octanol–air partition coefficient a better descriptor of the partitioning between gas phase and organic matter? Atmos. Environ. **37**, 2867–2878 (2003)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.