Henry's Law Constant – A General-Purpose Fragment Model to Predict Log K_{aw} From Molecular Structure

Ralf-Uwe Ebert,[a] Ralph Kühne,[a] Gerrit Schüürmann[a,b,*]

[a] UFZ Department of Ecological Chemistry, Helmholtz Centre for Environmental Research, Permoserstr. 15, 04318 Leipzig, Germany

[b] Institute of Organic Chemistry, Technical University Bergakademie Freiberg, Leipziger Str. 29, 09596 Freiberg, Germany

Corresponding author: Gerrit Schüürmann

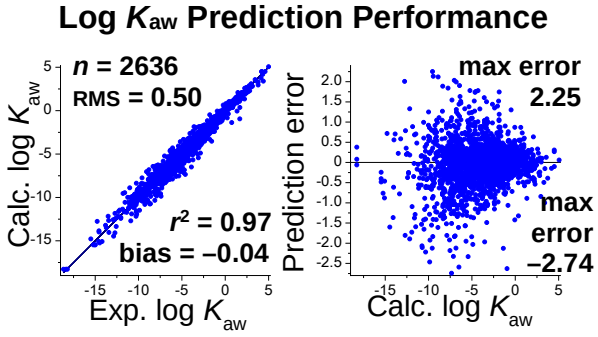
Emails gerrit.schuurmann@ufz.de, gerrit.schuurmann@chemie.tu-freiberg.de

WORD count

(w/o title page, TOC Graphic, SI information, References):

- Main text: 4816 words
- 2 tables, 2 figures

TOC Graphic


Log K_{aw} Prediction Performance
n = 2636
RMS = 0.50
r^2 = 0.97
bias = −0.04
Calc. log K_{aw}
Exp. log K_{aw}
Prediction error
max error
2.25
max
error
−2.74
Calc. log K_{aw}

ABSTRACT

Henry's law constant is important for assessing the environmental fate of organic compounds, including polar accumulation, indoor contamination and the impact of airborne predominance on persistence. Moreover, it can be used in the context of alternative 3R bioassays to inform about the compound loss through volatilization as confounding factor. For 2636 compounds, curated experimental log K_{aw} (air/water partition coefficient) data at 25° covering 23.6 orders of magnitude (from –18.6 to 5.0) have been collected from literature. Subsequently, a new fragment model for predicting log K_{aw} from molecular structure has been developed. According to the root-mean-squared error RMS and the maximum negative and positive errors MNE and MPE, this general-purpose model outperforms COSMOtherm, EPISuite HENRYWIN, OPERA and LSER with calculated input parameters significantly (RMS 0.50 vs 0.92 vs 1.25 vs 1.28 vs 1.38, MNE –2.74 vs –6.78 vs –9.11 vs –6.24 vs –6.27, MPE 2.25 vs 6.22 vs 8.27 vs 11.5 vs 7.69 log units). Initial separation into a training and prediction set (80%:20%), mutual leave-50%-out validation and target value scrambling (temporarily wrong compound-K_{aw} allocations) demonstrate the prediction capability, statistical robustness and mechanistically sound basis of the fragment scheme. The new model is available to the public in fully computerized form through the ChemProp software, and can be combined with a separate existing model to extend the log K_{aw} prediction to temperatures different from 25° C.

Key Words:

Henry's law constant, air/water partition coefficient, QSAR, fragment model, ChemProp

SYNOPSIS

Henry's law constant is important for assessing the environmental and bioassay fate of organic compounds. This study introduces a new respective fragment model, outperforming existing alternatives significantly.

INTRODUCTION

Henry's law constant *H* is one of the key physicochemical parameters governing the environmental fate of organic compounds.[1] Examples include the global distribution, deposition and bioaccumulation in polar regions,[2-4] the screening for environmental persistence in line with the precautionary principle,[5] profiling volatility-dependent differences in environmental compound partitioning between water-rich and water-poor regions,[6] and indoor contamination by xenobiotic vapors.[7-9]

Besides environmental fate assessment and modelling, *H* or its dimensionless form as air/water partition coefficient

$$K_{\mathrm{aw}} = \frac{H}{RT} \qquad (1)$$

(*R* gas constant, *T* temperature) are required for controlling the compound loss through volatilization from in vitro bioassays,[10-12] and also serve as link between gas-phase and solution-phase reaction thermodynamics.[13]

The current gold standard of fragment models for predicting log K_{aw} is the EPI-Suite HENRYWIN program from Meylan and Howard,[14] including significant extensions to their original bond contribution model published in 1991.[15] Following the description of the latest EPISuite version from 2015,[14] this widely used method covers 245 bond fragments and 43 correction factors. Of these 298 model parameters, 62 bond fragments and 33 correction factors had been calibrated through a training set of 442

compounds (squared correlation coefficient r^2 = 0.977, root-mean-squared error RMS = 0.40; experimental log K_{aw} from –11.6 to 2.9, molecular weight from 26 to 451.5 D), whereas the additional 203 parameters had been defined subsequently with a final performance check for further 1376 compounds (r^2 = 0.79, RMS = 1.54).

For the present study, we have extended the data base to 2636 organic compounds with curated experimental values, going significantly beyond the HENRYWIN database and earlier compilations.[14,16-18] The newly developed UFZ fragment scheme for predicting log K_{aw} from chemical structure comprises 209 parameters, outperforming HENRYWIN,[14] COSMOtherm,[19] OPERA,[20] and LSER[21,22] (linear solvation energy relationship) employing Platts-calculated[23] input parameters significantly; it is available to the public through our ChemProp software.[24]

MATERIAL AND METHODS

Dataset. For 2636 organic compounds, experimental H or K_{aw} at 25°C have been collected and curated from originally 2647 raw data. Molecular weight MW ranges from 16 to 959 D, and experimental log K_{aw} from –18.6 to 5.0.

Data curation has been carried out through individual expert judgement, proceeding along the following considerations: The most important criterion was qualitative consistency among structurally related compounds as expected from physical organic chemistry (e.g. homolog series, effect of halogenation and of aromatic substitution, alkane vs alkene vs alkyne, branched vs linear). In case of doubts or conflicting multiple data, COSMOtherm[19] has been used to cross-check expectation regarding relative (linear or non-linear) trends among compounds or compound series. Besides considering previous datasets,[14,16-18] original sources have been consulted for ca. 1000 compounds. In case of multiple data for single compounds, the most

reasonable individual value according to these criteria has been taken. This approach applied also to *H* determined indirectly from other experimental properties (see eqs. 2 and 3 below). For very low K_{aw}, the associated data quality is expected to be lower, which has been tolerated to some extent (except removing unplausible cases) in order to increase the chemical domain as much as reasonably possible.

Besides 1234 directly measured values, *H* was obtained for 1402 solutes from experimental vapor pressure P_v and water solubility S_w through

$$H \approx \frac{P_v}{S_w} \quad (2)$$

as approximate relationship.

Instead of S_w, the mole fraction solubility x_w^s defined through the activity coefficient at saturation in aqueous solution, γ_w^s,

$$x_w^s = \frac{1}{\gamma_w^s} \quad (3)$$

can be used (where $\gamma_w \rightarrow 1$ when the solution becomes the pure solute according to Raoult's law convention). With M_w = 55.56 mol/L as molarity of water and the common replacement of γ_w^s by the infinite-dilution activity coefficient γ_w^∞ (that is similar to γ_w^s for low-soluble compounds), experimental P_v and γ_w^∞ yield

$$H \approx \frac{P_v \cdot \gamma_w^\infty}{M_w} \quad (4)$$

that has been applied for five compounds. In eq. 4, it is assumed that the volume of the aqueous solution is 1 L as of pure water.

Finally, K_{aw} was obtained from the experimental K_{ow} (octanol/water partition coefficient) and K_{oa} (octanol/air partition coefficient) through

$$K_{aw} \approx \frac{K_{ow}}{K_{oa}} \quad (5)$$

for 46 compounds. Eq. 5 neglects that K_{ow} refers to octanol and water saturated by each other, whereas the water phase of K_{aw} contains the solute as only non-water chemical.

As will be shown below, the model performance is slightly inferior for the subset of 1402 compounds with indirectly measured Henry's law constants. Nevertheless, inclusion of the latter expands the chemical domain significantly, outweighing the fact that their error tends to be a bit larger than from direct measurement. Typically, compilations of experimental Henry's law constants include also indirectly measured *H*.

Model Development. Initially, the total compound set was subdivided into a training set and a prediction set, comprising 80% and 20% of the 2636 compounds, respectively. To this end, the following stratified selection procedure was used: All compounds were allocated to structural classes with increasing complexity (see below), and within each class ordered by increasing MW. Subsequently, from each structural class 20% of the compounds were randomly selected for the prediction set, thus leading to 527 prediction set and 2109 training set compounds, respectively. The stratified compound ordering regarding structural class and MW ensures that both training and prediction set obtain similar portions of more complex compounds (and in fact of all compound types), enabling to seriously test the prediction performance of the training-set-derived model.

For the new model, we developed the following modular fragmentation scheme, designed to efficiently limit the number of fragment parameters and thus to avoid overfitting. First, basis fragments were identified that represent atom types in a specific hybridization (e.g. sp^3 carbon, sp^2 nitrogen) and simple functional groups (e.g. –OH, $–NH_2$; see SI). If a given compound cannot be decomposed completely into basis fragments of the UFZ model, it is outside the model domain. Second, adjacent basis frag-

ments resulting in specific structural situations (e.g. π-electron conjugation) and composite functional groups (e.g. –COOH) are addressed by associated correction factors (see SI) in those cases where the basis fragment increments alone would lead to significant prediction errors. In this way, correction factors necessarily refer to substructures containing more than one basis fragment. For example, –COOH is built from the basis fragments C=O (fragment #17, see SI) and –OH (#14), augmented by correction factor #127 to account for the adjacent interaction C(=O)–OH. Multiple occurrences of fragments or corrections factors are addressed additively. If chemically different basis fragments or corrections factors resulted in sufficiently similar increment values, they were grouped together and thus allocated to joint increment values (see SI).

Development of this fragmentation scheme was confined to the training set through manual selection of substructural features as model parameters and respective stepwise multilinear regressions. Overall, this comprised the following five major steps in terms of sequentially trained subsets: (i) Hydrocarbons and simple monofunctional derivatives, including calibration of the model regression constant; (ii) monofunctional compounds with complex functional groups; (iii) compounds with multiple occurrence(s) of one type of functional group; (iv) compounds with two types of functional groups; (v) compounds with more than two types of functional groups.

In this way of sequential subgroup-specific calibration in the order of increasing structural complexity, increments of simple(r) functional groups are not contaminated by additional intramolecular effects and/or by additional solute-water interactions driven by more complex functional groups. Accordingly, the resultant total model error is slightly larger than with global (as opposed to sequential) least-squared error minimization. This, however, is more than outweighed by mechanistically sound (contamination-free) fragment values that are likely superior for true predictive applications.

Upon completion of the basis fragment and correction factor identification and calibration through using the training set, the model was applied to the untrained prediction set to inform about its prediction capability. Subsequently, only the increments of all basis fragments and correction factors were re-calibrated in the same sequential order as described above for the total set of 2636 compounds, leading to the final model (see statistics for n = 2636 in Table 1 below, and the SI for all final model parameters).

Overall, the sequential calibration yielded 38 basis fragments associated with additive increment values, 170 correction factors accounting for substructural features containing already defined fragments, and a regression constant (see the SI for details). As such, the resultant increment model can be termed an additive-constitutive calculation scheme.

Application domain. The structural applicability of the fragmentation scheme is addressed through our atom-centered fragment (ACF) approach[25] as implemented in the ChemProp software.[24] For a given target compound, its first-order ACFs (unique atom-centered fragment confined to one non-hydrogen atom and its first bonding neighbors)[25] and second-order ACFs (confined to first and second neighbors along each bonding direction)[25] are checked for respective occurrences in the total set of 2636 compounds. Within each ACF order, two ACFs are considered different in case of differences regarding atom type, exact number of attached hydrogen atoms, aromaticity (yes or no), ring atom (yes or no), total number of bonded neighbors including H atoms, bond type (nonaromatic: single, double, triple; aromatic), and ring closure (either inside the ACF or including up to second-order neighbors outside the ACF).

Mutual Leave-50%-Out Validation. Originally termed simulated external validation,[26] this approach can be viewed as an extended version of a leave-50%-out

cross-validation, focusing on 50%-subset-specific training and prediction performances and their differences rather than on respective statistical averages. More specifically, two 50%-subsets are used for separate calibrations and mutual predictive applications, respectively. As such, it informs about the model robustness and prediction capability when reducing the training set significantly.

The two complementary subsets were generated as follows: First, starting from the top entry of the stratified compound ordering as described above, every 2nd compound was allocated to a 50%-subgroup (*group2*, 1318 compounds), with the remaining 1318 compounds forming a complementary 50%-subgroup (*group1*). Second, the fragment model parameters were re-calibrated separately for *group1* and *group2* in the above-described 5-step sequential manner. Third, the resultant subgroup-specific regression models were used for quasi-external predictions of log K_{aw} of the complementary subgroup compounds.

This way of constructing *group1* and *group2* assures that the associated chemical domains are (almost) as similar as possible. Note, however, that the derivation of two subgroup-specific fragment schemes with newly identified basis fragments and correction factors would not have been possible adequately, because restriction to 50% of the compounds would imply to reduce the number of fragment model parameters correspondingly.

Permutation Test. Target value scrambling informs about whether a given model is overfitted and thus memorizes individual cases rather than mapping a mechanistically sound relationship between model parameters and target value. To this end, the target value (here: log K_{aw}) is allocated randomly to compounds while keeping their correct model parameter values. In our variant with systematically varying the degree of permutation,[27] the degree of target value scrambling ranges from 0% (original = true relationship between compounds and their target values) to 100% permutation (all

compounds with wrong target values but correct model parameter values) in steps of 10%.

Target value scrambling can be combined with cross-validation such as leave-10%-out (model calibrated with 90% applied to randomly selected 10% of the compounds), expecting an increasing difference between calibration and cross-validation performance with increasing degree of permutation.[27]

Performance Statistics. The conventional calibration r^2 (squared correlation coefficient) quantifies the goodness of fit. It ranges from 0 to 1, and automatically corrects for systematic errors. By contrast, the prediction performance is quantified by the predictive squared correlation coefficient q^2.[28] Here, untrained experimental data are confronted with the original model output without post-model scaling (through Y = a • [model output] + b as built in r^2), with q^2 ranging from –∞ (completely useless model) over 0 (model output as good as taking the experimental mean as predictor throughout) to 1 (perfect model).[28] In particular, differences between q^2 and r^2 inform about the extent of bias associated with a given model when used for external prediction.

The root-mean-squared error RMS provides a quantification of the scatter, and the bias represents the systematic error obtained through adding up all individual prediction errors (including their signs). Note further that full least-squares error regression yields q^2 identical to r^2 when applied for the same set, and in this case q^2 does not add information beyond r^2.

RESULTS AND DISCUSSION

Global Model Performance. For the total compound set of 2636 compounds, the UFZ model statistics (RMS 0.499, MNE –2.74, MPE 2.25) is significantly superior to the ones of COSMOtherm[19] (RMS 0.915, MNE –6.78, MPE 6.22), HENRYWIN[14] (RMS

1.25, MNE –9.11, MPE 8.27), OPERA[20] (not applicable to inorganics: n=2602, RMS 1.28, MNE –6.24, MPE 11.5), and LSER[21,22] with Platts-calculated[23] input parameters. Multi-linear regression of the latter (obtainable for 2587 compounds) yields RMS 1.38, MNE –6.27, MPE 7.69 (LSER parameters S, A, B, V, L) and RMS 1.43, MNE –5.74, MPE 7.61 (LSER parameters E, S, A, B, V), respectively.

On the one hand, the new UFZ model had been developed with 80% of the present dataset, whereas the respective dataset fractions used for training OPERA and LSER are unknown (COSMOtherm has not been trained for predicting log K_{aw}). On the other hand, the present dataset exceeds previous collections significantly regarding both the number of compounds and the chemical domain, and thus represents a true challenge for the prediction performance of any model meeting a reasonable compound-to-parameter ratio. More details regarding COSMOtherm and LSER predictions including analyses for the structurally simpler and much smaller subsets with experimental LSER parameters will be reported elsewhere.

Table 1 compares the performances of the UFZ and HENRYWIN fragment schemes in more detail. For all 2636 compounds, the UFZ model RMS is only 40% of the HENRYWIN counterpart, which holds similarly for the various subsets analyzed. Moreover, largest outliers reduce from HENRYWIN –9.11 (fipronil) to UFZ model –2.74 (brofluthrinate) and from HENRYWIN 8.27 (hexamethylenetetramine) to UFZ model 2.25 (sedaxane), respectively.

As indicated above, the 1402 compounds with indirectly measured K_{aw} yield a larger RMS than the 1234 directly measured K_{aw} data (UFZ: RMS 0.60 vs 0.35; HENRYWIN: RMS 1.49 vs 0.88), but without unusually large UFZ model outliers and a negligible UFZ model bias (–0.05). Regarding physical condition, the predictions are best for liquids (UFZ model RMS 0.41) followed by gases (0.47) and solids (0.58). The latter might also be affected by a possibly lower experimental accuracy with solids (direct

measurement: nominal vs actual S_w possibly affecting c_w and thus K_{aw}; indirect measurement: low P_v, and again nominal vs actual S_w that both affect P_v/S_w).

Table 1. Performance Statistics of the UFZ and HENRYWIN[14] Models for Predicting log K_{aw} From Molecular Structure.[a]

Model and data set	*n*	r^2	q^2	RMS	bias	MNE	MPE
UFZ							
Training set	2109	0.973	0.973	0.503	-0.034	-2.79	2.11
Prediction set	527	0.972	0.971	0.525	-0.006	-2.29	2.26
Total set	2636	0.974	0.974	0.499	-0.040	-2.74	2.25
Direct data	1234	0.978	0.977	0.354	-0.029	-2.63	2.03
Indirect data	1402	0.969	0.968	0.599	-0.050	-2.74	2.25
Solid	1279	0.963	0.963	0.578	-0.052	-2.46	2.25
Liquid	1272	0.971	0.970	0.409	-0.026	-2.74	1.78
Gas	85	0.930	0.928	0.466	-0.073	-2.63	0.87
HENRYWIN[b]							
Training set	2106	0.851	0.828	1.271	-0.176	-9.11	8.27
Prediction set	527	0.870	0.863	1.141	-0.052	-5.87	7.82
Total set	2633	0.855	0.835	1.250	-0.151	-9.11	8.27
Direct data	1231	0.872	0.859	0.883	-0.028	-6.70	6.76
Indirect data	1402	0.829	0.804	1.490	-0.259	-9.11	8.27
Solid	1279	0.763	0.727	1.570	-0.291	-9.11	8.27
Liquid	1272	0.894	0.881	0.809	-0.022	-4.45	7.98
Gas	82	0.533	0.499	1.220	0.044	-5.50	1.70

[a] The statistical parameters are: *n* = number of compounds, r^2 = squared correlation coefficient, q^2 = predictive squared correlation coefficient,[28] RMS = root-mean-squared error, bias = systematic prediction error, MNE = maximum negative error (maximum underestimation), MPE = maximum positive error (maximum overestimation). Direct vs direct data: directly vs indirectly measured (see text).

[b] The statistics refer to the application of HENRYWIN to the UFZ datasets, partly with reduced numbers of compounds due to respective restrictions of the HENRYWIN applicability.

Similar trends are observed for HENRYWIN, but with lower statistical performances (Table 1). For the total set, however, the HENRYWIN bias is pleasingly small (–0.15) despite large individual outliers. This shows that the main current challenge for

HENRYWIN are compounds with structural complexity beyond its original data set, thus leading to possibly large individual prediction errors without introducing a significant global bias.

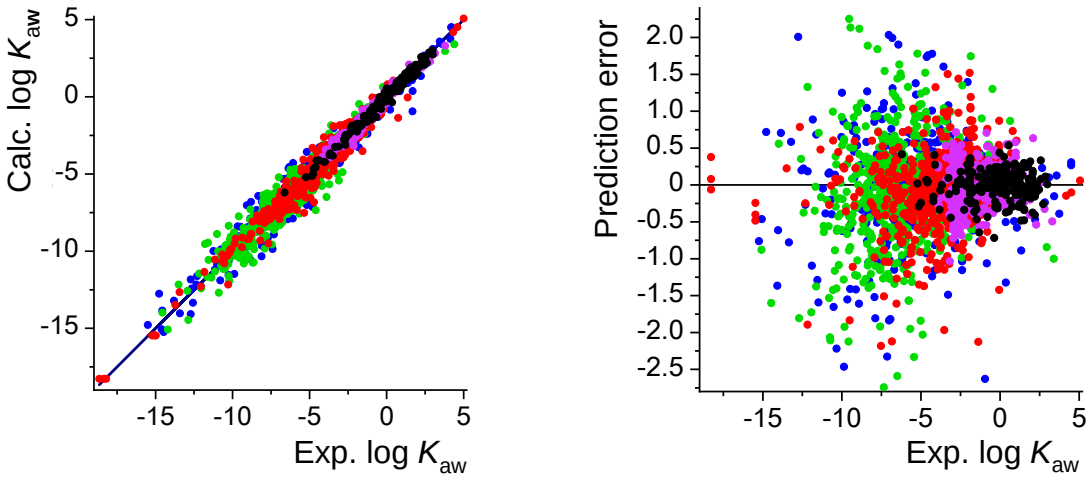


Figure 1. Calculated vs experimental log K_{aw} (left) and prediction error vs. experimental log K_{aw} (right) for the UFZ model applied on all 2636 compounds (see Table 1 for the statistics). The color-coded filled circles denote the following atomic composition of the compounds, where in each class beyond hydrocarbons heteroatom addition refers to at least one of the lower-class subsets. **Black** = hydrocarbon (HC, 265 compounds); **magenta** = halogenated HC (Hal-HC, 430); **red** = HC and Hal-HC augmented by O (867); **green** = HC, Hal-HC, O-HC and Hal-O-HC augmented by N (726); **blue** = all previous groups augmented by S and/or P and/or any other type of heteroatom (348).

Figure 1 shows the plot of predicted vs experimental log K_{aw} (left) and prediction error vs predicted value (right), color-coded according to atomic composition (heteroatoms halogen, O, N, S, P, and "other" covering the elements Si, As, Se, Hg, Sn, and Pb). Note that for analyzing a possible dependence of prediction error on target value (here: log K_{aw}), plotting prediction error vs predicted value is preferred over plotting prediction error vs experimental value. This holds because (multi)linear regression implies a correlation between prediction error and experimental value, with a squared

correlation coefficient of $1–r^2$ for r^2 as respective model calibration value.[29] Overall, Figure 1 (right) shows that the absolute prediction error increases with predicted log K_{aw} decreasing from ca. 5 to –10, and then – surprisingly – becomes smaller again.

Model Performance vs Structural Complexity. Due to the nature of fragment models, the prediction error is expected to increase with increasing molecular size and structural complexity. The reason is that except for fortuitous error compensations, small errors associated with fragment parameters typically add up with increasing numbers of fragments, and also with increasingly complex fragments due to their generally lower occurrence and accordingly lower level of model training.

Table S1 in the SI shows the impact of the number of heteroatoms (top), of functional groups (middle), and of molecular polarity (bottom; see explanation below) on the model performance. As expected, for both the UFZ model and HENRYWIN the prediction error generally increases with increasing structural complexity. More specifically, the UFZ model RMS in log K_{aw} units ranges from 0.19 (0 heteroatoms and 0 functional groups, 265 compounds) to 0.79 (≥ 9 heteroatoms, 151 compounds) and 0.97 (≥ 4 types of functional groups), with corresponding HENRYWIN RMS values increasing from 0.44 to 2.23 (≥ 9 heteroatoms) and 2.41 (≥ 4 types of functional groups), respectively.

Regarding polarity, our 4-group classification is as follows: Nonpolar and weakly polar compounds (NWP) comprise hydrocarbons and halogenated hydrocarbons (with Abraham H-bond parameters < 0.15).[30] The other three polarity groups considered are: H-bond donors (HBD that are also H-bond acceptors), H-bond acceptors (HBA) without H-bond donor capability, and silanes and other metalorganics without H-bonding. Now, the polarity-specific RMS ranges are: 0.25 (NWP, 697 compounds), 0.56 (1 HBD, 597 compounds), 0.59 (≥ 2 HBD, 146 compounds), 0.39 (1 HBA, 402 compounds, 0.48 (2 HBA, 269 compounds), 0.66 (3-5 HBA, 390 compounds), 0.78 (≥ 6

353 HBA, 119 compounds), and 0.76 (16 metalorganics), with again throughout larger
354 HENRYWIN RMS values (see Table S1).

355 Indeed, the log K_{aw} prediction error depends also on MW. For our new model,
356 the range-specific RMS errors are 0.36 (MW ≤ 200 D, 1371 compounds) and 0.62 (MW
357 > 200 D, 1255 compounds), respectively.

358 **Mutual Leave-50%-out Validation.** According to our atom-centered fragment
359 (ACF) analysis,[25] the total compound set of 2636 compounds is structurally quite div-
360 erse. It covers 1864 first-order and 7408 second-order ACFs.

361 On the one hand, this suggests a substantial chemical domain and an accord-
362 ingly broad application domain of the model. On the other hand, the two 50%-subsets
363 *group1* and *group2* (see above) contain only 1387 vs 1378 first-order and 4782 vs
364 4725 second-order ACFs, respectively. The latter implies that log K_{aw} prediction for the
365 *group2* compounds by the *group1*-calibrated model is confronted with quite some
366 *group2* ACF features outside the *group1* ACF domain, and vice versa. Since ignoring
367 chemical domain violation is clearly not recommended, we apply the mutual leave-
368 50%-out validation both without and with exclusion of compounds outside the sub-
369 group-specific (ACF-defined)[25] chemical domains.

370 Table 2 summarizes the performance of this simulated external validation. For
371 *group1* and *group2* that each contain 1318 compounds, the calibration RMS is 0.50 and
372 0.47 log K_{aw} units, respectively. Application of the *group1* model for predicting log K_{aw}
373 of all or only the ACF-domain-inside[25] *group2* compounds yields RMS values of 0.61 vs
374 0.50 (916 compounds), and when applying the *group2* model for predicting *group1*
375 log K_{aw} the corresponding RMS values are 0.62 and 0.47 (906 compounds), respect-
376 ively. Note that upon excluding outside-domain compounds, RMS comes close to its
377 training set (calibration) value. This confirms that within ACF-defined model domains
378 external prediction can be expected to be similar to the training set calibration quality.[25]

Moreover, bias, MNE and MPE remain sufficiently small (Table 2), providing further support for the statistical robustness and true prediction capability of the presently introduced log K_{aw} model.

Table 2. Mutual Leave-50%-Out Cross-Validation Statistics.[a]

Model activity and subset	*n*	r^2	q^2	RMS	bias	MNE	MPE
Calibration *group1*	1318	0.973	0.973	0.502	-0.048	-2.83	2.40
Prediction *group2*	1318		0.960	0.612	-0.021	-3.52	3.09
Only inside ACF domain	916		0.968	0.495	-0.037	-2.46	2.68
Calibration *group2*	1318	0.977	0.977	0.469	-0.031	-2.03	2.17
Prediction *group1*	1318		0.959	0.620	-0.057	-3.84	3.09
Only inside ACF domain	906		0.970	0.471	-0.021	-2.88	2.74

[a] See Table 1 for the statistical parameters. For the 50%-subsets *group1* and *group2*, the log K_{aw} fragment scheme was calibrated separately, and applied to externally predict log K_{aw} of the complementary subset (*group1* model on *group2* compounds and vice versa), without and with exclusion of prediction-set compounds outside the ACF domain[25] of the respective calibration set (see text).

Permutation Test. Target value scrambling was performed in a step-wise manner, controlling the degree of permutation from 0% (no permutation = correct compound-K_{aw} allocation throughout) to 100% (each compound randomly allocated to a K_{aw} from a different compound) in intervals of 10%. Here, q^2 is used to quantify the degree of agreement between true and (partly) permuted log K_{aw} for all 2636 compounds.

The results are shown in Table S2 (SI) and summarized in Figure 2. Calibration r^2 decreases with increasing degree of permutation (from 0.974 without scrambling to 0.010-0.030 with 100% scrambling), where for 100% permutation five different randomly (but completely) wrong compound-K_{aw} allocations have been generated with resultant q^2 (true vs permuted log K_{aw}) from –0.961 to –1.040. This result confirms that

the fragment scheme reflects a mechanistically sound relationship between molecular structure and log K_{aw}, and in particular is not affected by memorizing individual cases through overfitting.

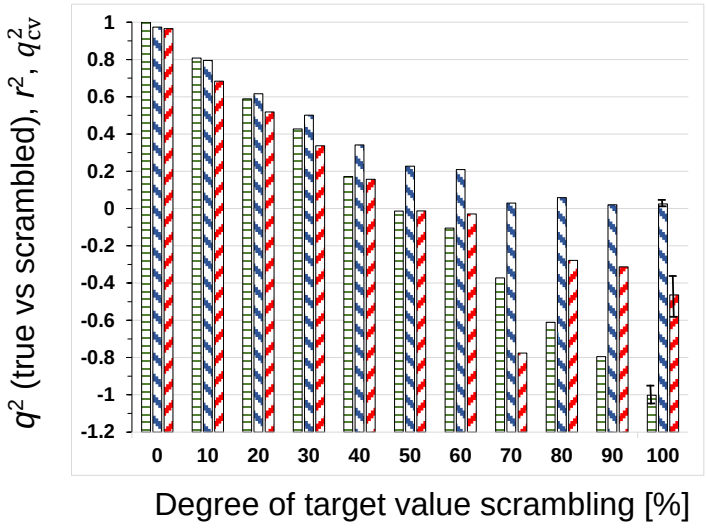


Figure 2. Permutation test statistics of the UFZ log K_{aw} model (see text and Table S2 for details). For varying degrees of target value scrambling (permutation) from 0% to 100% (x axis), the following statistical results are shown (y axis): q^2 quantifying the degree of agreement between original (true) and permuted log K_{aw} (green bars, left), regression r^2 with (not, partially and fully) permuted log K_{aw} data (blue bars, middle), and cross-validated q^2_{cv} quantifying the leave-10%-out prediction capability with (not, partially and fully) permuted log K_{aw} data (red bars, right). For 100% permutation, the five different runs performed yield results within the intervals overlaying the three respective bars that here indicate the respective medians.

Besides evaluating r^2, the permutation test was extended to check also the prediction capability resulting from models trained with scrambled log K_{aw}. To this end, leave-10%-out cross-validation was employed with the respective q^2_{cv} as measure of the prediction quality (keeping in mind different variants of q^2_{cv}).[28] As can be seen from Table S2 and its graphical summary in Figure 2, q^2_{cv} decreases from 0.966 (no permutation) to –0.667 (lowest value for 100% permutation). These results demonstrate that

the prediction capability rapidly decreases with increasing degree of log K_{aw} scrambling. Moreover, q^2_{cv} is below r^2 for all models trained with permutation, with no setting where fortuitous error compensation might have resulted in an artificially good prediction despite (partly) wrong compound-K_{aw} allocations.

EPISuite Dataset. The current EPISuite HENRYWIN dataset[14] contains 1829 compounds including 6 duplicates, 36 UVCBs (unknown or variable composition, complex reaction products or biological materials), 6 organic salts, 7 compounds representing structurally unique cases (CS_2, COS, CO_2, SF_6, SO_2F_2, triazoxide as only aromatic N-oxide, ziram as only organic Zn compound), phenylmercury dimethyldithiocarbamate with unresolved covalent vs ionic bonding), 1 structure without experimental value, and 170 compounds with no alternative literature data instead of the experimental values not meeting our quality criteria. Removal of these 227 compounds yields a curated subset of 1602 EPISuite compounds with a structural variation significantly below the one of the UFZ dataset (1417 vs 1864 first-order and 4948 vs 7408 second-order ACFs). Now, application of HENRYWIN (n=1601, one compound not calculable) results in r^2 0.856, RMS 1.185, bias –0.117, MNE –8.42, MPE 8.50 when using the UFZ-curated experimental data.

Using 80% of these compounds as training set (n=1282, selected according to our approach as described above), the UFZ fragment and correction factor calibration was possible for 197 model parameters, yielding r^2 0.977, RMS 0.439, bias –0.012, MNE –2.41, MPE 2.25, and a 20% prediction set (n=320) performance of r^2 0.968, RMS 0.520, bias –0.010, MNE –3.11, MPE 2.76, respectively. Finally, recalibration with all 1602 compounds gave r^2 0.976, RMS 0.449, bias –0.018, MNE –3.11, MPE 2.32 (see the SI for more details). These results illustrate the general modelling capability associated with the presently introduced fragmentation approach.

453 **Outlier Example.** The SI presents manual applications of the UFZ model with
454 three examples, and in this way provides further insight into the mechanistic basis of
455 the fragmentation scheme (with all model details being documented in the SI). 1,4-
456 Benzoquinone (**4** in Scheme S1) belongs to the outliers of the model. This substance
457 is both redox-active and electrophilic as Michael acceptor with respective toxicological
458 concern, and applied in industrial syntheses as Diels-Alder dienophile and as oxidant.
459 As compared to its indirect experimental log K_{aw} value of –5.243, the UFZ model
460 prediction of –6.839 yields an error of –1.596 that is outside ±3 • RMS (= 1.497 for all
461 2636 compounds, see Table 1).

462 At present, the UFZ model does not contain any parameter specific for the ben-
463 zoquinone electronic structure. The latter comprises two Michael acceptors bonded to
464 each other in a ring with strongly polarized but non-aromatic π-bonded carbons, with
465 four resonance opportunities of the type –HC=C(H)–(R)C=O ↔ –HC$^+$CH=C(R)–O$^-$**.** In-
466 stead, data from naphtho- and anthraquinones enabled derivation of a respective cor-
467 rection factor that is currently used generically for all quinones. Accordingly, the
468 log K_{aw} calculation of **4** invokes the model fragments 4 x C= (fragment #3, 4 • –0.547
469 = –2.188; see SI), 4 x H attached to sp^2-C (#9, 4 • 0.405 = 1.62), and 2 x O=C (#17, 2
470 • –3.995 = –7.99), augmented by the two correction factors 4 x cyclic C (#76, 4 •
471 –0.064 = –0.256) and quinone (#82, 1.52) besides the regression constant.

472 Since the experimental value was derived as P_V/S_W (eq. 2) with P_V = 3.9 Pa and
473 S_W = 0.275 mol/L, we checked the latter two values against separate model predictions.
474 For the solid state of **4**, EPISuite MPBPWIN[31] and WSKOWIN[32] calculate P_V = 3.42 Pa
475 and S_W = 0.689 mol/L. Noting that 1 atm = 101325 Pa, eq. 2 yields $H \approx P_V/S_W$ =
476 (3.42/[101325 • 0.689]) atm • L/mol = 4.90 • 10^{-5} atm • L/mol, and division through RT
477 = 24.465 atm • L/mol at 25°C leads to log K_{aw} = –5.70 (see eq. 1). This value is lower
478 than the indirect experimental value by 0.46 log units, and thus a bit closer to – but still

479 significantly different from – our UFZ model result. Interestingly, the HENRYWIN result
480 for **4** is –7.302 and thus below the value from separate EPISuite P_v and S_w predictions
481 by 1.602 log units (and underestimates experimental log K_{aw} by 2.059).
482 COSMOtherm[19] based on quantum chemistry and statistical thermodynamics
483 (with an only small number of parameters) provides subcooled-liquid predictions for **4**
484 of P_v = 107.6 Pa and S_w = 1.61 mol/L, leading to $H \approx 0.659 \bullet 10^{-3}$ atm • L/mol that finally
485 gives log K_{aw} = –4.57. In this case, predicted P_v and S_w result in a log K_{aw} larger by
486 0.67 than the indirect experimental value. Note further that the COSMOtherm-predict-
487 ed log K_{ow} of 0.25 comes close to the experimental log K_{ow} of 0.20, suggesting that the
488 electronic structure of **4** is at least in principle accounted for properly. With EPISuite
489 KOWWIN,[33] the log K_{ow} of 0.22 for **4** agrees almost perfectly with experiment, but this
490 provides no information about how well **4** is covered by other EPISuite fragment mo-
491 dels for other properties (such as log K_{aw}).
492 Thus, K_{aw} obtained from predicted P_v and S_w yields inconclusive results for **4**.
493 Because data for further 1,4-benzoquinones are not available as basis for calibrating
494 a respective correction factor, **4** remains an outlier of the UFZ model at this point in
495 time, which holds correspondingly for another 57 outliers listed in the SI.
496 Overall, the newly developed fragment scheme for predicting log K_{aw} at 25°C
497 from molecular structure is based on a substantial chemical domain as indicated
498 through the ACF analysis.[25] Considering its competitive performance statistics also as
499 compared to EPISuite HENRYWIN,[14] COSMOtherm,[19] OPERA,[20] and LSER[21,22] (with
500 Platts-calculated[23] input parameters), the new model may serve as general-purpose
501 tool for providing – in a fully computerized manner available to the public[24] – Henry's
502 law constant in case of missing experimental data.
503 As indicated above, there are currently 58 outliers with prediction errors ≥ ±3 •
504 RMS but still within the log interval [–2.74, 2.26]. Their improved treatment would require

505 further experimental data from sufficiently similar structures. For the time being, simi-
506 larity checks with these outliers may help to identify model predictions with lower levels
507 of confidence that would go beyond the ChemProp routine check regarding the ACF-
508 defined chemical domain.[25]

509 Correction factors encode the impact of bonding across or interaction between
510 polar groups on log K_{aw}. As opposed to the volatility-lowering impact of heteroatom-
511 associated local atomic charge (e. g. increments of the fragments OH, O, SO_2, NH/NH_2
512 as presented above), most of the correction factors (160 vs 10) increase log K_{aw} pos-
513 sibly for one of the two following reasons: First, intramolecular polar interactions are
514 favored over respective solute-solvent interactions if geometrically feasible. This is
515 illustrated by the positive log K_{aw} contribution of intramolecular H bonds at the cost of
516 solute-water H bonding. Second, intramolecular electron delocalization tends to reduce
517 local atomic charges and thus their disposition for polar interactions with water.

518 In case of interest in Henry's law constants at temperatures different from 25°C,
519 the current scheme can be combined with a separate model – again programmed for
520 fully automatized use in ChemProp[24] – addressing the respective temperature varia-
521 tion.[27]

522

523

524 **ASSOCIATED CONTENT**

525 **Supporting Information.** Impact of structural complexity on model performance
526 statistics, permutation test statistics, UFZ and HENRYWIN model performances with
527 EPISuite HENRYWIN dataset, data set of UFZ model, UFZ model outliers, and
528 fragments and correction factors of the UFZ model.

529

REFERENCES

(1) Mackay, D.; Celsie, A. K. D.; Parnis, J. M. The evolution and future of environmental partition coefficients. *Environ. Rev.* **2016**, *24* (1), 101-113.

(2) Wania, F.; Mackay, D. Tracking the Distribution of Persistent Organic Pollutants. *Environ. Sci. Technol.* **1996**, *30* (9), 390A-396A.

(3) Casal, P.; Casas, G.; Vila-Costa, M.; Cabrerizo, A.; Pizarro, M.; Jiménuz, B.; Dachs, J. Snow Amplification of Persistent Organic Pollutants at Coastal Antarctica. *Environ. Sci. Technol.* **2019**, *53* (15), 8872-8882.

(4) Chen, Y.; Lei, Y. D.; Wensvoort, J.; Gourlie, S.; Wania, F. Probing the Thermodynamics of Biomagnification in Zoo-Housed Polar Bears by Equilibrium Sampling of Dietary and Fecal Samples. *Environ. Sci. Technol.* **2022**, *56* (13), 9497-9504.

(5) Gouin, T.; Mackay, D.; Webster, E.; Wania, F. Screening Chemicals for Persistence in the Environment. *Environ. Sci. Technol.* **2000**, *34* (5), 881-884.

(6) Breitkopf, C.; Kühne, R.; Schüürmann, G. Multimedia Level-III Partitioning and residence times of xenobiotics in water-rich and water-poor environments. *Environ. Toxicol. Chem.* **2000**, *19* (5), 1430-1440.

(7) Li, L; Arnot, J. A.; Wania, F. How are Humans Exposed to Organic Chemicals Released to Indoor Air? *Environ. Sci. Technol.* **2019**, *53* (19), 11276-11284.

(8) Schwartz-Narbonne, H.; Abbatt, J. P. D.; DeCarlo, P. F.; Farmer, D. K.; Mattila, J. M.; Wang, C.; Donaldson, D. J.; Siegel, J. A. Modeling the Removal of Water-Soluble Trace Gases from Indoor Air via Air Conditioner Condensate. *Environ. Sci. Technol.* **2021**, *56* (16), 10987-10993.

(9) Wu, S.; Hayati, S. K.; Kim, E.; de la Mata, P.; Harynuk, J. J.; Wang, C.; Zhao, R. Henry*s Law Constant and Indoor Partitioning of Microbial Volatile Organic Compounds. *Environ. Sci. Technol.* **2022**, *56* (11) 7143-7152.

(10) Schramm, F.; Müller, A.; Hammer, H.; Paschke, A.; Schüürmann, G. Epoxide and thiirane toxicity in vitro with the ciliates *Tetrahymena pyriformis*: Structural Alerts Indicating Excess Toxicity. *Environ. Sci. Technol.* **2011,** *45* (13), 5812-5819.

(11) Knöbel, M.; Busser, F. J. M.; Rico-Rico, Á.; Kramer, N. I.; Hermens, J. L. M.; Hafner, C.; Tanneberger, K.; Schirmer, K.; Scholz, S. Predicting adult fish acute lethality with the zebrafish embryo: relevance of test duration, endpoints, compound properties, and exposure concentration analysis. *Environ. Sci. Technol.* **2012,** *46* (17), 9690-9700.

(12) Tanneberger, K.; Knöbel, M.; Busser, F. J. M.; Sinnige, T. L.; Hermens, J. L. M.; Schirmer, K. Predicting fish acute toxicity using a fish gill cell line-based toxicity assay. *Environ. Sci. Technol.* **2013,** *47* (2), 1110-1119.

(13) Trogolo, D.; Arey, J. S. Equilibria and Speciation of Chloramines, Bromamines, and Bromochloramines in Water. *Environ. Sci. Technol.* **2017**, *51* (1) 128-140.

(14) US EPA **2015**. Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11, module HENRYWIN v. 3.21. United States Environmental Protection Agency, Washington, DC, USA.

(15) Meylan, W. M.; Howard, P. H. Bond contribution method for estimating Henry's law constants. *Environ. Toxicol. Chem.* **1991**, *10* (10), 1283-1293.

(16) Mackay, D.; Shiu, W. Y. A Critical Review of Henry's Law Constants for Chemicals of Environmental Interest. *J. Phys. Chem. Ref. Data* **1981**, *10* (4) 1175-1199.

(17) Staudinger, J.; Roberts, P. V. A critical compilation of Henry's law constant temperature dependence relations for organic compounds in dilute aqueous solutions. *Chemosphere* **2001**, *44* (4), 561-576.

(18) Sander, R. Compilation of Henry's law constants (version 4.0) for water as solvent. *Atmos. Chem. Phys.* **2015**, *15* (8), 4399-4981, augmented by a corrigendum.

(19) COSMOlogic GmbH Co. KG, a Dassault Systèmes company 2019. COSMOthermX, version 19.0.4, http://www.cosmologic.de.

(20) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *J. Cheminform.* **2018**, 10, 10.

(21) Abraham, M. H. Scales of Solute Hydrogen-bonding: Their Construction and Application to Physicochemical and Biochemical Processes. *Chem. Soc. Rev.* **1993**, *22* (2), 73-83.

(22) Abraham, M. H., Gola, J. M. R.; Cometto-Muniz, J. E.; Cain, W. S. The solvation properties of nitic oxide. *J. Chem. Soc. Perkin Trans. 2*, **2000**, (*10*), 2067-2070.

(23) Platts, J. A., Butina, D.; Abraham, M. H., Hersey, A. Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 835-845.

(24) UFZ Department of Ecological Chemistry **2021**. ChemProp 7.1.1. http://www.ufz.de/ecochem/chemprop.

(25) Kühne, R.; Ebert, R.-U.; Schüürmann, G. Chemical Domain of QSAR Models from Atom-Centered Fragments. *J. Chem. Inf. Model.* **2009**, *49* (12), 2660-2669.

(26) Boháč, M.; Loeprecht, B.; Damborský, J.; Schüürmann, G. Impact of orthogonal signal correction (OSC) on the predictive ability of CoMFA models for the ciliate toxicity of nitrobenzenes. *Quant. Struct.-Act. Relat.* **2002**, *21* (1), 3-11.

(27) Kühne, R.; Ebert, R.-U.; Schüürmann, G. Prediction of the temperature dependence of Henry's law constant from chemical structure. *Environ. Sci. Technol.* **2005**, *39* (17), 6705-6711.

(28) Schüürmann, G.; Ebert, R.-U.; Chen, J.; Wang, B.; Kühne, R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient – Test Set Activity Mean vs Training Set Activity Mean. *J. Chem. Inf. Model.* **2008**, *48* (11), 2140-2145.

(29) Fox, J. Regression Diagnostics. Sage Publications, New Bury Park, CA (USA), **1991**, 92 pp.

(30) ACD **2020**. Percepta, Labs Release 2020.1.2. http://www.acdlabs.com

(31) US EPA **2015**. Estimation Programs Interface Suite™ for Microsoft® Windows, v. 4.11, module MPBPWIN v. 1.44. United States Environmental Protection Agency, Washington, DC, USA.

(32) US EPA **2015**. Estimation Programs Interface Suite™ for Microsoft® Windows, v. 4.11, module WSKOWWIN v. 1.43. United States Environmental Protection Agency, Washington, DC, USA.

(33) US EPA **2015**. Estimation Programs Interface Suite™ for Microsoft® Windows, v. 4.11, module KOWWIN v. 1.69. United States Environmental Protection Agency, Washington, DC, USA.