Preprint

Applicability Domain of Polyparameter Linear Free Energy Relationship Models for Predicting Equilibrium Partition Coefficients

Preprint 5th January 2022 version

Satoshi Endo*[a,b]

Polyparameter linear free energy relationships (PP-LFERs) are accurate and robust models to predict equilibrium partition coefficients (*K*) of organic chemicals. The accuracy of predictions by a PP-LEFR depends on the composition of the respective calibration data set. It is generally expected that extrapolation outside the model calibration domain is less accurate than interpolation. In this study, the applicability domain (AD) of PP-LFERs is systematically evaluated by calculation of the leverage (*h*), a measure of distance from the calibration set in the descriptor space. Repeated simulations with experimental data show that the root mean squared error of predictions increases with *h*, and that large prediction errors (>3 $SD_{training}$, the standard deviation of training data) occur more frequently when *h* exceeds the common threshold of 3 h_{mean}, where h_{mean} is the mean *h* of all training compounds. Nevetheless, analysis also shows that well-calibrated PP-LFERs with many (e.g., 100), diverse, and accurate training data are highly robust against extrapolation; extreme prediction errors (> 5 $SD_{training}$) are rare. For such PP-LFERs, 3 h_{mean} may be too strict as the cutoff for AD. Evaluation of published PP-LFERs in terms of their AD using 25 chemically diverse, environmentally relevant chemicals as AD probes indicated that many reported PP-LFERs do not cover organosiloxanes, per- and polyfluorinated alkylsubstances, highly polar chemicals, and/or highly hydrophobic chemicals in their AD. It is concluded that calculation of *h* is useful to identify model extrapolations as well as the strengths and weaknesses of the trained PP-LFERs.

1. Introduction

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

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

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

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

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

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

Because PP-LFERs are derived from a multiple linear regression, their applicability domain (AD) is related to the training (or calibration) set of chemicals. It is generally expected that extrapolation (i.e., prediction beyond the calibrated domain) tends to be less accurate than interpolation (i.e., prediction within the calibrated domain). Moreover, a long-range extrapolation is expected to be more error-prone than a short-range extrapolation. However, in a multidimensional space (here 5 descriptors) it is not obvious how the terms interpolation and extrapolation can be defined and how a quantitative relationship between the extent of extrapolation and prediction accuracy may be established. It is also important that an extrapolation can be less accurate but is not necessarily

[a] *Health and Environmental Risk Division, National Institute for Environmental Studies (NIES), Onogawa 16-2, 305-8506 Tsukuba, Ibaraki, Japan*
[b] *Graduate School of Engineering, Osaka City University, Sugimoto 3-3-138, Sumiyoshi, 558-8585 Osaka, Japan*
*Corresponding author address: Health and Environmental Risk Division, National Institute for Environmental Studies (NIES), Onogawa 16-2, 305-8506 Tsukuba, Ibaraki, Japan, phone: ++81-29-850-2695, email: endo.satoshi@nies.go.jp
Electronic Supplementary Information (ESI) available

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

Among various approaches, calculation of the leverages is useful to define and evaluate the AD for linear regression models.[6-8] The leverage is a quantitative measure of the distance of a data from the entire set of calibration data and is calculated solely with independent variables. Leverage calculation is often 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. Usually, a threshold value is set to draw a line between interpolation and extrapolation. Since PP-LFERs are a linear regression model, calculation of the leverage should give an insight into their AD.

The purposes of this work are two-fold: (1) To demonstrate quantitatively how the prediction accuracy of a PP-LFER decreases when moving away from a specific domain of calibration defined by the leverage; and (2) to evaluate several calibration sets for PP-LFERs in terms of their AD using a newly proposed concept of AD probes. On the basis of these, it is discussed how the AD should be defined and evaluated for PP-LFER models. The information is also helpful for future development of PP-LFERs because it guides the way to an optimized calibration data set and informs us of missing experimental data to construct such a data set.

2. Methods

2.1 Leverage calculation

A matrix expression of the PP-LFER regression appears,

$$y = X\,\beta + \varepsilon \qquad (4)$$

y is the vector of log K observations. X is the design matrix containing solute descriptors of n training chemicals.

$$X = \begin{bmatrix} 1 & E_1 & S_1 & A_1 & B_1 & V_1 \\ \vdots & & \vdots & \vdots & & \vdots \\ 1 & E_n & S_n & A_n & B_n & V_n \end{bmatrix} \qquad (5)$$

β is the vector of regression coefficients. ε is the error vector. The hat matrix (H) is defined as,

$$H = X\,(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} \qquad (6)$$

The diagonals of H (i.e., h_{ii}) are referred to as the leverages and infer the distance of each calibration compound from all the others in terms of the solute descriptor combination. h_{ii} is constrained to between 0 and 1, and the sum of h_{ii} for the n training chemicals is equal to the number of fitting parameters p, which is 6 for PP-LFERs (including the fitting constant). Typically, $h_{ii} > 3h_{mean}$ is considered too high,[6-8] where h_{mean} is the mean of h_{ii} for all calibration chemicals and is equal to p/n. Too high h_{ii} means that the respective calibration compound is an outlier in terms of its descriptors and has a strong influence on the regression coefficients; removing such a compound from the calibration set is then advised.

To check extrapolation for a compound j which is not included in the calibration set, one calculates h as,

$$h = x_j\,(X^{\mathrm{T}}X)^{-1}\,x_j^{\mathrm{T}} \qquad (7)$$

where,

$$x_j = [1 \quad E_j \quad S_j \quad A_j \quad B_j \quad V_j] \qquad (8)$$

Too high h indicates that compound j is distant from the calibration data set in terms of descriptor values and that the prediction for y_j by the trained model is an extrapolation. In a similar way to identification of outliers in the training set described above, $3h_{mean}$ is considered the threshold value for extrapolation of missing data. However, it has not been debated to what extent the prediction deteriorates above this threshold in case of PP-LFER models. In this work, the following two tests were performed to make use of h for delineating the AD of PP-LFERs.

2.2 Test 1: Comparison of leverages and prediction errors

In the first test, it was examined how prediction errors vary with h. For this test, six experimental data sets of partition coefficients were used: octanol/water (K_{ow}, n = 314),[9] air/water (K_{aw}, n = 390),[10] oil/water (K_{oilw}, n = 247),[11] soil organic carbon/water (K_{oc}, n = 79),[12] phospholipid liposome/water (K_{lipw}, n = 131),[13] and bovine serum albumin/water (K_{BSAw}, n = 82)[14] partition coefficients. These data sets comprise a relatively large number of compounds, have environmental and toxicological relevance, and represent both homogeneous and heterogeneous phases. Regarding the last point, K_{ow}, K_{aw}, and K_{oilw} are partition coefficients between two homogeneous solvents, whereas K_{oc}, K_{lipw}, and K_{BSAw} involve a heterogeneous or anisotropic phase. The K values and solute descriptors were

Table 1. Ranges of partition coefficients and solute descriptors (min/max) considered in this study.

	n of compounds	Log *K*	Descriptors						SD[a]	Ref
			E	*S*	*A*	*B*	*V*	*L*		
Log K_{ow}	314	-1.38/5.65	-0.60/2.81	-0.20/1.91	0.00/0.82	0.00/0.84	0.17/1.67	-0.82/8.83	0.154	9
Log K_{aw}	390	-8.07/2.32	-0.60/1.67	-0.20/1.91	0.00/0.82	0.00/1.06	0.17/1.67	-0.80/6.92	0.156	10
Log K_{oilw}	247	-2.66/9.88	-0.79/2.81	-0.30/1.72	0.00/0.76	0.00/0.97	0.25/2.36	-0.82/8.83	0.286	11
Log K_{oc}	79	0.64/4.39	-0.24/2.06	0.00/1.95	0.00/0.99	0.00/1.10	0.64/2.56	1.95/11.11	0.250	12
Log K_{lipw}	131	-0.79/7.86	0.00/4.07	0.00/3.29	0.00/1.14	0.00/1.63	0.31/2.62	0.97/13.26	0.285	13
Log K_{BSAw}	82	1.48/4.76	-0.24/4.07	0.00/2.05	0.00/0.99	0.00/1.38	0.71/2.28	1.75/13.45	0.422	14

[a] Standard deviation of PP-LFER (eq 1) when all compounds are used for model training.

taken from the references cited above and are listed in the electronic supplementary information (ESI), S-1, Tables S1–S6. A summary of the data is provided in Table 1.

To evaluate prediction accuracy, each data set of *K* was divided into training and test sets. Training compounds were randomly selected from the whole data set. The number of the training compounds ($n_{training}$) was 20, 30, 40, 50, 75 or 100. Rather small $n_{training}$ of 20–30 was also included in this test to simulate cases of insufficient calibration. All compounds that were not selected as training compounds were used as test compounds. The training set was used to calibrate eq 1 and generate the hat matrix. The calibrated equation was used to predict log *K* for the test compounds and the hat matrix to derive *h* values. Prediction errors (predicted log *K* minus experimental log *K*) were then calculated and compared with *h*. For each data set of *K* and each $n_{training}$, the cycle of "random generation of a training set", "calibration of the PP-LFER", and "prediction for the test set" was repeated 500 times. This number is rather arbitrary but appears to be enough for stable statistics. All calculations were performed with *R* software.

2.3 Test 2: Evaluating reported PP-LFERs for their domain of applicability

In the second test, *h* calculation was applied to evaluate the AD of reported PP-LFER equations. In this test, the calibration chemicals used to derive the respective PP-LFER were extracted from the literature and their solute descriptors were used to calculate the hat matrix. Then, using the hat matrix, *h* values for 25 selected chemicals were calculated. These 25 chemicals, referred to as AD probes here, were selected from their wide variations in descriptor values, structural diversity, and environmental relevance (Table 2). They represent aliphatic and aromatic, polar and nonpolar, and small to large compounds. The 25 AD probes additionally include multifunctional polar compounds such as various pesticides and pharmaceuticals as well as neutral highly fluorinated and organosilicon compounds. The AD was judged large if *h* values of many of these 25 AD probes were low. Solute descriptors for AD probes were obtained from the UFZ-LSER database.[15] Note that Test 2 does not use *K* values at all. The evaluation is solely based on the solute descriptors. Note also that there exist chemicals that

Table 2. Twenty-five applicability domain (AD) probes used for testing the reported PP-LFERs.

			E	*S*	*A*	*B*	*V*	*L*
aliphatic	nonpolar	dichloromethane	0.39	0.57	0.10	0.05	0.494	2.019
		hexachloroethane	0.68	0.68	0.00	0.00	1.125	4.718
		n-hexadecane	0.00	0.00	0.00	0.00	2.363	7.714
	H-acceptor	methyl *tert*-butyl ether	0.02	0.28	0.00	0.54	0.872	2.270
		molinate	0.88	1.09	0.00	0.70	1.547	6.578
		tri-*n*-butyl phosphate	-0.10	0.90	0.00	1.21	2.239	7.370
	H-donor	*tert*-butyl alcohol	0.18	0.30	0.31	0.60	0.731	1.963
		decanoic acid	0.12	0.64	0.62	0.45	1.592	5.698
aromatic	nonpolar	benzene	0.61	0.52	0.00	0.14	0.716	2.786
		hexachlorobenzene	1.49	0.99	0.00	0.00	1.451	7.624
		phenanthrene	2.06	1.29	0.00	0.29	1.454	7.632
		PCB 180	2.29	1.87	0.00	0.09	2.181	10.415
		benzo[*ghi*]perylene	3.61	2.11	0.00	0.44	2.084	12.707
	H-acceptor	nitrobenzene	0.87	1.11	0.00	0.28	0.891	4.557
		benzophenone	1.45	1.50	0.00	0.50	1.481	6.955
		di-*n*-butyl phthalate	0.69	1.30	0.00	0.94	2.274	8.553
	H-donor	phenol	0.81	0.89	0.60	0.30	0.775	3.766
		pentachlorophenol	1.22	0.87	0.96	0.01	1.387	6.822
		bisphenol A	1.61	1.56	0.99	0.91	1.864	9.603
multifunctional polar		caffeine	1.50	1.82	0.08	1.25	1.363	7.838
		metolachlor	1.15	1.01	0.07	1.38	2.281	8.863
		diuron	1.28	1.60	0.57	0.70	1.599	8.060
		estradiol	1.80	1.77	0.86	1.10	2.199	11.107
neutral PFAS		8:2 FTOH	-1.56	0.14	0.62	0.31	2.220	3.470
organosilicon		D5	-0.70	-0.10	0.00	0.50	2.931	5.242

8:2 FTOH, 1H,1H,2H,2H-perfluorodecan-1-ol; D5, decamethylcyclopentasiloxane.

have extreme descriptor values which are not covered by the 25 AD probes. Extreme examples include an antibiotic erythromycin (E = 2.90, S = 3.73, A = 1.25, B = 4.96, V = 5.773) and a cardiac glycoside digoxin (E = 3.67, S = 4.68, A = 1.58, B = 5.07, V = 5.753),[16] both of which have exceptionally high S, B and V values. Such chemicals are not used for calibration, are always out of the calibration domain, and thus not necessary to include specifically in the evaluation here.

3. Results and discussion

3.1 Increase in prediction error with h (Test 1)

Fig. 1 (and its extended version, Fig. S1) show the RMSEs for training and testing sets randomly generated 500 times. Test chemicals were grouped into several bins according to h/h_{mean} before RMSEs were calculated. As mentioned above, h/h_{mean} = 3 is the common threshold for extrapolation. For a given data set of K and $n_{training}$, the RMSE for the test chemicals increases with h. The increasing trend of RMSE with h is particularly clear for simulations with small $n_{training}$ (i.e., 20, 30) but is rather unclear for high $n_{training}$ (i.e., 75, 100). A reason for the unclear trend may be that, when $n_{training}$ is large, n_{test} becomes small and cannot provide a representative RMSE particularly for high h/h_{mean} bins. Also, the increase of RMSE with h is clearer for log K_{ow} and log K_{aw} than for the other partition coefficients. This may also be because of the larger number of data (thus large n_{test}) available for log K_{ow} and log K_{aw} as well as the good fit of the PP-LFER equation to these data, both of which prevent the RMSE from being influenced substantially by a few specific chemicals.

With increasing $n_{training}$, the training RMSE increases and the test RMSE decreases slightly. As a result, the difference in RMSE between training and test sets become smaller. Fig. 2 and S2 show the RMSE values of test data relative to that of training data. Taking the "2 < h/h_{mean} < 3" bin of log K_{ow} test data as example, we can see that the test RMSEs are higher than the training RMSE by a factor of 1.75, 1.52, 1.38, and 1.30 for $n_{training}$ = 20, 40, 75, and 100, respectively. For any partition coefficient considered and with $n_{training}$ ≥ 75, the test-to-training RMSE ratio < 1.1 for "h/h_{mean} < 1" and < 1.3 for "1 < h/h_{mean} < 2". Thus, if the PP-LFER is trained with a sufficiently large number of data, we can expect that RMSE for clear interpolations (i.e., h/h_{mean} < 2) well resembles that for the training set. Even for extrapolation cases (i.e., "3 < h/h_{mean} < 4"), the relative RMSE < 1.5 when $n_{training}$ ≥ 50 and < 2 when $n_{training}$ ≥ 20. Thus, although PP-LFER predictions become less accurate with increasing h, RMSE is still less than two times the training RMSE, suggesting that the model is highly robust against short-range extrapolation up to h/h_{mean} = 4.

Fig. 1 and S1 also illustrate the difference in quality of PP-LFER fitting between partition coefficients. Training RMSEs for log K_{ow} and log K_{aw} are generally low (ca 0.1). In contrast, the other four systems show larger training RMSEs (0.2–0.4) due to heterogeneous phases or data inaccuracy. While 3 is often used as the h/h_{mean} threshold for extrapolation,[6-8] the actual threshold value can be adapted to the required accuracy of predictions. For example, if the required accuracy is 0.3 log units, which is typically the level of accuracy of contaminant fate models,[17] then fairly far extrapolations (even h/h_{mean} of > 4) of the PP-LFERs for log K_{ow} and log K_{aw} can be allowed, according to Fig. 1. In contrast, a strict threshold, say h/h_{mean} of < 1 or 2,

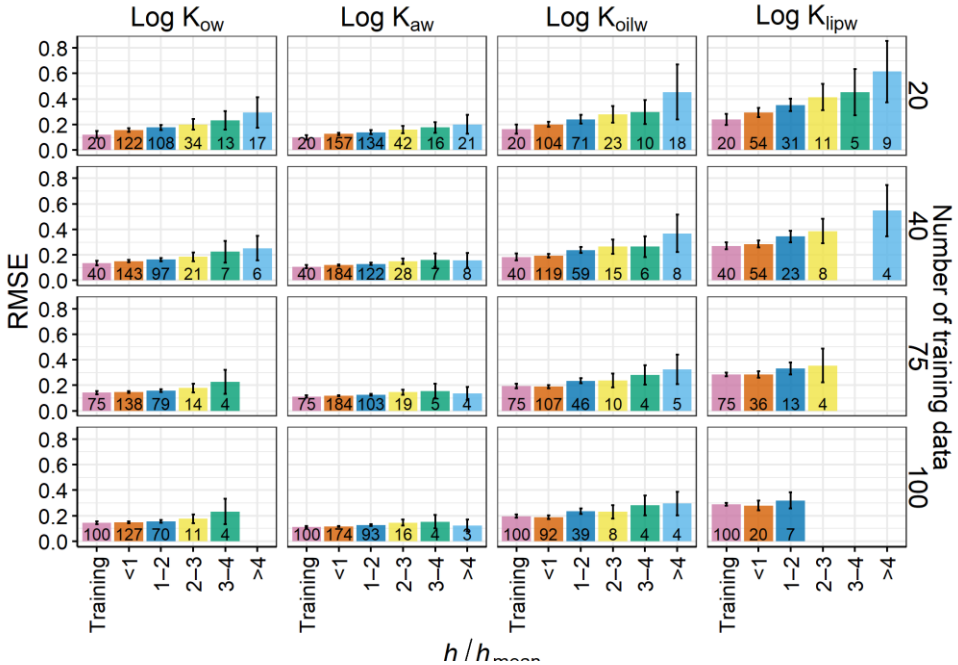

Fig. 1. RMSEs for training and test data. Columns and error bars indicate the means and the standard deviations, respectively, for 500 repeated simulations. Test data were sorted into five bins according to their h/h_{mean}. Numbers given in the plot show the mean number of data. Columns with the mean number of data < 3 are not shown. The plots for $n_{training}$ = 30 and 50 as well as for log K_{oc} and log K_{BSAw} are given in the ESI, S-2, Fig. S1.

should be set to log K_{oc}, log K_{lipw}, and log K_{BSAw} to comply the criterion of 0.3 log unit RMSE, depending on the quality of PP-LFER fit and $n_{training}$ (see Fig. S1).

Test 1 was also performed with eq 3, the PP-LFER equation that uses L instead of E but the results were similar to those of eq 1 and thus are not discussed here.

3.2 Large prediction errors (Test 1)

Along with an average of prediction errors (e.g., RMSE) discussed above, the risk of an extremely inaccurate prediction is also of particular importance. In Fig. 3, prediction errors for individual data of log K_{ow} and log K_{lipw} are plotted against h/h_{mean}. All other data are presented in the ESI, S-3, Fig. S3. The same data sets as above (generated by 500 times model training and testing) are used. Prediction errors are normalized to the SD of the trained PP-LFER ($SD_{training}$), as the SD is the most frequently provided error estimate for trained PP-LFERs. Fig. 4 provides percentages of large prediction errors, as defined by the absolute residuals being > 3 times $SD_{training}$, for both interpolation and extrapolation. Interesting observations in Fig. 3 and 4 include:

With small $n_{training}$ (e.g., 20, 30), both h/h_{mean} (x-axis) and the normalized error (y-axis) for test data distribute widely (Fig. 3). Very large errors (|error/$SD_{training}$| > 5 or even > 10) sometimes occur when $n_{training}$ is small and h/h_{mean} is large (> 10). In contrast, when $n_{training}$ is high (e.g., 75, 100), training and test data distribute similarly in terms of both h/h_{mean} and the normalized error.

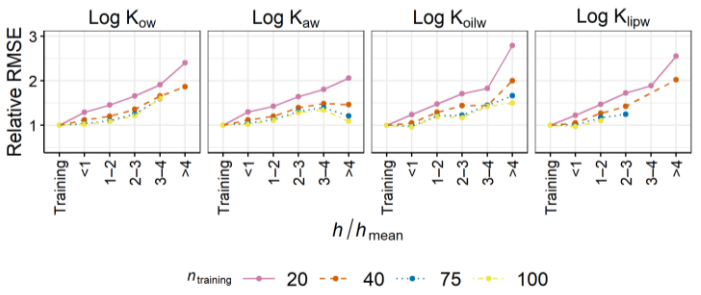

Fig. 2. RMSEs of test data, sorted according to h/h_{mean}, relative to RMSE of training data. The plots for $n_{training}$ = 30 and 50 as well as for log K_{oc} and log K_{BSAw} are given in the ESI, S-2, Fig. S2.

The percentage of large prediction errors (|error/$SD_{training}$| > 3) is clearly higher in the case of extrapolation (h/h_{mean} > 3) than interpolation (h/h_{mean} < 3) (Fig. 4). The percentage decreases with $n_{training}$. Taking log K_{ow} as example: When $n_{training}$ = 20, 2.8% of interpolations and 15% of extrapolations suffered from large prediction errors. In contrast, when $n_{training}$ = 100, 1.0% of interpolations and 5.8% of extrapolations resulted in large prediction errors, which conversely means that 94% of extrapolations ended up with errors within 3 $SD_{training}$.

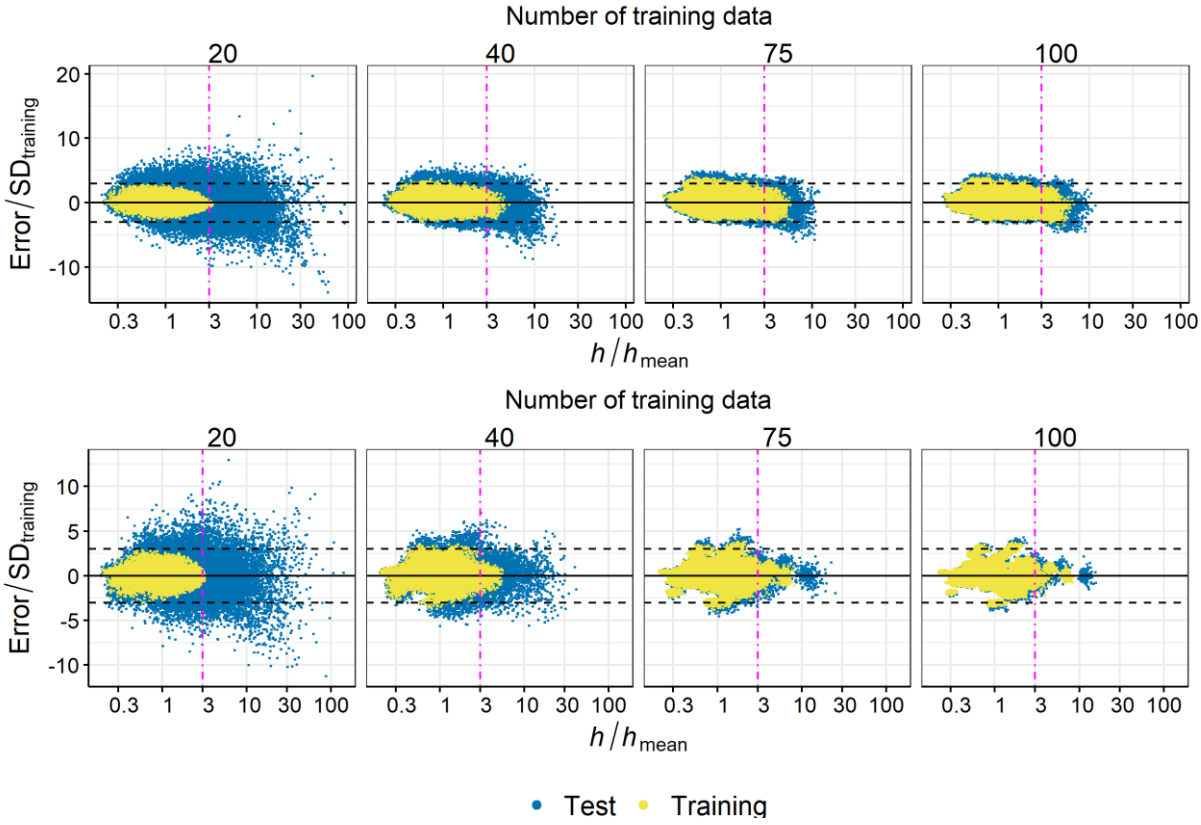

Fig. 3. Prediction errors normalized to $SD_{training}$ plotted against h/h_{mean}. All results from 500 simulations are shown. h/h_{mean} = 3 is indicated with a dash-dotted vertical line. Dashed horizontal lines indicate errors (or residuals) being 3 times $SD_{training}$. Top, log K_{ow}; bottom, log K_{lipw}. More data are in ESI, S-3.

K_{lipw} data scatter more than K_{ow} data, reflecting the heterogeneous phase of liposomes, but the general trends as described in (1) and (2) still apply.

Altogether, it can be said that the h/h_{mean} cutoff of 3 is useful to identify "risky predictions" that more frequently cause extreme inaccuracy. However, h/h_{mean} alone is not a versatile criterion, as $n_{training}$ appears to have an influence on the level of prediction errors as well. Plots in Fig. 3, 4 and S3 suggest that, in case $n_{training}$ is large, h/h_{mean} = 3 might be too strict a threshold, and that h/h_{mean} up to ~5 could be accepted. It should be noted that not only the number, but the diversity and quality of training data should also have an influence on the prediction accuracy on the absolute scale. The data sets used here consist of carefully evaluated experimental K data and cover various compound groups (see below for evaluation of the AD with AD probes). Although randomly selected, 75 out of 79–390 quality data can rarely lead to extremely biased calibration. Therefore, the observed dependency on $n_{training}$ can be somewhat specific to the data sets considered in this work. That means, even if the number of data is large, highly biased training data or with inaccurate experimental K data could generate a biased PP-LFER and lead to frequent occurrence of large prediction errors.

3.3 Per- and polyfluoroalkyl substances and organosilicon compounds

Per- and polyfluoroalkyl substances (PFASs) and organosilicon compounds (OSCs) have extremely weak van der Waals interaction properties, and hence, the E and L values are comparatively low for their molecular sizes. Therefore, PP-LFERs often have to be extrapolated to predict K values for these compounds unless PFASs and/or OSCs are included in the training data set. Using the 500 trained PP-LFERs for log K_{ow}

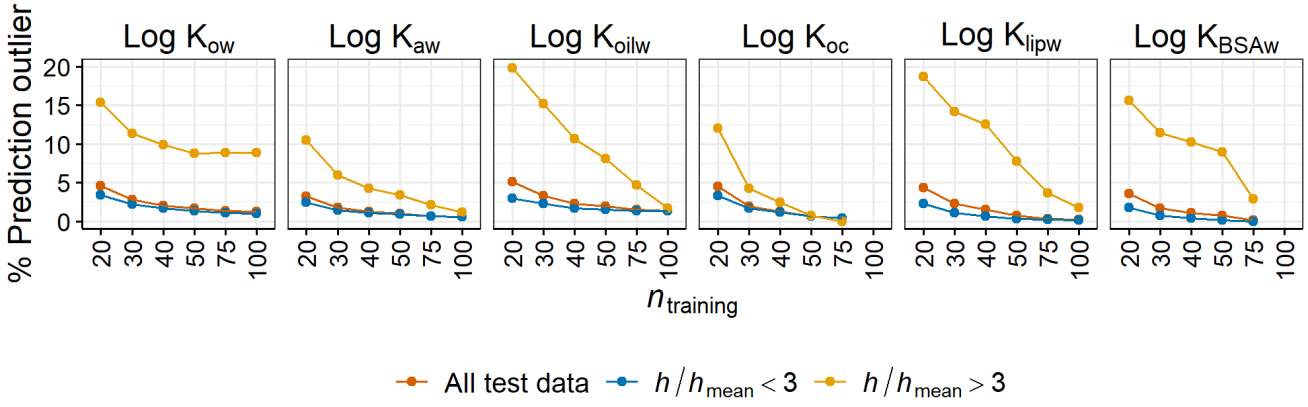

Fig. 4. Percentage of prediction outliers (i.e., absolute error > 3 times $SD_{training}$) in the 500 repeated simulations.

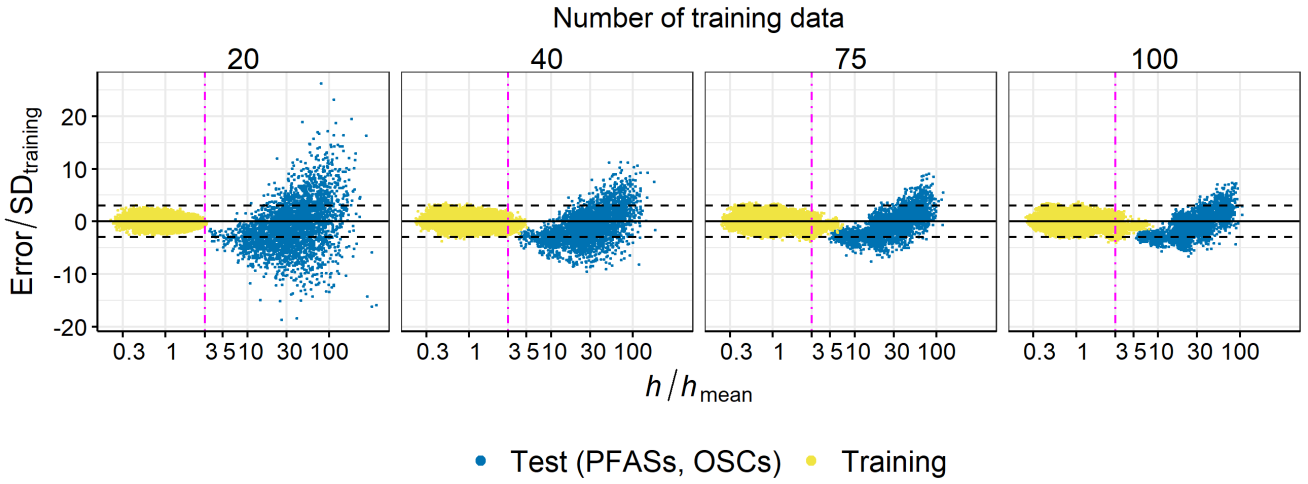

Fig. 5. Prediction errors for PFASs and OSCs normalized to $SD_{training}$ plotted against h/h_{mean}. Log K_{ow} data were used. h/h_{mean} = 3 is indicated with a dash-dotted vertical line. Dashed horizontal lines indicate errors being 3 times $SD_{training}$. Equation 3 was used for this plot (see the text for more details). More data are in the ESI, S-4 and S-5.

generated above, log K_{ow} of 3 PFASs (4:2 fluorotelomer alcohol (FTOH), 6:2 FTOH, 8:2 FTOH) and 3 OSCs (octamethylcyclotetrasiloxane (D4), decamethylcyclopentasiloxane (D5), dodecamethylcyclohexasiloxane (D6)) were predicted and compared to experimental data (Fig. 5). Note that, for Fig. 5, eq 3 instead of eq 1 was used, because the latter does not fit data for PFASs and OSCs well (ref 18; also see the ESI, S-4, S-5). h/h_{mean} for these 6 chemicals were 3–100 with any $n_{training}$ used, indicating strong extrapolations. That said, predictions appear to improve with increasing $n_{training}$. When $n_{training}$ = 100, for instance, even largely extrapolated FTOHs (h/h_{mean} ~ 30) were often predicted within 3 $SD_{training}$.

Lessons that can be learned from predicting PFASs and OSCs are that (i) long-range extrapolation (h/h_{mean} > 10) can be detrimental to the prediction accuracy, but that (ii) a large number of training data (with high data quality and diversity) improves the quality of the PP-LFER and can minimize the errors associated with long-range extrapolation. It is obvious, however, that including data for PFASs and OSCs in the calibration set is desirable for accurately predicting *K*'s of PFASs and OSCs.[18]

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

Using 25 AD probes, training sets of eight published PP-LFER equations[9-14,19,20] including some of those presented above were evaluated (Fig. 6).

First of all, none of the training sets considered clearly encompassed all 25 AD probes within their AD (h/h_{mean} < 3). Particularly, 8:2 FTOH and D5 always appeared as highly extrapolated chemicals (h/h_{mean} = 10–100), reflecting the fact that PFASs and OSCs are neither included in any of the training sets nor are well represented by other training chemicals.

Within each class of chemicals, small chemicals (e.g., dichloromethane, methyl *tert*-butyl ether, benzene) show smaller h/h_{mean} than large chemicals (e.g., hexadecane, tri-*n*-

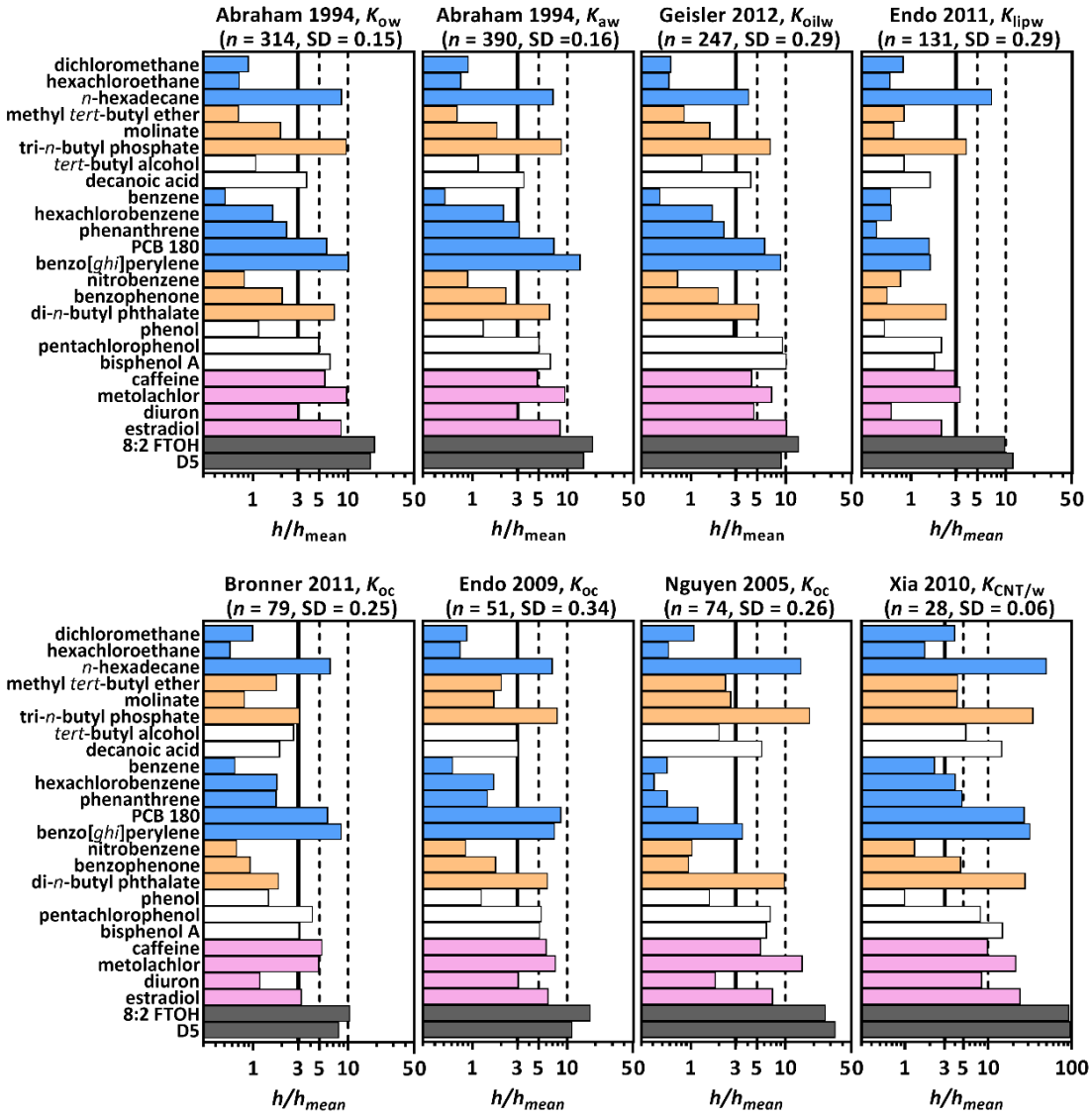

Fig. 6. h/h_{mean} of 25 applicability domain (AD) probes calculated with the training data sets of 8 literature PP-LFERs.

butyl phosphate, benzo[*ghi*]perylene). Relatively small chemicals are often easy to measure and their data are present in the training set, whereas obtaining data for large chemicals is typically more challenging. Accordingly, PP-LFERs must be extrapolated for large chemicals.

The data sets for log K_{ow},[9] log K_{aw},[10] and log K_{oilw}[11] exhibited a similar pattern in Fig. 6. Small chemicals are well within the AD ($h/h_{mean} < 3$), and large chemicals also are within $h/h_{mean} < 10$. Also considering the large number of training data (n = 247–390) and the relatively low training RMSE for these PP-LFERs, it is expected that a vast range of neutral chemicals of environmental interest can be predicted with acceptable errors. Large hydrophobic chemicals, large polar chemicals, and most importantly, PFAS and OSCs are expected to show larger errors.

The data set for log K_{lipw}[13] enjoys an excellent coverage of the AD probes; 21 out of 25 AD probes showed $h/h_{mean} < 3$. Only hexadecane, 8:2 FTOH, and D5 showed $h/h_{mean} > 5$. The data set covers aliphatic and aromatic as well as polar and nonpolar compounds with varying sizes. A wealth of data for hydrophobic chemicals (PCBs, PAHs), substituted phenols, hormones and pharmaceuticals are characteristic of the log K_{lipw} data set and contribute to the low h/h_{mean} of many AD probes.

Three data sets for log K_{oc} are included in Fig. 6. Bronner's set[12] is well-balanced, covering 18 AD probes within the $h/h_{mean} < 3$ limit. Large hydrophobic chemicals show $h/h_{mean} > 5$, as such chemicals are absent in the data set. Bronner et al.[12] used an HPLC retention method to measure K_{oc}, and the method is not suitable for highly hydrophobic chemicals that are strongly retained by the soil column. Nguyen's data[19] for K_{oc} has the opposite characteristics. The data set covers nonpolar aromatic chemicals very well, whereas polar chemicals (e.g., tri-*n*-butyl phosphate, metolachlor) exhibit large h/h_{mean}. Nguyen's data set is a selection of literature data, which are predominated with nonpolar aromatic compounds. Endo's data set[20] shows relatively high h/h_{mean} for large chemicals, irrespective of polarity, reflecting the types of chemicals included in the data set.

The data set for carbon nanotube/water partition coefficient ($K_{CNT/w}$)[21] is an example of insufficient calibration. All chemicals but one in the data set are aromatic, and all are small and have a simple molecular structure. As a result, h/h_{mean} is < 3 for only 3 probes and > 10 for 9 probes. Together with low $n_{training}$, it is expected that elevated prediction errors may occur to many chemicals. Conversely, the leverage calculation presented here is most useful for such a low-calibrated PP-LFER, as it can identify chemicals for which predictions are considered reliable.

4. Conclusions

This study showed that h calculation helps evaluate the AD of trained PP-LFERs and identify extrapolations in terms of the descriptor space. Test 1 demonstrated that extrapolation is particularly error-prone when the number of training data is limited (e.g., <30) and/or h/h_{mean} value is extremely high (e.g., >10). In contrast, well-calibrated PP-LFERs with many and diverse training data (e.g., 100) are surprisingly robust against extrapolation. For partition coefficients between well-defined, homogeneous phases (e.g., K_{ow}), many extrapolations will result in prediction errors within 3 times SD of the training data. The 25 AD probes are useful to illustrate the strength and weakness of calibrated PP-LFERs. Missing classes of chemicals in the training data, e.g., large hydrophobic chemicals and multifunctional polar chemicals, can be identified by h/h_{mean} values of the 25 AD probes. PFASs and OSCs are often the furthest from the domain of calibration. Including these classes of chemicals in the training data set, which requires their determination of accurate partition coefficients, would substantially enlarge the AD of the existing PP-LFERs.

Conflicts of interest

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

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP18K05204 and JP16K16216 as well as 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.

References

1. M. H. Abraham, A. Ibrahim and A. M. Zissimos, Determination of sets of solute descriptors from chromatographic measurements, *J. Chromatogr. A*, 2004, **1037**, 29-47.
2. K.-U. Goss and R. P. Schwarzenbach, Linear free energy relationships used to evaluate equilibrium partitioning of organic compounds, *Environ. Sci. Technol.*, 2001, **35**, 1-9.
3. S. Endo and K.-U. Goss, Applications of Polyparameter Linear Free Energy Relationships in Environmental Chemistry, *Environ. Sci. Technol.*, 2014, **48**, 12477-12491.
4. C. F. Poole, T. C. Ariyasena and N. Lenca, Estimation of the environmental properties of compounds from chromatographic measurements and the solvation parameter model, *J. Chromatogr. A*, 2013, **1317**, 85-104.
5. K.-U. Goss, Predicting the equilibrium partitioning of organic compounds using just one linear solvation energy relationship (LSER), *Fluid Phase Equilib.*, 2005, **233**, 19-22.
6. T. I. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M. T. Cronin, P. Gramatica, J. S. Jaworska, S. Kahn, G. Klopman, C. A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G. Y. Patlewicz, R. Perkins, D. Roberts, T. Schultz, D. W. Stanton, J. J. van de Sandt, W. Tong, G. Veith and C. Yang, Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52, *ATLA Altern. Lab. Anim.*, 2005, **33**, 155-173.

7. J. Jaworska, N. Nikolova-Jeliazkova and T. Aldenberg, QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review, *ATLA Altern. Lab. Anim.*, 2005, **33**, 445-459.
8. P. Gramatica, Principles of QSAR models validation: Internal and external, *QSAR Comb. Sci.*, 2007, **26**, 694-701.
9. M. H. Abraham, H. S. Chadha, G. S. Whiting and R. C. Mitchell, Hydrogen bonding. 32. An analysis of water-octanol and water-alkane partitioning and the Δlog P parameter of seiler, *J. Pharma. Sci.*, 1994, **83**, 1085-1100.
10. M. H. Abraham, J. Andonian-Haftvan, G. S. Whiting, A. Leo and R. S. Taft, 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, 1777-1791.
11. A. Geisler, S. Endo and K.-U. Goss, Partitioning of Organic Chemicals to Storage Lipids: Elucidating the Dependence on Fatty Acid Composition and Temperature, *Environ. Sci. Technol.*, 2012, **46**, 9519-9524.
12. G. Bronner and K.-U. Goss, Predicting sorption of pesticides and other multifunctional organic chemicals to soil organic carbon, *Environ. Sci. Technol.*, 2011, **45**, 1313-1319.
13. S. Endo, B. I. Escher and K.-U. Goss, Capacities of Membrane Lipids to Accumulate Neutral Organic Chemicals, *Environ. Sci. Technol.*, 2011, **45**, 5912-5921.
14. S. Endo and K.-U. Goss, Serum Albumin Binding of Structurally Diverse Neutral Organic Compounds: Data and Models, *Chem. Res. Toxicol.*, 2011, **24**, 2293-2301.
15. N. Ulrich, S. Endo, T. N. Brown, N. Watanabe, G. Bronner, M. H. Abraham and K. U. Goss, UFZ-LSER database v 3.2 [Internet], 2017.
16. M. H. Abraham, A. Ibrahim and W. E. Acree, 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**, 478-485.
17. D. Mackay and J. A. Arnot, The Application of Fugacity and Activity to Simulating the Environmental Fate of Organic Contaminants, *J. Chem. Eng. Data*, 2011, **56**, 1348-1355.
18. S. Endo and K.-U. Goss, Predicting Partition Coefficients of Polyfluorinated and Organosilicon Compounds using Polyparameter Linear Free Energy Relationships (PP-LFERs), *Environ. Sci. Technol.*, 2014, **48**, 2776-2784.
19. T. H. Nguyen, K.-U. Goss and W. P. Ball, 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**, 913-924.
20. S. Endo, P. Grathwohl, S. B. Haderlein and T. C. Schmidt, LFERs for soil organic carbon-water distribution coefficients (K_{OC}) at environmentally relevant sorbate concentrations, *Environ. Sci. Technol.*, 2009, **43**, 3094-3100.
21. X.-R. Xia, N. A. Monteiro-Riviere and J. E. Riviere, An index for characterization of nanomaterials in biological systems, *Nat. Nanotechnol.*, 2010, **5**, 671-675.