Applicability Domain of Polyparameter Linear Free Energy Relationship Models Evaluated by Leverage and Prediction Interval Calculation

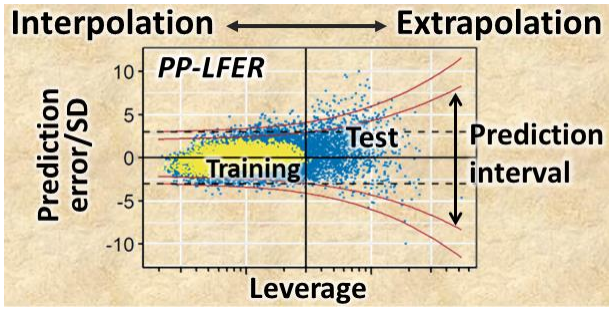
Satoshi Endo[1,2]

[1] Health and Environmental Risk Division, National Institute for Environmental Studies (NIES), Onogawa 16-2, 305-8506 Tsukuba, Ibaraki, Japan

[2] Graduate School of Engineering, Osaka City University, Sugimoto 3-3-138, Sumiyoshi, 558-8585 Osaka, Japan

Contact information:
Satoshi Endo
Health and Environmental Risk Division
National Institute for Environmental Studies (NIES)
Onogawa 16-2, 305-8506 Tsukuba, Ibaraki
Japan
Phone: ++81-29-850-2695
Fax: ++81-29-850-2870
Email: endo.satoshi@nies.go.jp

TOC graphic


Interpolation
Extrapolation
PP-LFER
Prediction error/SD
10
5
0
-5
-10
Test
Training
Prediction interval
Leverage

Abstract

Polyparameter linear free energy relationships (PP-LFERs) are accurate and robust models employed to predict equilibrium partition coefficients (*K*) of organic chemicals. The accuracy of predictions by a PP-LFER depends on the composition of the respective calibration data set. Generally, extrapolation outside the model calibration domain is likely to be less accurate than interpolation. In this study, the applicability domain (AD) of PP-LFERs was systematically evaluated by calculating the leverage (*h*) and prediction interval (PI). Repeated simulations with experimental data showed that the root mean squared error of predictions increased with *h*. However, the analysis also showed that PP-LFERs calibrated with a large number (e.g., 100) of training data were highly robust against extrapolation error. For such well-calibrated PP-LFERs, the common definition of extrapolation ($h > 3\ h_{mean}$, where h_{mean} is the mean *h* of all training compounds) may be excessively strict. Alternatively, the PI is proposed as a metric to define the AD of PP-LFERs, as it provides a concrete estimate of the error range that agrees well with the observed errors, even for extreme extrapolations. Additionally, published PP-LFERs were evaluated in terms of their AD using the new concept of AD probes, which indicated the varying predictive performance of PP-LFERs in existing literature for environmentally relevant compounds.

Keywords

Applicability domain, linear solvation energy relationship, extrapolation, property prediction, partition coefficient, QSAR, QSPR, perfluoroalkyl substances

Synopsis

Calculating the prediction intervals delineates the applicability domain of polyparameter linear free energy relationship models.

53 # 1. Introduction

54 Equilibrium partition coefficients largely determine the environmental distribution of organic
55 contaminants and are crucial parameters for environmental risk assessments. Among various
56 models, the linear solvation energy relationships (LSERs),[1] or generally, polyparameter linear
57 free energy relationships (PP-LFERs) that use Abraham's solute descriptors have been
58 confirmed to be accurate and robust for predicting partition coefficients.[2] The PP-LFERs cover
59 the intermolecular interactions relevant to the phase partitioning of neutral organic
60 compounds. Their successful environmental applications have been previously reviewed.[3,4]

61 PP-LFERs are multiple linear regression models that typically use five solute
62 descriptors. The following three types of equations are most often applied.[1,5]

63 $$\text{Log } K = c + eE + sS + aA + bB + vV \quad (1)$$

64 $$\text{Log } K = c + eE + sS + aA + bB + lL \quad (2)$$

65 $$\text{Log } K = c + sS + aA + bB + vV + lL \quad (3)$$

66 The symbols denote the following: *K*, partition coefficient; *E*, excess molar refraction; *S*, solute
67 polarizability/dipolarity parameter; *A*, solute hydrogen (H)-bond donor property; *B*, solute H-
68 bond acceptor property; *V*, McGowan's molar volume; and *L*, logarithmic hexadecane/air
69 partition coefficient. The lowercase letters are regression coefficients and are typically trained
70 with several tens of compounds for which experimental log *K* and the solute descriptors (i.e.,
71 *E*, *S*, *A*, *B*, *V*, and *L*) are available. The fitting of the PP-LFERs is high even to data that are highly
72 diverse in size and polarity. For solvent/water and solvent/air partition coefficients, the
73 calibration typically results in a standard deviation (SD) of 0.2 or below for the log *K* values.[1]
74 Partition systems that involve a heterogeneous phase (e.g., natural organic matter) can
75 exhibit a lower quality of fit (SD, 0.3–0.5 log units).[3]

76 PP-LFERs are derived from a multiple linear regression; therefore, their applicability
77 domain (AD) is related to the training (calibration) set of compounds. Generally, extrapolation
78 (i.e., prediction beyond the calibrated domain) is likely to be less accurate than interpolation.
79 Moreover, a long-range extrapolation is expected to be more error-prone than a short-range
80 extrapolation. However, in a multidimensional space (here, 5 descriptors), it is unclear how
81 the terms interpolation and extrapolation can be defined and how a quantitative relationship
82 between the extent of extrapolation and prediction accuracy may be established. Notably, an
83 extrapolation can be less accurate but is not necessarily inaccurate or unreliable. The required

accuracy depends on the purpose of the model use, and extrapolation can be acceptable within the range where its accuracy is satisfactory.

Among various approaches, calculation of the leverages has been considered to define and evaluate the AD for linear regression models.[6-9] The leverage is a quantitative measure of the distance from the entire set of calibration data. Leverage calculation is applied to identify outliers within the calibration set, and it can also be used to quantitatively define extrapolation in the prediction. A large leverage value indicates a long distance from the calibrated domain and thus an extrapolation with the possibility of increased error.

The prediction interval (PI) is the range of values where future model predictions are expected to fall at a given frequency. Typically, 95 or 99% PIs are calculated. Although PIs are frequently calculated for predictions by a simple linear regression model, they are not commonly presented for multiple linear regression models, including PP-LFERs. However, the PI can be more useful than the leverage, as the PI considers both the distance from the calibration set and the quality of the model fitting (see Section 2.2 for more details).

The purposes of this study are three-fold: (i) To quantitatively demonstrate how the prediction accuracy of a PP-LFER decreases when moving away from a specific domain of calibration defined by the leverage, (ii) to compare actual prediction errors with error margins expected by PIs, and (iii) to evaluate several calibration sets for PP-LFERs in terms of their AD using a new concept of AD probes. On the basis of these, a discussion is presented on the definition and evaluation of AD for PP-LFER models. The information should also be helpful for the future development of PP-LFERs because it ensures an optimized calibration data set.

2. Methodology

2.1 Definition and calculation of the leverage and PI

The definition and calculation of the leverage and PI are described in full in SI-1 of the Supporting Information (SI) and only briefly here.

The PP-LFER regression can be expressed in matrix form as follows,

$$y = X\beta + \varepsilon \tag{4}$$

where y is the vector of observations for log K, β is the vector of regression coefficients, and ε is the error vector. X is the design matrix containing solute descriptors of n training compounds. The hat matrix (H) can be derived from X, and the diagonals of H (i.e., h_{ii}) are

referred to as the leverages and infer the distance of each calibration compound from the others in terms of the solute descriptor combination. h_{ii} is between 0 and 1, and the sum of h_{ii} for the n training compounds is equal to the number of fitting parameters p, which is 6 for the PP-LFERs (including the regression constant). An overly high h_{ii} indicates that the respective calibration compound is an outlier in terms of its descriptors. Typically, $h_{ii} = 3h_{mean}$ is considered a threshold value,[6-9] where h_{mean} is the mean of h_{ii} for all calibration compounds and is equal to p/n. To evaluate the extrapolation for compound j, which is not included in the calibration set, h is calculated as,

$$h = x_j^T (X^T X)^{-1} x_j \quad (5)$$

where x_j is the column vector containing the solute descriptors of j. Analogous to the identification of outliers in the training set, $h = 3h_{mean}$ is typically considered the threshold value for extrapolation.[6-9]

The PI of the PP-LFER can be expressed as [log K_j – Δ(log K), log K_j + Δ(log K)], where log K_j is the value for compound j predicted with eq 4 (i.e., log $K_j = x_j^T \beta$) and Δ(log K) is half the width of the PI. Δ(log K) is calculated as,

$$\Delta(\log K) = t_{\alpha/2,n-k-1}\mathrm{SD}_{\mathrm{training}}\sqrt{1 + x_j^T (X^T X)^{-1} x_j} \quad (6)$$

$$= t_{\alpha/2,n-k-1}\mathrm{SD}_{\mathrm{training}}\sqrt{1 + h} \quad (7)$$

where $t_{\alpha/2,n-k-1}$ is the two-tailed t-value for a given confidence level (α, e.g., 95%), number of training data (n), and number of independent variables (k; 5 for PP-LFERs). $\mathrm{SD}_{training}$ is the standard deviation of the PP-LFER model fitted to the training data. Δ(log K) may be normalized to $\mathrm{SD}_{training}$, as

$$\Delta(\log K)/\mathrm{SD}_{\mathrm{training}} = t_{\alpha/2,n-k-1}\sqrt{1 + h} \quad (8)$$

In this study, the following two tests were performed to discuss the use of h and the PIs to delineate the AD of PP-LFERs.

2.2 Test 1: Comparison of prediction errors with *h* and the PIs

In the first test, the variation of actual prediction errors by PP-LFERs with h and the PIs was examined. Six experimental data sets of partition coefficients from existing literature were used: octanol/water (K_{ow}, n = 314);[10] air/water (K_{aw}, n = 390);[11] oil/water (K_{oilw}, n = 247);[12] soil organic carbon/water (K_{oc}, n = 79);[13] phospholipid liposome/water (K_{lipw}, n = 131);[14] and bovine serum albumin/water (K_{BSAw}, n = 82).[15] These data sets comprise a relatively large

145 number of compounds and exhibit environmental and toxicological relevance. K_{ow}, K_{aw}, and
146 K_{oilw} were partition coefficients between two homogeneous solvents, whereas K_{oc}, K_{lipw}, and
147 K_{BSAw} involved a heterogeneous or anisotropic phase. The *K* values and solute descriptors
148 were obtained from the aforecited references, are listed in Tables S1–S6, and are summarized
149 in Table S7 (SI-2 of the SI)

150 To evaluate prediction accuracy, the *K* data of each set were divided into training and
151 test sets. Training compounds were randomly selected from the entire data set. The number
152 of the training compounds ($n_{training}$) was 20, 30, 40, 50, 75 or 100. Rather small values of $n_{training}$
153 were also included in this test to simulate cases of insufficient calibration. The compounds
154 that were not selected as training compounds were used as test compounds. The PP-LFER in
155 the form of eq 1 was calibrated with the training data and was used to predict log *K* for the
156 test compounds. Prediction errors (predicted log *K* − experimental log *K*) were calculated and
157 compared with *h* and Δ(log *K*). For each combination of the *K* set and $n_{training}$, the cycle of
158 "random generation of a training set," "calibration of the PP-LFER," and "prediction for the
159 test set" was repeated 200 times. This number was arbitrary but appeared sufficient for stable
160 results.

161 Additionally, using the 200 calibrated PP-LFERs for each case, the experimental log *K*
162 values of per- and polyfluoroalkyl substances (PFASs) and organosilicon compounds (OSCs)
163 were predicted. PFASs and OSCs possess extremely weak van der Waals interaction
164 properties; thus, the *E* and *L* values are comparatively low for their molecular sizes.[16]
165 Therefore, PP-LFERs often have to be extrapolated to predict *K* values. These classes of
166 compounds are not present in the data set of any considered PP-LFER and are used to
167 evaluate the influences of extrapolation on the prediction accuracy.

168 All calculations mentioned above were performed with *R* software.

169 **2.3 Test 2: Evaluating reported PP-LFERs with AD probes**

170 In the second test, *h* and PI calculation was applied to evaluate the AD of reported PP-LFER
171 equations. Here, *n*, $SD_{training}$, and the solute descriptors of the calibration compounds were
172 extracted from existing literature and used to calculate *h* and PIs for 25 selected compounds
173 (Table S8, SI-3). These compounds, referred to as AD probes herein, were selected because
174 of their wide variations in descriptor values, structural diversity, and environmental relevance.
175 They represented aliphatic and aromatic, polar and nonpolar, and small and large compounds

176 and included multifunctional polar compounds such as various pesticides and
177 pharmaceuticals, a neutral PFAS, and an OSC. Solute descriptors for the AD probes were
178 obtained from the UFZ-LSER database and listed in Table S8 (SI-3).[17] Test 2 did not require the
179 experimental K values of the AD probes, and only solute descriptors were used for the
180 calculation. As the SI, an Excel file with a macro is provided that calculates h, h/h_{mean}, and
181 $\Delta(\log K)$ for the AD probes and any desired chemical based on the user-entered training data.
182 Note that there exist compounds with extreme descriptor values that are not covered by the
183 25 AD probes proposed here. For example, an antibiotic erythromycin (E = 2.90, S = 3.73, A =
184 1.25, B = 4.96, V = 5.773)[18] exhibits exceptionally high S, B and V values. However, such
185 compounds are rarely used for calibration and are always out of the calibration domain;
186 therefore, they are not necessary specifically in this evaluation.
187

188 # 3. Results and discussion

189 ## 3.1 Prediction errors compared to *h* and the PIs (Test 1)

190 Figure S1 (SI-4) shows the root mean squared errors (RMSEs) for training and testing sets
191 randomly generated 200 times. The test compounds were grouped into several bins according
192 to the h normalized to h_{mean} (h/h_{mean}) before the RMSEs were calculated. The observed RMSE
193 for the test compounds increased with h for a given K data set and $n_{training}$. The increasing
194 trend of RMSE with h was particularly clear for simulations with small $n_{training}$ values (i.e., 20,
195 30). The trend was sometimes unclear for simulations with high $n_{training}$ values, likely because
196 large $n_{training}$ resulted in a relatively small n_{test}, which may not be able to provide
197 representative RMSEs, particularly for high h/h_{mean} bins.

198 To demonstrate the increase in RMSE with h/h_{mean} more clearly, the RMSE values for
199 the test data relative to the RMSE for the training data were calculated (Figure 1, Figure S2 in
200 SI-4). The relative RMSE generally increased with h/h_{mean} but to a lesser extent when $n_{training}$
201 was large. For example, the relative RMSEs of log K_{ow} data in the "$2 < h/h_{mean} < 3$" bin were
202 1.75, 1.52, 1.42, and 1.34 for $n_{training}$ = 20, 40, 75, and 100, respectively. This result suggests
203 that if the PP-LFER is trained with a sufficient size of data, the RMSEs for interpolations (i.e.,
204 $h/h_{mean} < 3$) will resemble the RMSE for the training set. Noteworthily, even for the "$3 <$
205 $h/h_{mean} < 4$" bin (i.e., extrapolation), the relative RMSE for any K considered was < 1.5 when
206 $n_{training} \geq 50$, and < 2.2 when $n_{training} \geq 20$. These RMSEs can be sufficiently accurate for various

purposes. Although $h = 3h_{mean}$ is the common definition of extrapolation, the actual threshold of h may be adapted to the required accuracy of predictions, depending on the quality of the PP-LFER fit and $n_{training}$. For example, if the required accuracy is 0.3 log units, which is typically the level of accuracy of contaminant fate models,[19] then extrapolations by the PP-LFERs for log K_{ow} and log K_{aw} up to an h/h_{mean} of 4 can be allowed, according to the results of Test 1 (Figure S1). In contrast, a stricter threshold, e.g., $h/h_{mean} < 2$ or even < 1, should be set to log K_{oc}, log K_{lipw}, and log K_{BSAw} to comply with the criterion of 0.3 log unit RMSE. Alternative AD thresholds are further discussed in Section 3.3.

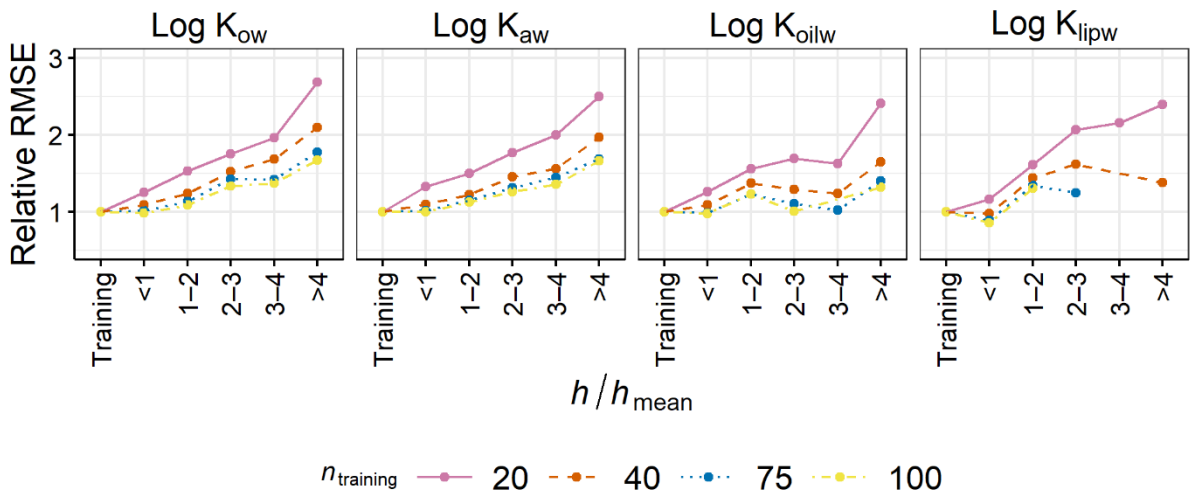


Figure 1. RMSEs of the test data, sorted according to h/h_{mean}, relative to the RMSE of the training data. The plots for $n_{training}$ = 30 and 50 and log K_{oc} and log K_{BSAw} are available in the Figure S2 (SI-4).

Along with average errors, such as RMSEs, the risk of an extremely inaccurate prediction is of interest. Individual data of Test 1 for log K_{ow} and log K_{lipw} were plotted against h (Figure 2). All other data are shown in Figure S3 (SI-5). When $n_{training}$ was small (e.g., 20, 30), both h (x-axis) and prediction errors (y-axis, normalized to $SD_{training}$) for the test data were widely distributed. Extremely large errors ($|error/SD_{training}| > 5$) occasionally occurred, particularly if h was large ($> 10h_{mean}$). In contrast, when $n_{training}$ was large (e.g., 75, 100), the training and test data were similarly distributed in terms of h and the prediction errors.

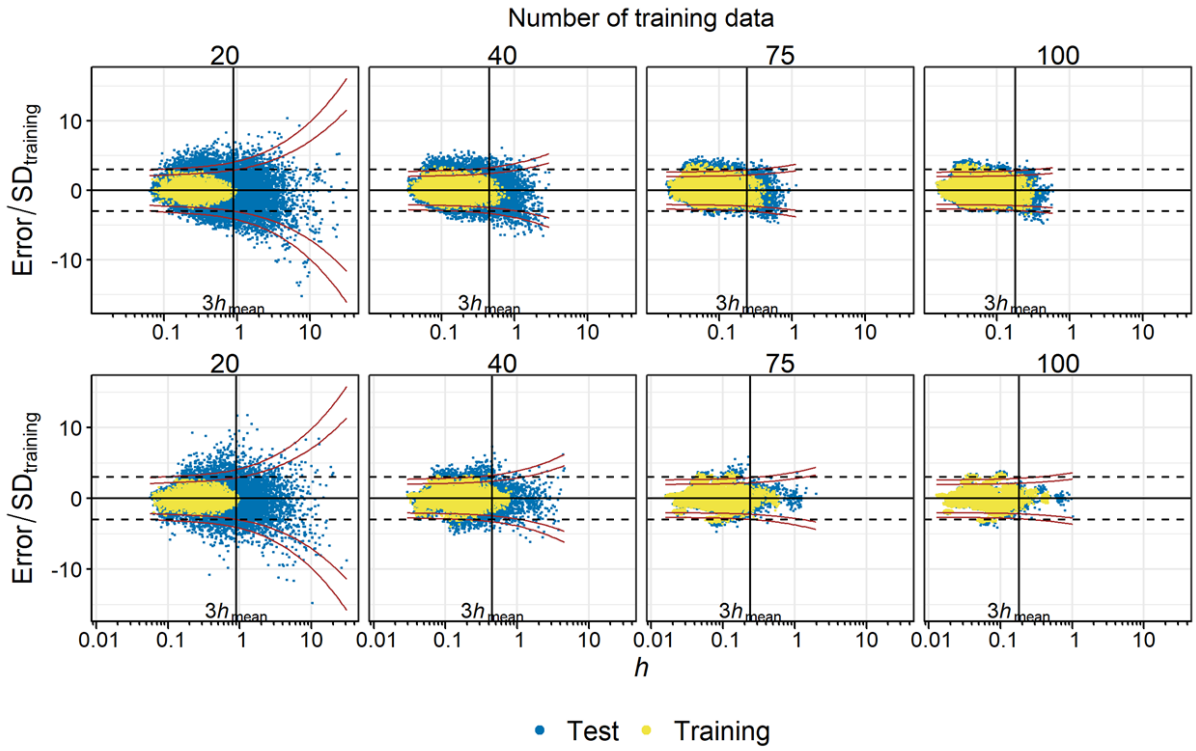


Figure 2. Prediction errors normalized to $SD_{training}$ plotted against *h*. Results from 200 simulations are shown. The vertical line indicates $3h_{mean}$. The dashed horizontal lines indicate errors that are 3 times the $SD_{training}$. The curves indicate the 95% (inside) and 99% (outside) prediction intervals. Top, log K_{ow}; bottom, log K_{lipw}. All other data are shown in Figure S3 (SI-5).

The percentage of large prediction errors, defined by $|error/SD_{training}| > 3$, was generally higher for extrapolation ($h/h_{mean} > 3$) than interpolation ($h/h_{mean} < 3$) (Figure S4, SI-6). However, the percentage strongly decreased with $n_{training}$. As an example: for log K_{ow}, when $n_{training}$ = 20, 3.3% of the interpolations and 17% of the extrapolations suffered from large prediction errors. In contrast, when $n_{training}$ = 100, 0.94% of the interpolations and 4.7% of the extrapolations resulted in large prediction errors, which conversely indicated that 94% of the extrapolations ended up with errors within 3 $SD_{training}$.

Figures 2 additionally shows the 95% and 99% PIs as a function of *h*. The PIs were narrow up to $h \sim 1$ and diverged with *h*, as expected from eq 8. The extent of divergence was large when $n_{training}$ was small, which can be explained by a large $t_{\alpha/2,n\text{-}k\text{-}1}$ in eq 8. The data points from Test 1 were within the PIs with a few outliers. The percentage of the test data within a

248 given PI agrees with the theoretical expectations; e.g., ca 95% of the test data are within the
249 95% PI, independent of $n_{training}$ (Figure S5, SI-7).
250 Overall, Test 1 demonstrated that the mean prediction error increased with h and
251 could be used to identify "risky predictions" that frequently cause high inaccuracy. However,
252 a threshold of $3h_{mean}$ did not appear to be versatile in defining the AD, as the $n_{training}$ appeared
253 to influence the range of prediction errors. The plots in Figures 1, 2, and S1–S5 suggested that,
254 when $n_{training}$ was large, $h = 3h_{mean}$ might be overly strict as a threshold, because prediction
255 errors were often similar in magnitude even when $h > 3h_{mean}$. Note that Test 1 was also
256 performed with eq 3, the PP-LFER equation that uses L instead of E. However, the results were
257 similar to those of eq 1 and are thus not discussed herein.
258
259 **3.2 PFASs and OSCs**
260 Using 200 trained PP-LFERs, log K_{ow} of 3 PFASs (4:2 fluorotelomer alcohol (FTOH), 6:2 FTOH,
261 and 8:2 FTOH) and 3 OSCs (octamethylcyclotetrasiloxane (D4), decamethylcyclopentasiloxane
262 (D5), and dodecamethylcyclohexasiloxane (D6)) were predicted and compared to the
263 experimental data (Figure 3; additional data in Figure S6, SI-8.[16] For this comparison, eq 3
264 instead of eq 1 was used because the latter is known to be unsuitable for PFASs and OSCs (ref
265 16; also compare Figures S6 and S7 in SI-8 and SI-9, respectively). The h/h_{mean} ratios for these
266 six chemicals were always above 3 with any $n_{training}$ used and were up to 300, indicating strong
267 extrapolations. The predictions were highly inaccurate when the $n_{training}$ was small. However,
268 the predictions appeared to improve with an increase in $n_{training}$. When $n_{training} = 100$, even
269 largely extrapolated FTOHs ($h \sim 2$, $h/h_{mean} \sim 33$) were frequently predicted within 3 $SD_{training}$.
270 The dependence of the prediction error on h was well captured by the PIs; the majority of the
271 data were within the 99% PIs, and this was the case for extreme extrapolations as well (Figures
272 3, S6). The results for PFASs and OSCs can be considered another indication that well-
273 calibrated PP-LFERs are robust against extrapolation and that $h = 3h_{mean}$ as the cutoff criterion
274 is excessively strict if the $n_{training}$ is large. Notably, although well-calibrated PP-LFERs appear
275 to bear extrapolation, the inclusion of PFASs and OSCs in the calibration set is the first choice
276 to develop PP-LFERs that work for these classes of chemicals, as that substantially decreases
277 h for PFASs and OSCs.[16]

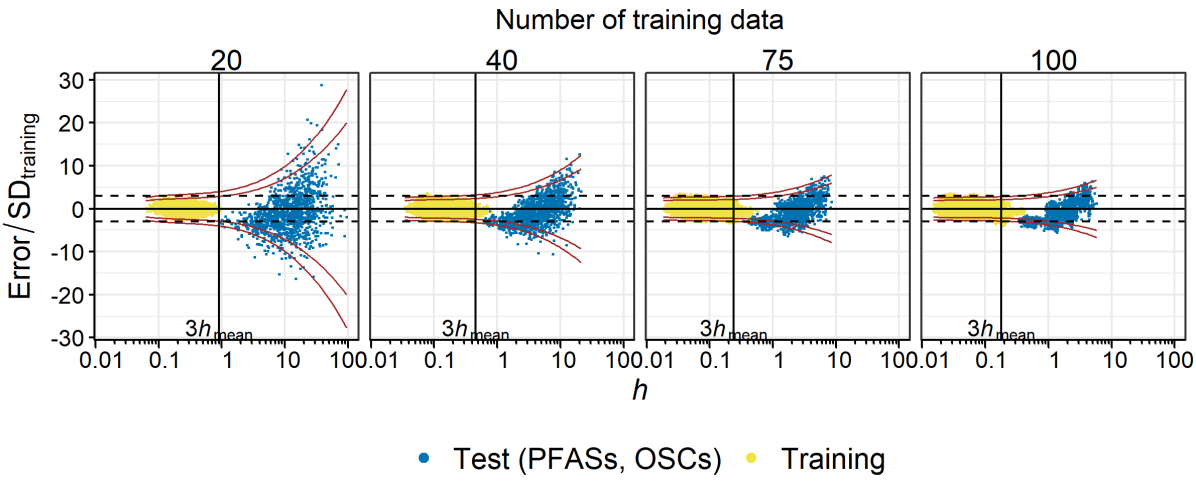


Figure 3. Prediction errors for log K_{ow} of PFAS and OSCs normalized to $SD_{training}$ plotted against *h*. The results from 200 simulations are shown. The lines indicate the same as in Figure 2. Equation 3 was used for this plot (see text for more details). Additional data are in Figure S6 (SI-8).

3.3 How can we define the AD of PP-LFERs?

In previous discussions regarding the AD of quantitative structure activity relationships (QSARs), the use of *h* with a cutoff value of $3h_{mean}$ has been frequently presented. As shown in Test 1 of this study, however, this cutoff may excessively limit the potential of well-calibrated PP-LFERs to predict a broad range of compounds above the $3h_{mean}$ threshold. The use of the PI, in contrast, has rarely been investigated in the context of QSAR development but may be more practical for multiple linear regression models, such as PP-LFERs, because the PI encompasses the distance (*h*), quality of model fit ($SD_{training}$), and size of training data (influencing *h* and $t_{\alpha/2,n-k-1}$) and provides a concrete estimate of the error range (eq 7). To use the PI to define the AD, an upper threshold for Δ(log *K*) must be set. Here, two ways that may be acceptable are discussed.

(A) Set the Δ(log K) threshold at a multiple of $SD_{training}$. The AD may be defined by a Δ(log *K*) threshold that is a multiple of $SD_{training}$. An example of such a criterion is $\Delta(\log K)_{99\%PI} < 3SD_{training}$. According to eq 8, this condition corresponds to,

$$t_{99/2,n-k-1}\sqrt{1+h} < 3 \qquad (9)$$

Inequality 9 describes the two intersections in Figures 2 and 3 where the curves for the 99% PI meet the horizontal lines for $\pm 3SD_{training}$. By solving this inequality for *h*, we obtain,

$$h < \left(\frac{3}{t_{99/2,n-k-1}}\right)^2 - 1 \quad (10)$$

Inequality 10 describes a new h threshold that is derived from "$\Delta(\log K)_{99\%PI} < 3SD_{training}$" and is a function of $t_{\alpha/2,n-k-1}$. As $t_{\alpha/2,n-k-1}$ is dependent on $n_{training}$, this h threshold is also dependent on $n_{training}$ (Figure 4). For example, if $n_{training} = 50$, the new threshold is $h < 0.24$, which is $h/h_{mean} < 2.0$. If $n_{training} = 100$, the threshold is $h < 0.30$, which is $h/h_{mean} < 5.0$. The common threshold of $h/h_{mean} < 3$ can be derived when $n_{training} = 66.6$. Thus, the new threshold is stricter if $n_{training} \leq 66$ and less strict if $n_{training} \geq 67$, compared with the $3h_{mean}$ rule.

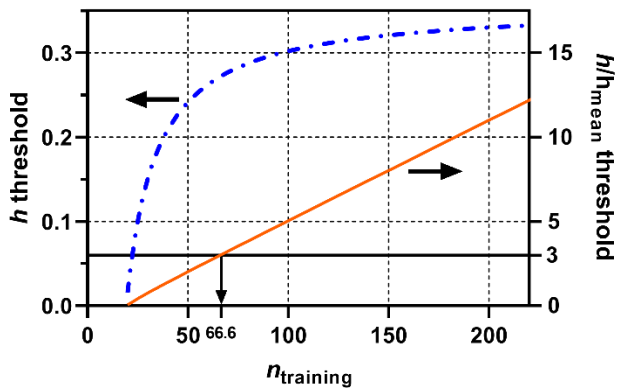


Figure 4. New thresholds of h and h/h_{mean} derived from $\Delta(\log K)_{99\%PI} < 3SD_{training}$ as a criterion (eq 10).

(B) Set the Δ(log K) threshold at a certain value. In the second approach, the AD is defined in such a way that the PI becomes narrower than a certain range. For example, we may consider $\Delta(\log K)_{99\%PI} < 0.5$ (i.e., a factor of 3 for K) as an acceptable error margin, then eq 7 becomes,

$$t_{99/2,n-k-1}\mathrm{SD}_{\mathrm{training}}\sqrt{1+h} < 0.5 \quad (11)$$

which can be rewritten as,

$$h < \left(\frac{0.5}{t_{99/2,n-k-1}\mathrm{SD}_{\mathrm{training}}}\right)^2 - 1 \quad (12)$$

Using the $SD_{training}$ value for the PP-LFER of log K_{ow} (Table S7, SI-1) as an example, we can derive a threshold of h specific to log K_{ow}. By inserting $SD_{training} = 0.154$ and $t_{99/2,n-k-1} = 2.59$ (with $n = 314$) in inequality 12, we obtained $h < 0.57$ (i.e., $h/h_{mean} < 30$). Note that if $SD_{training}$ is high (e.g., 0.285 for log K_{lipw}), "$\Delta(\log K)_{99\%PI} < 0.5$" is not achievable no matter how large $n_{training}$ is, because $t_{99/2,n-k-1}$ is > 2.58 regardless of $n_{training}$ and the righthand side of inequality

12 is always negative. The difficulty associated with this approach to define the AD may be to set the acceptable $\Delta(\log K)_{99\%PI}$ level such that it is both useful and achievable.

3.4 Evaluating AD of published PP-LFERs with AD probes (Test 2)

Using the 25 AD probes, 10 published PP-LFER equations[10-15,20-23] including those used in Test 1 were evaluated (Figure 5, Figure S8 in SI-10).

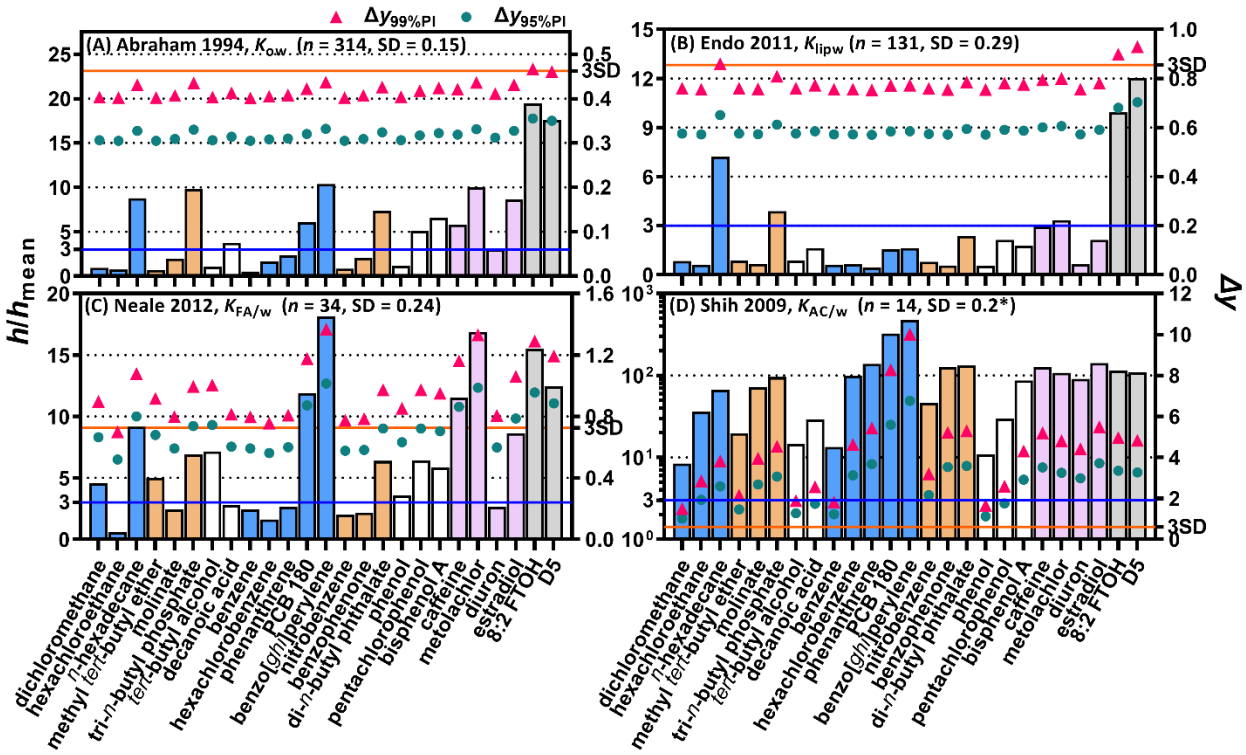


Figure 5. Leverage (bars) and prediction intervals (triangles and circles) of 25 applicability domain (AD) probes calculated with the training data sets of four PP-LFERs. Solid horizontal lines indicate h/h_{mean} = 3 and $\Delta(\log K)$ = 3SD. *The cited reference does not give SD but a "mean error" of 0.2, which was used here. Plots for all 10 PP-LFERs are shown in Figure S8, SI-10.

The h calculation showed that none of the 10 training sets considered encompassed all the 25 AD probes within the $3h_{mean}$ domain. This indicates that certain environmentally relevant compounds must be extrapolated with these PP-LFERs. Particularly, 8:2 FTOH and D5 always appeared as highly extrapolated chemicals (h/h_{mean} = 8–50), reflecting the fact that PFASs and OSCs were not included in any of the training sets and indicating that these

342 compounds were not well represented by other training compounds. For each type of
343 chemical, the small compounds (e.g., dichloromethane, methyl *tert*-butyl ether, benzene)
344 exhibited lower h/h_{mean} ratios than the large compounds (e.g., hexadecane, tri-*n*-butyl
345 phosphate, benzo[*ghi*]perylene). Generally, relatively small compounds are easy to measure,
346 and their data are present in the training set, whereas obtaining data for large compounds
347 tends to be more challenging. Consequently, PP-LFERs must be frequently extrapolated for
348 large compounds.

349 The data sets for log K_{ow}[10] and log K_{aw}[11] exhibited similar patterns for h/h_{mean} and Δ(log
350 *K*). Thus, the h/h_{mean} ratios of the small compounds were < 3 (interpolation) and those of the
351 large compounds were in the range of 3–15 (extrapolation) (Figure 5A). However, the Δ(log
352 *K*) values were not largely different across the 25 AD probes. Although 12 out of 25 AD probes
353 exhibited h/h_{mean} > 3, $\Delta(\log K)_{95\%PI}$ and $\Delta(\log K)_{99\%PI}$ were ~ 0.3 and ~ 0.4, respectively, for all
354 the AD probes. Even for strongly extrapolated 8:2 FTOH, $\Delta(\log K)_{95\%PI}$ and $\Delta(\log K)_{99\%PI}$ of log
355 K_{ow} predictions were 0.36 and 0.47, respectively. These relatively low Δ(log *K*) values for the
356 extrapolated compounds originated from the substantial size of training data for K_{ow} and K_{aw}.
357 The log K_{oilw}[12] data set resulted in similar patterns for h/h_{mean} and Δ(log *K*), but the values of
358 Δ(log *K*) were higher than those of log K_{ow} and log K_{aw} because of the higher $SD_{training}$ of log
359 K_{oilw} (Figure S8).

360 The data set for log K_{lipw}[14] had the benefit of excellent coverage of the AD probes; only
361 5 out of 25 AD probes exhibited h/h_{mean} > 3 (Figure 5B). A wealth of data for hydrophobic
362 compounds (e.g., PAHs), substituted phenols, hormones, and pharmaceuticals in addition to
363 simple aliphatic and aromatic and polar and nonpolar compounds with varying sizes resulted
364 in the low h/h_{mean} for the AD probes. Because of the low h/h_{mean} and high *n*, the Δ(log *K*) values
365 were similar for all AD probes. Nevertheless, the values of $\Delta(\log K)_{95\%PI}$ and $\Delta(\log K)_{99\%PI}$ (~ 0.6
366 and ~ 0.8, respectively) for log K_{lipw} were higher than those for log K_{ow} by a factor of ~ 2,
367 because the $SD_{training}$ of log K_{lipw} was higher by the same factor.

368 Figures 5C and 5D show illustrative examples of PP-LFERs with limited training data.
369 The data set of fulvic acid/water partition coefficients ($K_{FA/w}$)[20] comprised 34 training data,
370 and 16 out of 25 AD probes were extrapolated (h/h_{mean} > 3). The major difference from log
371 K_{ow} and log K_{lipw} was the wide range of Δ(log *K*); the $\Delta(\log K)_{95\%PI}$ and $\Delta(\log K)_{99\%PI}$ values for
372 log $K_{FA/w}$ were in the range of 0.5–1.0 and 0.7–1.4, respectively. The data set of activated
373 carbon/water partition coefficients ($K_{AC/w}$)[23] was a clearer example of insufficient calibration.

374 It only contained 14 training data, and all AD probes were considered extrapolated (h/h_{mean},
375 8–480). Although the model fitting seemed to be good (the stated mean error, 0.2),[23] the PIs
376 were extremely broad, with $\Delta(\log K)_{95\%PI}$ and $\Delta(\log K)_{99\%PI}$ being 1.0–6.8 and 1.5–10,
377 respectively. These results indicate that PP-LFERs from such small training sets will have a
378 limited predictive ability for external compounds. Conversely, the calculation of *h* and the PIs
379 will be most useful for such poorly calibrated PP-LFERs, as they can identify compounds for
380 which the precision of prediction is still acceptable.

381 In SI-10 of the SI, a comparative discussion is provided for three data sets of log
382 K_{oc}[13,21,22] in terms of their ADs. These data sets possessed different characteristics, which were
383 demonstrated by the AD probes.

384 Overall, it can be concluded that the 25 AD probes are useful in illustrating the
385 strength and weakness of calibrated PP-LFERs. The missing classes of compounds in the
386 training data, e.g., large hydrophobic compounds and multifunctional polar compounds, can
387 be identified using the h/h_{mean} values, and the associated elevation of error margins can be
388 evaluated by calculating the PIs.

389

390 **3.5 Practical implications**

391 This study demonstrated that extrapolation was error-prone when the number of training
392 data was limited and the h/h_{mean} value was extremely high. In contrast, well-calibrated PP-
393 LFERs with many training data (e.g., 100) were highly robust against extrapolation. For
394 partition coefficients between solvent phases or solvent and air such as K_{ow} and K_{aw}, the data
395 are typically accurate and abundant. Thus, extrapolations can frequently result in low
396 prediction errors. Extrapolation matters for heterogeneous environmental, biological, and
397 technical phases, because the data are often limited, and $SD_{training}$ tends to be large.

398 The commonly used threshold of $h < 3\ h_{mean}$ appeared not to be useful in defining the
399 AD of PP-LFER models. Alternatively, two possible ways were proposed in this article to define
400 the AD based on the calculation of the PI. For practical purposes, presenting the PIs for each
401 time of prediction may be highly recommended. For example, using the PP-LFER, log K_{ow} for
402 hexachlorobenzene is predicted as 5.49 with a 95% PI of [5.16, 5.81]. With these PI values,
403 the model user can appreciate the reliability of the prediction and decide whether the value
404 is taken or not, following the accuracy required for the given model use. It could be claimed

that calculating the PI each time is more important and useful than seeking a strict definition of the AD, because the former presents a quantitative estimate of the error range, while the latter is a qualitative, binomial indicator with an arbitrary cutoff in the end.

To develop a robust PP-LFER, the training set should contain (A) a large number (>60, preferably >100) of (B) accurate experimental *K* data for (C) diverse compounds with (D) accurate descriptors available. (A) decreases $t_{\alpha/2,n-k-1}$ and *h*, (B) and (D) decrease $SD_{training}$, and (C) decreases *h* in eq 7, all contributing to tight PIs. The predictive performance of an empirical model is always restricted by the quality and quantity of the underlying experimental data. The improvement in data accuracy and availability will contribute to the further development of PP-LFER approaches.

Extended use of the PI may be considered for evaluating the AD of QSARs that are derived by the multiple linear regression analysis. The calculation of the PI is no more complex than *h* is, but the former provides far more insights into the reliability of predictions, as discussed above. Noteworthily, the success of applying the PI for PP-LFERs may be partially related to the excellent linearity of the PP-LFER descriptors to log *K*. The suitability of the PI for various existing QSAR descriptors and properties warrants future investigation.

Associated content

Supporting information

The Supporting Information is available free of charge at …

Additional explanations for *h* and PIs, tables listing the used *K* data and AD probes, additional figures for Tests 1 and 2 (PDF)

MS Excel file with a macro to calculate *h* and the PIs (XLSM)

Conflicts of interest

The author has no conflicts of interest associated with this article.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP18K05204 and JP16K16216 and by the MEXT/JST Tenure Track Promotion Program. Kai-Uwe Goss and Jort Hammer are thanked for their valuable comments on an earlier version of this manuscript.

436

437 ## References

438 1. Abraham, M. H.; Ibrahim, A.; Zissimos, A. M., Determination of sets of solute
439 descriptors from chromatographic measurements. *J. Chromatogr. A* **2004,** *1037*, (1-2), 29-47.
440 2. Goss, K.-U.; Schwarzenbach, R. P., Linear free energy relationships used to evaluate
441 equilibrium partitioning of organic compounds. *Environ. Sci. Technol.* **2001,** *35*, (1), 1-9.
442 3. Endo, S.; Goss, K.-U., Applications of Polyparameter Linear Free Energy Relationships
443 in Environmental Chemistry. *Environ. Sci. Technol.* **2014,** *48*, (21), 12477-12491.
444 4. Poole, C. F.; Ariyasena, T. C.; Lenca, N., Estimation of the environmental properties of
445 compounds from chromatographic measurements and the solvation parameter model. *J.*
446 *Chromatogr. A* **2013,** *1317*, 85-104.
447 5. Goss, K.-U., Predicting the equilibrium partitioning of organic compounds using just
448 one linear solvation energy relationship (LSER). *Fluid Phase Equilib.* **2005,** *233*, (1), 19-22.
449 6. Netzeva, T. I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M. T.; Gramatica, P.;
450 Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.;
451 Patlewicz, G. Y.; Perkins, R.; Roberts, D.; Schultz, T.; Stanton, D. W.; van de Sandt, J. J.; Tong,
452 W.; Veith, G.; Yang, C., Current status of methods for defining the applicability domain of
453 (quantitative) structure-activity relationships. The report and recommendations of ECVAM
454 Workshop 52. *ATLA Altern. Lab. Anim.* **2005,** *33*, (2), 155-73.
455 7. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T., QSAR Applicability Domain
456 Estimation by Projection of the Training Set in Descriptor Space: A Review. *ATLA Altern. Lab.*
457 *Anim.* **2005,** *33*, (5), 445-459.
458 8. Gramatica, P., Principles of QSAR models validation: internal and external. *QSAR Comb*
459 *Sci.* **2007,** *26*, (5), 694-701.
460 9. Gramatica, P.; Giani, E.; Papa, E., Statistical external validation and consensus
461 modeling: A QSPR case study for K_{oc} prediction. *J. Mol. Graph. Model.* **2007,** *25*, (6), 755-766.
462 10. Abraham, M. H.; Chadha, H. S.; Whiting, G. S.; Mitchell, R. C., Hydrogen bonding. 32.
463 An analysis of water-octanol and water-alkane partitioning and the Δlog P parameter of seiler.
464 *J. Pharm. Sci.* **1994,** *83*, (8), 1085-100.

11. Abraham, M. H.; Andonian-Haftvan, J.; Whiting, G. S.; Leo, A.; Taft, R. S., Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapors in water at 298 K, and a new method for its determination. *J. Chem. Soc. Perkin Trans. 2* **1994**, (8), 1777-91.

12. Geisler, A.; Endo, S.; Goss, K.-U., Partitioning of Organic Chemicals to Storage Lipids: Elucidating the Dependence on Fatty Acid Composition and Temperature. *Environ. Sci. Technol.* **2012,** *46*, (17), 9519-9524.

13. Bronner, G.; Goss, K.-U., Predicting sorption of pesticides and other multifunctional organic chemicals to soil organic carbon. *Environ. Sci. Technol.* **2011,** *45*, (4), 1313-1319.

14. Endo, S.; Escher, B. I.; Goss, K.-U., Capacities of Membrane Lipids to Accumulate Neutral Organic Chemicals. *Environ. Sci. Technol.* **2011,** *45*, (14), 5912-5921.

15. Endo, S.; Goss, K.-U., Serum Albumin Binding of Structurally Diverse Neutral Organic Compounds: Data and Models. *Chem. Res. Toxicol.* **2011,** *24*, (12), 2293-2301.

16. Endo, S.; Goss, K.-U., Predicting Partition Coefficients of Polyfluorinated and Organosilicon Compounds using Polyparameter Linear Free Energy Relationships (PP-LFERs). *Environ. Sci. Technol.* **2014,** *48*, (5), 2776-2784.

17. Ulrich, N.; Endo, S.; Brown, T. N.; Watanabe, N.; Bronner, G.; Abraham, M. H.; Goss, K. U., UFZ-LSER database v 3.2 [Internet]. **2017**.

18. Abraham, M. H.; Ibrahim, A.; Acree, W. E., Jr., Air to lung partition coefficients for volatile organic compounds and blood to lung partition coefficients for volatile organic compounds and drugs. *Eur. J. Med. Chem.* **2008,** *43*, (3), 478-485.

19. Mackay, D.; Arnot, J. A., The Application of Fugacity and Activity to Simulating the Environmental Fate of Organic Contaminants. *J. Chem. Eng. Data* **2011,** *56*, (4), 1348-1355.

20. Neale, P. A.; Escher, B. I.; Goss, K.-U.; Endo, S., Evaluating dissolved organic carbon–water partitioning using polyparameter linear free energy relationships: Implications for the fate of disinfection by-products. *Water Res.* **2012,** *46*, (11), 3637-3645.

21. Nguyen, T. H.; Goss, K.-U.; Ball, W. P., Polyparameter linear free energy relationships for estimating the equilibrium partition of organic compounds between water and the natural organic matter in soils and sediments. *Environ. Sci. Technol.* **2005,** *39*, (4), 913-924.

22. Endo, S.; Grathwohl, P.; Haderlein, S. B.; Schmidt, T. C., LFERs for soil organic carbon-water distribution coefficients (K_{OC}) at environmentally relevant sorbate concentrations. *Environ. Sci. Technol.* **2009,** *43*, (9), 3094-3100.

23. Shih, Y.-h.; Gschwend, P. M., Evaluating Activated Carbon−Water Sorption Coefficients of Organic Compounds Using a Linear Solvation Energy Relationship Approach and Sorbate Chemical Activities. *Environ. Sci. Technol.* **2009,** *43*, (3), 851-857.