QMRF Title: **IFSQSAR PPLFER logKow v2; QSPR for log K_{OW}, Brown, Armitage, Sangion, and Arnot 2024**
Contact: **Trevor N. Brown – trevor.n.brown@gmail.com**
Date: **9 March 2025**

1. QSAR identifier

1.1. QSAR identifier (title):

IFSQSAR PPLFER logKow v2; QSPR for log K_{OW}, Brown, Armitage, Sangion, and Arnot 2024

1.2. Other related models:

Versions of this QSPR:

- IFSQSAR PPLFER logKow v1
  - A curated dataset of solute descriptor was used to calibrate the QSPRs and the PPLFER equations for v1 [2].
- IFSQSAR PPLFER logKow v2 (current)
  - Some data for PFAS used to calibrate the solute descriptor QSPRs and the PPLFER equations were identified as unreliable and removed, and new reliable data for PFAS [3] were added to the v2 training and external validation datasets [1].

1.3. Software coding the model:

IFSQSAR python package, logKow v2 is included in versions 1.1.1 and up.

Model is coded in python, with openbabel used for chemistry and numpy used for mathematics.

2. General information

2.0. Abstract

Edited text from the papers [1, 2]: *This study describes the development and evaluation of six new models for predicting physical–chemical properties, including log K_{OW}, that are highly relevant for chemical hazard, exposure, and risk estimation. These models are implemented as Poly-Parameter Linear Free Energy Relationship (PPLFER) equations which combine experimentally calibrated system parameters and solute descriptors predicted with QSPRs. New experimental data from Endo 2023 [3] are used to improve QSPRs for predicting poly-parameter linear free energy relationship (PPLFER) descriptors for calculating the octanol-water (K_{OW}) partition ratios. A key PPLFER descriptor for PFAS is the molar volume term and this work compares different versions and makes recommendations for obtaining the best property predictions. The results from the new IFSQSAR models show improvements for predicting PFAS properties.*

2.1. Date of QMRF

9 March 2025

2.2. QMRF author(s) and contact details

Trevor N. Brown – trevor.n.brown@gmail.com

2.3. Date of QMRF update(s)

NA

2.4. QMRF update(s)

NA

2.5. Model developer(s) and contact details

Trevor N. Brown – ARC Arnot Research & Consulting, Toronto, ON, Canada
trevor.n.brown@gmail.com

James A. Armitage – AES Armitage Environmental Sciences, Ottawa, ON, Canada

Alessandro Sangion – ARC Arnot Research & Consulting, Toronto, ON, Canada

Jon A. Arnot – ARC Arnot Research & Consulting, Toronto, ON, Canada; Department of Physical and Environmental Science, University of Toronto, ON, Canada; Department of Pharmacology and Toxicology, University of Toronto, Toronto, ON, Canada

2.6. Date of model development and/or publication

2024

2.7. Reference(s) to main scientific papers and/or software package

[1] Brown, T.N., et al., Improved prediction of PFAS partitioning with PPLFERs and QSPRs. Environmental Science: Processes & Impacts, 2024. 26(11): p. 1986-1998.

[2] Brown, T.N., A. Sangion, and J.A. Arnot, Identifying uncertainty in physical-chemical property estimation with IFSQSAR. J Cheminform, 2024. 16(1): p. 65.

[3] Endo, S., Intermolecular Interactions, Solute Descriptors, and Partition Properties of Neutral Per- and Polyfluoroalkyl Substances (PFAS). Environmental Science & Technology, 2023. 57(45): p. 17534-17541.

[4] Goss, K.-U., Predicting the equilibrium partitioning of organic compounds using just one linear solvation energy relationship (LSER). Fluid Phase Equilibria, 2005. 233(1): p. 19-22.

[5] Brown, T.N., Predicting hexadecane-air equilibrium partition coefficients (L) using a group contribution approach constructed from high quality data. SAR and QSAR in Environmental Research, 2014. 25(1): p. 51-71.

[6] Brown, T.N., J.A. Arnot, and F. Wania, Iterative fragment selection: a group contribution approach to predicting fish biotransformation half-lives. Environmental Science & Technology, 2012. 46(15): p. 8253-60.

[7] Brown, T.N., J.M. Armitage, and J.A. Arnot, Application of an Iterative Fragment Selection (IFS) Method to Estimate Entropies of Fusion and Melting Points of Organic Chemicals. Molecular Informatics, 2019. 38(8-9): p. e1800160.

2.8. Availability of information about the model

The model is non-proprietary: the model and data are available to use for free on the online platform EAS-E Suite: https://arnotresearch.com/eas-e-suite/.

2.9. Availability of another QMRF for exactly the same model

NA

3. Defining the endpoint - OECD Principle 1: "A DEFINED ENDPOINT"

3.1. Species

NA

3.2. Endpoint

Log of wet (practical) octanol-water partition coefficient (log K_{OW}, sometimes log P) - updated for PFAS

3.3 Comment on endpoint

The endpoint is for "wet" log K_{OW}, where the octanol and water in the measurements are in direct contact and are mutually saturated, as opposed to "dry" log K_{OW} where the octanol and water are both pure phases. Recalibrated to exclude unreliable values for PFAS and include new measured data from [3].

3.4. Endpoint units

mol/$L_{[octanol]}$ / mol/$L_{[water]}$ = $L_{[water]}/L_{[octanol]}$; though frequently referred to as "unitless"

3.5. Dependent variable

$\log_{10} K_{OW}$

3.6. Experimental protocol

Examples of the types of measured data:

EC A.8 Partition Coefficient OECD 123 Partition Coefficient (n-Octanol/Water): Slow-Stirring Method (2006)

OECD 107 Partition Coefficient (n-octanol/water); Shake Flask Method (1981 & 1995)

3.7. Endpoint data quality and variability

NA

4. Defining the algorithm - OECD Principle 2: "AN UNAMBIGUOUS ALGORITHM"

4.1. Type of model

PPLFER – Poly-Parameter Linear Free Energy relationship. PPLFERs are calibrated using only experimentally determined descriptors, in contrast to QSPRs which are calibrated using theoretical descriptors. PPLFERs include solute descriptors which are different for each chemical, and system parameters which are calibrated once for each system, in this case octanol-water partitioning as log K_{OW}. When available in the IFSQSAR internal database

experimental values for solute descriptors are used, returning empirical predictions. When experimental data is not available IFSQSAR returns predictions that are a combination of empirical system parameters and QSPR predictions for the solute descriptors. Both PPLFERs and the solute descriptor QSPRs are multiple linear regression, ordinary least squares models.

4.2. Explicit algorithm

The PPLFER equation used in the model is:

$\log_{10} K_{OW} = -1.219 \cdot S - 0.058 \cdot A - 3.579 \cdot B + 2.702 \cdot Vf + 0.341 \cdot L + 0.326$

S, A, B, V, and L are the solute descriptors used in the model. The solute descriptor predictions are calculated as the linear sum of fragment counts (f) and regression coefficients (a).

S, A, B, V, or $L = a_0 + a_1 f_1 + a_2 f_2 + \ldots + a_n f_n$

4.3. Descriptors in the model

The five PPLFER solute descriptors are S, A, B, V, and L, each of which is predicted with a different QSPR when required. The descriptors for the QSPRs are all molecular fragments specified as SMARTS strings. The inputs used in the algorithm shown in 4.2 are the counts of each fragment in a chemical structure. The QSPRS for S, A, B, and L contain 345, 144, 359, 279 fragments respectively, plus the intercept (a_0) in the model for the L QSPR. The Vf descriptor is a modified version of V, also known as McGowan volume and is considered a theoretical descriptor.

4.4. Descriptor selection

For the solute descriptor QSPRs the preliminary descriptor pool is generated using custom code by recursively fragmenting the training dataset to obtain all possible molecular fragments, which typically number in the tens of thousands, a ratio of up to 50:1 vs. the number of training data. After consolidating fragment with colinear counts and applying other filters such as a cut off for correlation vs. the dependent variable the number of fragments in final descriptor pool is typically at a ratio of about 10:1 vs. the number of training data points.

4.5. Algorithm and descriptor generation

The solute descriptors in the PPLFER equations have been set based on the best available science [4]. For the solute descriptor QSPRs fragments are selected from several sub-pools of descriptors proceeding iteratively from simple fragments to complex fragments. Fragments in each sub-pool are selected by iterative forward selection and backwards removal, with replacement of fragments in the model with fragments from the sub-pool also considered in the forward selection. Fragment selection and removal are chosen by selecting or removing the fragment which improves the goodness-of-fit (GoF) the most. GoF is the Akaike Information Criteria Corrected for dataset size (AICC), which uses as input the predictive sum of squares (PRESS) calculated from a k-fold cross validation (typically k = 10).

4.6. Software name and version for descriptor generation

IFSQSAR custom development code, see 4.4 and 4.5.

4.7. Chemicals/Descriptors ratio

For the solute descriptor QSPRs the ratio of training data to selected descriptors are in the range of about 9:1 to 12:1.

5. Defining the applicability domain - OECD Principle 3: "A DEFINED DOMAIN OF APPLICABILITY"

5.1. Description of the applicability domain of the model

Three methods of assessing Applicability Domain (AD) are applied simultaneously in IFSQSAR:

1. The leverage approach [5] which is essentially a metric that quantifies the distance in the model descriptor space between a chemical the training data.
2. Chemical similarity score (CSS) [6] which is an k-nearest neighbours approach (k=5) that also incorporates how well the model fits the training data of the nearest neighbours.
3. Other structural or property alerts [7]:
    a. None of the fragments in the model are in the chemical and the prediction is just the intercept.
    b. A check for atom or bond types not found in the training dataset.
    c. Boundary condition violations, i.e. XXX.

The first two methods are applied in a complementary way to define chemicals as in AD, borderline cases which are in AD but have more uncertainty, out of AD, and cases of egregious extrapolation which are very uncertain. These designations are overridden by alerts of the third method which identify very uncertain predictions.

The uncertainty of predictions is quantified as the standard error of prediction (SEP), which along with the predicted value can be used to calculate a 95% prediction interval.

These AD methods are applied for the individual solute descriptor QSPRs, and then aggregate AD and uncertainty are calculated for the final log K_{OW} predictions based on propagation of error rules. If any of the solute descriptor QSPRs are considered to be out of the AD then the aggregate AD is also out of the AD.

5.2. Method used to assess the applicability domain

It is observed for this model and others that the uncertainty as quantified using the external validation dataset increases proceeding from: in AD, borderline, out AD, egregious extrapolation, missing atoms or bonds. This pattern of increasing validated uncertainty with increasingly dire AD warnings builds confidence that the AD methods are effective and robust. The reliability of the uncertainty metric and the various AD methods was extensively evaluated and validated vs. a novel external dataset [2]. The uncertainty metric is scaled by a factor 1.25 to ensure that 95% of the predictions contain the expected value within the 95% prediction interval.

5.3. Software name and version for applicability domain assessment

IFSQSAR custom code, see 5.1.

5.4. Limits of applicability

Boundaries for the individual solute descriptors QSPRs are shown below.

Boundaries for the leverage approach, where p' is the number of model parameters including the intercept, and n is the number of training data:

- Borderline: leverage > 0.228, 0.089, 0.214, 0.259 for S, A, B, L, defined as $1.5 \cdot p'/n$
- Out of AD: leverage > 0.437, 0.170, 0.411, 0.484 for S, A, B, L, defined as $3 \cdot p'/n$
- Egregious extrapolation: leverage > 1

Boundaries of the CSS approach where CSS is in the range 0 (no similarity) to 1 (very similar):

- Out of AD: CSS < 0.156, 0.065, 0.163, 0.126 for S, A, B, L, defined as lower CSS than 95% of the training data.
- Borderline: CSS < 0.313, 0.130, 0.325, 0.253 for S, A, B, L, defined as lower CSS than 75% of the training data.

A chemical is assigned in AD, borderline, out of AD, or egregious extrapolation based on the most conservative (poorest) result from the leverage and CSS approaches, so a chemical is only assigned in AD if both approaches assign a chemical as in AD.

Out of AD due to no fragment overlap: counts for all fragments in the model are zero.

Out of AD due to novel bonds or atoms: at least one bond or atom type not found in the training dataset.

Out of AD due to lower boundary violation: the A and B solute descriptors have lower boundaries of 0.

The external validation of the aggregate AD showed that the prediction intervals for chemicals had comparable accuracy in all cases except for the leverage cutoff for egregious extrapolation, and the alerts for novel bonds or atoms, and boundary violations. The definition of AD is "the chemical space in which predictions of a given accuracy are made", and by this definition the aggregate the AD for log K_{OW} predictions is considered to be out of only for the mentioned cases where the prediction interval is not reliable.

6. Defining goodness-of-fit and robustness (internal validation) – OECD Principle 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY"

6.1. Availability of the training set

Data are available in the database of the online platform EAS-E Suite: https://arnotresearch.com/eas-e-suite/.

6.2. Available information for the training set

a) Chemical names: No

b) CAS numbers: Yes

c) SMILES: Yes

d) InChI codes: No

e) MOL files: No

f) Structural formula: No

g) Nanomaterials: No

h) test chemical purity: No

i) Any other structural information: No

6.3. Data for each descriptor variable for the training set

No, descriptor values are not included.

6.4. Data for the dependent variable for the training set

See 6.1.

6.5. Other information about the training set

The training dataset for the solute descriptor QSPRs contains 3310 chemicals. The training dataset for the PPLFER equation contains n= 607 chemicals with log K_{OW} in the range -2.5 to 10.5.

6.6. Pre-processing of data before modelling

NA

6.7. Statistics for goodness-of-fit

For the solute descriptor QSPRs:

S: $r^2 = 0.964$, RMSE = 0.144

A: $r^2 = 0.891$, RMSE = 0.133

B: $r^2 = 0.976$, RMSE = 0.088

L: $r^2 = 0.996$, RMSE = 0.235

For the PPLFER equation: RMSE = 0.162

6.8. Robustness - Statistics obtained by leave-one-out cross-validation

NA

6.9. Robustness - Statistics obtained by leave-many-out cross-validation

NA

6.10. Robustness - Statistics obtained by Y-scrambling

Not performed for the solute descriptor QSPRs, but y-scrambling was done with 50 iterations while developing the algorithm and it was found that the average y-scrambled r^2 and RMSE were 0.086 and 1.10 compared to fitted values of 0.789 and 0.526 [6].

6.11. Robustness - Statistics obtained by bootstrap

NA

6.12. Robustness - Statistics obtained by other methods

NA

7. Defining predictivity (external validation) – OECD Principle 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY"

7.1. Availability of the external validation set

Data are available in the database of the online platform EAS-E Suite: https://arnotresearch.com/eas-e-suite/.

7.2. Available information for the external validation set

a) Chemical names: No

b) CAS numbers: Yes

c) SMILES: Yes

d) InChI codes: No

e) MOL files: No

f) Structural formula: No

g) Nanomaterials: No

h) test chemical purity: No

i) Any other structural information: No

7.3. Data for each descriptor variable for the external validation set

No, descriptor values are not included.

7.4. Data for the dependent variable for the external validation set

See 7.1.

7.5. Other information about the external validation set

The external validation dataset for the solute descriptor QSPRs contains n=1649 chemicals. The external validation dataset for log K_{OW} dataset contains 8416 chemicals.

7.6. Experimental design of test set

For the solute descriptor QSPRs, the training data is initially seeded with the chemicals that have the highest and lowest expected values. In this case all the splitting was done simultaneously for all five solute descriptors S, A, B, and L, so the highest and lowest were designated as those chemicals with the lowest and highest sums of the rank orders of their solute descriptors. Chemicals are then alternately added to the training and validation datasets until all chemicals are assigned to a dataset. When adding to the training dataset the chemical which is the least similar to chemicals already in the training dataset and most similar to chemicals already in the external validation dataset is selected. When adding to the external validation dataset the chemical which is the least similar to chemicals already in the external validation dataset and most similar to chemicals already in the training dataset is selected. More details about quantifying the similarity are available in [5], but the method is similar to the definition of CSS, see 5.1.

The log K_{OW} external validation dataset was compiled from the OPERA internal database, first removing all the chemicals that are in the solute descriptor QSPR training or validation datasets.

7.7. Predictivity - Statistics obtained by external validation

For the solute descriptor QSPRs:

S: $r^2 = 0.872$, RMSE = 0.195

A: $r^2 = 0.778$, RMSE = 0.131

B: $r^2 = 0.942$, RMSE = 0.098

L: $r^2 = 0.989$, RMSE = 0.269

For the external log K_{OW} dataset: RMSE = 1.00

7.8. Predictivity - Assessment of the external validation set

For the solute descriptor QSPRs the external validation dataset comprises 33% of the available data, and the methodology to assign chemicals to the training and external validation datasets ensures representative dependent variable and fragment count distributions in both datasets.

The log K_{OW} external validation dataset represents an entirely separate dataset and chemical space from the training dataset. The training dataset are limited to chemicals with measured PPLFER solute descriptors which only partially overlaps with the chemical space of log K_{OW} measurements. This means that the external validation with log K_{OW} values is well representative of the model's predictive power for novel, data-poor chemicals.

7.9. Comments on the external validation of the model

NA

8. Providing a mechanistic interpretation - OECD Principle 5: "A MECHANISTIC INTERPRETATION, IF POSSIBLE"

8.1. Mechanistic basis of the model

PPLFER equations are based on a mechanistic model of molecular interactions. The experimentally determined solute descriptors correlate with molecular interactions: S correlates with polarity and polarizability, A correlates with hydrogen-bond donor strength B correlates with hydrogen-bond acceptor strength, V correlates with molecular size and cavitation energy, and L correlates with van der Waals interactions. This mechanistic basis allows the results to be interpreted in terms of molecular interactions.

8.2. A priori or a posteriori mechanistic interpretation

A priori for the meaning of the solute descriptors in the PPLFER equation, a posteriori for the individual fragments of the solute descriptor QSPRs.

8.3. Other information about the mechanistic interpretation

NA

9. Miscellaneous information

9.1. Comments

The inclusion in the model outputs of a quantitative uncertainty metric is useful for assessing the reliability of the predictions and for propagating uncertainty in applications of the model predictions.

9.2. Bibliography

NA

9.3 Supporting information

NA