Prediction of blood:air and fat:air partition coefficients of volatile organic compounds for the interpretation of data in breath gas analysis*

Journal of Breath Research

PAPER

Prediction of blood:air and fat:air partition coefficients of volatile organic compounds for the interpretation of data in breath gas analysis[6]

Christian Kramer[1,2], Paweł Mochalski[3], Karl Unterkofler[3,4], Agapios Agapiou[5], Veronika Ruzsanyi[3] and Klaus R Liedl[1]

[1] Institute of General, Inorganic and Theoretical Chemistry, University of Innsbruck, Innrain 82, A-6020 Innsbruck, Austria
[2] Present address: Fa. Hoffmann-La Roche Ltd, Pharma Research and Early Development, Therapeutic Modalities, Roche Innovation Center Basel, Grenzacherstrasse 124, CH-4070 Basel, Switzerland
[3] Breath Research Institute, Leopold-Franzens University Innsbruck, Rathausplatz 4, A-6850 Dornbirn, Austria
[4] Vorarlberg University of Applied Sciences, Hochschulstr. 1, A-6850 Dornbirn, Austria
[5] Department of Chemistry, University of Cyprus, P.O. Box 20537, Nicosia 1678, Cyprus

E mail: Christian.Kramer@roche.com and Pawel.Mochalski@uibk.ac.at

OPEN ACCESS

RECEIVED
17 August 2015

REVISED
25 October 2015

ACCEPTED FOR PUBLICATION
29 October 2015

PUBLISHED
27 January 2016

Keywords: blood:air partition coefficient, fat: blood partition coefficient, VOCs, volatile organic compounds, breath analysis, prediction, volatilome

Supplementary material for this article is available online

Abstract

In this article, a database of blood:air and fat:air partition coefficients ($\lambda_{b:a}$ and $\lambda_{f:a}$) is reported for estimating 1678 volatile organic compounds recently reported to appear in the volatilome of the healthy human. For this purpose, a quantitative structure-property relationship (QSPR) approach was applied and a novel method for Henry's law constants prediction developed. A random forest model based on Molecular Operating Environment 2D (MOE2D) descriptors based on 2619 literature-reported Henry's constant values was built. The calculated Henry's law constants correlate very well ($R^2_{test} = 0.967$) with the available experimental data. Blood:air and fat:air partition coefficients were calculated according to the method proposed by Poulin and Krishnan using the estimated Henry's constant values. The obtained values correlate reasonably well with the experimentally determined ones for a test set of 90 VOCs ($R^2 = 0.95$). The provided data aim to fill in the literature data gap and further assist the interpretation of results in studies of the human volatilome.

1. Introduction

It is well established that volatile organic compounds (VOCs) produced and then partially released by the human body have a great potential for diagnosis in physiology and medicine. In particular, this volatile chemical fingerprint can provide non-invasive and real-time information on infections, cancer development, metabolic disorders, progression of therapeutic intervention as well as individual's exposure to environmental pollutants, or toxins [1–3]. For instance, VOC patterns identified in breath proved to be useful for recognition of lung cancer [4–9], gastric cancer [10, 11], and breast cancer [12]. Despite this huge potential, the use of these patterns within a clinical setting is still rather limited. The main unresolved issue is the poor understanding of the origin, behavior, and metabolic fate of their constituents in the human organism. Within this framework, the knowledge of the fundamental physicochemical parameters of identified markers governing their distribution in human organism is highly desirable.

A recent review reported a database of 1764 volatiles appearing in different human body fluids [13]. Amongst these, 874 were detected in exhaled breath, 279 in urine, 504 in skin emanations, 353 in saliva, 130 in blood, and 381 in feces. These compounds belong to diverse chemical families and thereby exhibit very different physicochemical properties. In principle, two fundamental physicochemical parameters govern the behavior and fate of VOCs in humans. These are blood:air and fat:blood partition coefficients. In particular, the blood:air partition coefficient ($\lambda_{b:a}$) is a paramount determinant of pulmonary gas exchange, which together with ventilatory flow and cardiac output

[6] Dedicated to the memory of our friend, colleague, and mentor Anton Amann.

determines both the inhalational uptake of exogenous vapors and the elimination of endogenous compounds via exhalation. This parameter is particularly crucial in breath gas analysis, which usually aims at the identification and exploitation of volatile constituents of human breath for the diagnosis of disease states that frequently occur in distant parts of the body [1, 2]. More specifically, VOCs exhibiting lower affinity for blood ($\lambda_{b:a} < 10\,(\text{mol} \times L_b^{-1})/(\text{mol} \times L_a^{-1})$) exchange merely in the alveoli, whereas those with high blood affinity ($\lambda_{b:a} > 100$) exchange also in the airways [14–16]. Moreover, breath levels of poorly blood-soluble VOCs react very sensitively to changes in ventilation and perfusion, which can be misinterpreted as fluctuations in the endogenous (blood) levels [14, 17]. Blood affinity also influences the peripheral gas exchange via the relation $\lambda_{tissue:b} = \lambda_{tissue:a}/\lambda_{b:a}$. The tissue:blood partition coefficient ($\lambda_{tissue:b}$) is commonly used to describe a venous equilibrium between blood and the respective tissue. The fat:blood partition coefficient ($\lambda_{f:b}$), in turn, governs the distribution of VOCs between the blood compartment and fat tissue and lipophilic cell membranes. Lipophilic volatiles tend to accumulate in lipid membranes, or fat compartment, whereas, compounds with low $\lambda_{f:b}$ readily leave lipophilic cell membranes and drain into blood. Together $\lambda_{f:b}$ and $\lambda_{b:a}$ determine the equilibrium concentration of a given compound between breath, blood and fat and assist in modeling the uptake, distribution, and elimination of VOCs in the human organism [14–19].

The blood:air and the fat:blood partition coefficients of VOCs observed in the human volatilome can differ by more than 12 orders of magnitude [20, 21]. This means that species having comparable levels in exhaled breath can exhibit disparate concentrations in blood and fat. This effect can be illustrated using two volatiles that are omnipresent in human breath; isoprene and acetone [22–25]. The isoprene blood:air partition coefficient ($\lambda_{b:a}$) amounts to 0.95 $(\text{mol} \times L_b^{-1})/(\text{mol} \times L_a^{-1})$ [26], whereas, its $\lambda_{f:b}$ amounts to 82 $(\text{mol} \times L_f^{-1})/(\text{mol} \times L_b^{-1})$ [27]. Acetone, in turn, has a blood:air partition coefficient of ~340 $(\text{mol} \times L_b^{-1})/(\text{mol} \times L_a^{-1})$ [28] and a fat:blood partition coefficient of ~0.253 $(\text{mol} \times L_f^{-1})/(\text{mol} \times L_b^{-1})$ [29]. Assuming a concentration of both species in alveolar air of 200 ppb ($7.76 \times 10^{-9}\,\text{mol} \times L^{-1}$ at 37°C and 1 bar), their equilibrium concentrations in blood are

$$C_{blood}(\text{isoprene}) = 7.4 \times 10^{-9}\text{mol} \times L^{-1}$$
$$C_{blood}(\text{acetone}) = 2.6 \times 10^{-6}\text{mol} \times L^{-1}$$

and the equilibrium concentrations in fat

$$C_{fat}(\text{isoprene}) = 6.0 \times 10^{-7}\text{mol} \times L^{-1}$$
$$C_{fat}(\text{acetone}) = 6.7 \times 10^{-7}\text{mol} \times L^{-1}$$

Hence the concentrations of isoprene and acetone are very different in blood and similar in the fat compartment. Thus, the same concentration of VOCs in alveolar air may correspond to blood concentrations that vary more than 8 orders of magnitude. The same holds true for the concentrations in the fat compartment. These essential differences in physicochemical properties of compounds forming the human volatilome [30] may result in different exhalation kinetics manifested, e.g. via different responses of the breath constituents during the moderate workload ergometer challenge [14, 15, 17, 19, 31].

In this context, it becomes clear that the knowledge of reliable blood:air and fat:blood partition coefficients of VOCs observed in the human volatilome is of utmost importance for the understanding of their behavior in the human body, identification of their underlying biochemical pathways and assessment of their applicability in diagnosis and therapy monitoring. Currently, the methods for determining blood:air and fat:blood partition coefficients can be classified as experimental and predictive approaches. The experimental methods (mainly headspace techniques) employ direct measurements of the gas and blood/fat concentrations of an analyte in closed containers under equilibrium conditions [26, 29, 32–34]. It should be stressed here that the fat:blood partition coefficient is usually determined indirectly via measurements of the fat:air partition coefficient and dividing the latter by a respective blood:air partition coefficient. Unfortunately, the experimental approach suffers from various downsides related to tissue sampling and handling, analytical treatment (e.g. losses of analytes, contaminations, sample decomposition), or unavailability of reference materials. Moreover, this approach is relatively time and effort consuming and involves special (frequently sophisticated) analytical instrumentation. Consequently, experimentally determined values of the blood:air or fat:air partition coefficients are still lacking for many volatile compounds. On the other hand, predictive approaches calculate $\lambda_{f:b}$ and $\lambda_{b:a}$ using other physicochemical parameters of the compound under scrutiny, such as water:air and n-octanol:water partition coefficients ($\lambda_{w:a}$ and $\lambda_{o:w}$), vapor pressures, blood composition, or previously obtained $\lambda_{f:b}$ and/or $\lambda_{b:a}$ of homolog compounds [33, 35–40]. Predictive approaches can, however, fail, or lead to wrong results when the necessary physicochemical characteristics are not available or incorrectly determined, or unknown factors influencing the compound's solubility occur (e.g. VOC protein binding).

Within this context, the primary objective of this paper is to estimate blood:air and fat:blood partition coefficients for 1678 VOCs reported to occur in the human volatilome [13]. Thereby, we expect to fill the literature data gap and support the interpretation of results in studies on the human volatilome.

2. Methods

In principle, partition coefficients of organic compounds can be predicted according to two different

methods: (a) from their chemical structure or (b) by inference from other physicochemical properties.

If predictions are to be made based on the chemical structure, ab-initio methods based on primary physical principles (quantum mechanical calculations or force field simulations) or methods based on chemical similarity can be used. Ab-initio methods are usually very time-consuming, and their predictive accuracy varies with the target property. On the other hand, methods based on the chemical similarity can give very good results if a large enough set of training data or knowledge base is available. This is the realm of quantitative structure-property relationship (QSPR) models, which are used in daily routine in many different fields that require predictions of physicochemical properties. QSPR model predictions are usually very fast to calculate, and the uncertainty in the individual predictions can be quantified very well.

Predictions based on inference from other physicochemical properties (classic formulae) can be used if these properties and the scaling factors are known. For example, the blood:air and fat:air partition coefficients can be predicted from the octanol:water and the water:air partition coefficients according to the formula of Poulin & Krishnan [41]. While octanol:water partition coefficients for many different organic volatile compounds can be found in respective databases, the data on water:air partition coefficients are much more limited. Thus, a two stage strategy was used to derive estimates for blood:air and fat:air partition coefficients. First, water:air partition coefficients were calculated from the chemical similarity using a QSPR model, and then blood:air and fat:air partition coefficients were calculated by inference from tabulated octanol:water and estimated water:air partition coefficients.

2.1. Prediction of Henry's law constants

2.1.1. Training data set

The training dataset of organic compounds with measured Henry's constant values was assembled from four different sources:

(1) Compilation of measured Henry's law constants provided by Sander [42] (214 compounds).
(2) PHYSPROP database (411 compounds) containing data stemming from an article by Katriztky *et al* (1998) [43].
(3) Compilation of Henry's law constants published by Dearden *et al* (2007) containing directly and indirectly measured Henry's constant values [44] (940 compounds).
(4) Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds [45] (1940 compounds).

All values of Henry's constant were converted to atm $\times$ m^3 $\times$ mol^{-1}. This unit corresponds to the Henry's constant expressed as the ratio between gas partial pressure and its water concentration. This unit was used for all further modeling. Apart from this, all data given in the above compilations were used as given, no further inclusion/exclusion criteria of data points were applied.

2.1.2. Chemical structure assignment

In the four sources described above, the chemical identity of the compounds is either specified by the CAS (Chemical Abstracts Service) number, trivial name or IUPAC name. The chemical structures (as SMILES) of the compounds were derived via two different web services: the CACTVS chemical identifier resolver hosted by the NIH (http://cactus.nci.nih.gov/chemical/structure), and the ChemSpider API [46]. Structures obtained for CAS numbers were considered the most reliable. If no CAS number was available, structures obtained from the Chemical Identifier Resolver were prioritized over structures obtained from the ChemSpider API. Structures with chemical identifiers where the ChemSpider API yielded more than ten different solutions were individually checked. Identifiers that yielded zero hits in any web service were individually determined. For all chemical structures, the InCHI Key [47] was calculated using OpenBabel [48] in order to identify and remove duplicate entries within each datasets and between the datasets.

After removal of all duplicates and counterions, the structure assignment and cleanup procedure yielded 2619 different compounds with assigned Henry's constant values. A full list of those can be found in the supporting information.

2.1.3. Descriptors

In order to generate QSPR (quantitative structure-property relationship) models, the full set of 2D descriptors ($n = 192$) available from MOE [49] was calculated for all compounds. A table with the numerical values for all descriptors and compounds can be found in the supporting information. During the course of model development, different descriptor sets were also tried, namely the RDKit 2D descriptors [50] and the ParaSurf Surface Integral Model descriptors [51]. However, the models based on these descriptors did not perform better and are not reported in this publication.

2.1.4. Model training

Based on the descriptors and the measured Henry's constant values, random forest models [52] were trained to predict the Henry's constants. Random forests are a standard nonlinear machine learning tool in chemoinformatics for generating QSPR models. In large scale validations, random forests usually turn out to be among the best performing machine learning algorithms for QSPR [53].

In brief, a random forest is a collection of decision trees (in this study $n_{trees} = 500$), each trained on a bagged subset of the overall data (in this study the default 'sampling with replacement' was used, which leads to 63.2% of the overall dataset being used as

training set for each tree). On every split level, a randomly selected subset of the descriptors (the default for regression: $n_{\text{descriptors}}/3$) is evaluated for the split point that gives the largest improvement in root mean squared error (RMSE). For every tree, the out-of-bag test set is predicted and the final prediction is the average of the individual predictions of the out-of-bag predictions. No descriptor selection was applied, therefore the out-of-bag predictions represent the predictions for an independent validation set. In other words, individual compounds are not used to predict the partition coefficients of themselves at any stage, thus the results presented resemble a fully independent validation.

Sheridan *et al* [54] reported that random forests tend to underestimate extreme values and this behavior can be alleviated by rescaling the predictions. Thus, an additional crossvalidation loop was introduced to calculate scaling parameters and apply them to the predictions: the overall dataset was split in ten parts, and nine out of the ten parts were repeatedly used to train the model and to calculate the scaling parameters. The Henry's law constant values of the tenth part that was not used for any model building was then predicted and the prediction was rescaled.

During the course of the model development, different linear QSPR modeling approaches were tried, including PLS [55] and stepwise regression with descriptor pool size adjusted F-values [56]. However, those linear models performed slightly worse and are not reported in this publication.

2.1.5. Error model training

Sheridan has recently introduced the concept of building a separate error model to predict the confidence intervals for individual predictions [54]. In this approach, an additional random forest model is trained to predict the absolute error of the QSAR/QSPR model. Sheridan showed that the tree standard deviation and the absolute predicted value are two essential descriptors that have a high predictivity for the absolute error [57, 58]. In the initial experiments, it was found that the MOE descriptors add additional predictivity, improving the correlation between the predicted and the absolute error. Therefore all MOE descriptors from above plus the tree standard deviation and the predicted value were used to train error models.

To build the error models, the strategy outlined by Sheridan was followed, using the same random forest setup and the tenfold cross validation for scaling, as above. The error model predictions were scaled to reproduce the moving window RMSE (or standard deviation, this is the same here). In order to calculate the moving window RMSE, the compounds were sorted according to the predicted absolute error, and the experimental and predicted RMSE for each compound was calculated from all compounds within a window of 101 (50 to the left, 50 to the right). The RMSE within the moving window is termed 'local' RMSE, since it is different for every compound and depends on the order of the compounds.

For both QSPR model and error model, the R implementation 'randomForest' by Wiener and Liaw [59] was used to generate the models.

2.1.6. Performance metrics

The quality of the QSPR models is measured using the R^2, RMSE, and mean unsigned error (MUE). All metrics are calculated based on the predictions of the outer loop cross validation test folds.

$$R^2 = \left(\frac{\sum_{j=1}^{n} (\log_{10} H_{\text{exp},j} - \overline{\log_{10} H_{\text{exp}}})(\log_{10} H_{\text{pred},j} - \overline{\log_{10} H_{\text{pred}}})}{\sqrt{\sum_{j=1}^{n} (\log_{10} H_{\text{exp},j} - \overline{\log_{10} H_{\text{exp}}})^2} \sqrt{\sum_{j=1}^{n} (\log_{10} H_{\text{pred},j} - \overline{\log_{10} H_{\text{pred}}})^2}} \right) \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (\log_{10} H_{\text{exp},j} - \log_{10} H_{\text{pred},j})^2} \quad (2)$$

$$\text{MUE} = \frac{1}{n} \sum_{j=1}^{n} \left| \log_{10} H_{\text{exp},j} - \log_{10} H_{\text{pred},j} \right| \quad (3)$$

Here n is the number of compounds, $\log_{10} H_{\text{exp}}$ is the logarithm of the experimentally determined Henry's constant, $\log_{10} H_{\text{pred}}$ is the logarithm of the predicted Henry's constant, $\log_{10}(H_{\text{exp}})$ is the average of the experimentally determined Henry's constant and $\overline{\log_{10}(H_{\text{pred}})}$ is the average of the predicted Henry's constants.

2.1.7. Henry law constant prediction for the data base of VOCs

The final QSPR and error model were used to predict Henry's constant values and the prediction standard deviation for the 1741 compounds of the data base of volatile compounds. In addition, the structures of the data base of volatile compounds were compared with the 2619 compounds from the training set and 68 overlapping compounds were found. The predictions and the reported experimental values are given in the supporting information.

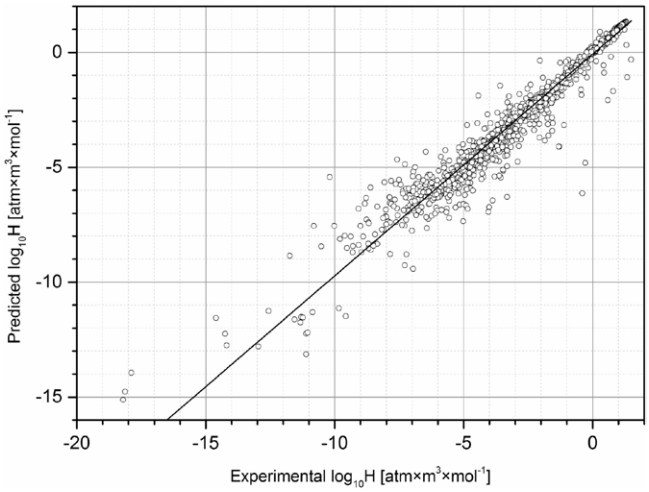

Figure 1. Plot of predicted versus measured Henry's constant value ($\log_{10}$H) for the 2619 Henry's constant literature values. Predictions are based on crossvalidation, so none of the compound has been used to predict its own value.

2.2. Prediction of blood-air and fat-air partition coefficients

Using the predicted values of the Henry law constant, the blood-air and the fat-air partition coefficients were estimated using the method of Poulin & Krishnan [41]. The blood-air partition coefficient can be estimated by the formula:

$$\lambda_{b:a} = \lambda_{o:w}\lambda_{w:a}(a + 0.3b) + \lambda_{w:a}(c + 0.7b) \quad (4)$$

Where, $a \approx 0.0033$ is the fraction of neutral lipids in blood, b ≈ 0.0024 is the fraction of phospholipids in blood, $c \approx 0.82$ is the fraction of water in blood, and $\lambda_{o:w}$ is the octanol:water partition coefficient. The $\lambda_{o:w}$ values can be taken from SciFinder (https://scifinder.cas.org). The fat-air partition coefficients $\lambda_{f:a}$ are estimated by an analogous method from Poulin & Krishnan [41] given by the equation:

$$\lambda_{f:a} = \lambda_{o:w}\lambda_{w:a}(A + 0.3B) + \lambda_{w:a}(C + 0.7B) \quad (5)$$

Where, $A \approx 0.798$ is the fraction of neutral lipids in adipose tissue (fat), $B \approx 0.002$ is the fraction of phospholipids in adipose tissue, and $C \approx 0.15$ is the fraction of water in adipose tissue.

3. Results

3.1. Exemplary calculations for CS_2

The calculations illustrated below are for carbon disulfide (CS_2). The predicted logarithm of Henry's law constant for CS_2 is -2.45 ($\log_{10}$[atm × m^3 × mol^{-1}]). The enthalpy of vaporization is 26.74 kJ mol^{-1} and the measured value of $\lambda_{o:w}$ is 1.94. The Henry's constant can be converted to the dimensionless Henry's law constant (C_{air}/C_{water}) at 310.15 K according to

$$H'_{TS} = \frac{H_r \exp(-\frac{\Delta H_{v,TS}}{R_c})(\frac{1}{T_S} - \frac{1}{T_R})}{RT_S} \quad (6)$$

where T_s is the body temperature (310.15 K), T_R is the reference temperature (298.15 K) H'_{TS} is the dimensionless value for the Henry's constant at 310.15 K, $\Delta H_{v,TS}$ is the enthalpy of vaporization in cal × mol^{-1}, H_r is Henry's law constant at 298.15 K in atm × m^3 × mol^{-1}, R_C (=1.9872 cal × mol^{-1} × K^{-1}) and R (8.205 × 10^{-5} atm × m^3 × mol^{-1} × K^{-1}) are the gas constant in different units. Inserting the above values into equation (6) yields

$$H'_{TS}(CS_2) = \frac{10^{-2.45}\exp\left(-\frac{26.74\left[\frac{kJ}{mol}\right] \times 239.005\,736\left[\frac{cal}{kJ}\right]}{1.9872}\right)\left(\frac{1}{310.15} - \frac{1}{298.15}\right)}{0.000\,082\,05 \times 310.15} \times \left[\frac{c_{air}}{c_{water}}\right] = 0.2128 \quad (7)$$

Thus, $\lambda_{w:a}$ is $\log_{10}(1/H'_{Ts}(CS_2)) = 0.672$. Using this value in equations (4) and (5) gives $\lambda_{b:a} = 0.56$ and $\lambda_{f:a} = 1.14$ (all dimensionless).

3.2. Henry's law constants

For the 2619 membered Henry's constant data set, a QSPR model for the Henry's constant values was generated with $R^2 = 0.967$, RMSE = 0.49 and MUE = 0.22. A plot of predicted versus measured Henry's constant is shown in figure 1.

The predicted values of Henry's constant span a range from roughly -15 to 1 atm × m^{-3} × mol^{-1} and agree well with literature values. Errors tend to become larger as the values of Henry's constant become extremely negative. The predicted local RMSE error of the hold-out test set, calculated with the moving window approach, is plotted against the empirical experimental local RMSE in figure 2.

Figure 2 shows that the predicted error correlates very well with the empirical local RMSE. This indicates that the individual error estimates are highly reliable. Overall the predicted error estimates span a range from 0.01 to 3.14 $\log_{10}$ Henry's constant values. The error estimates represent the standard deviation of the

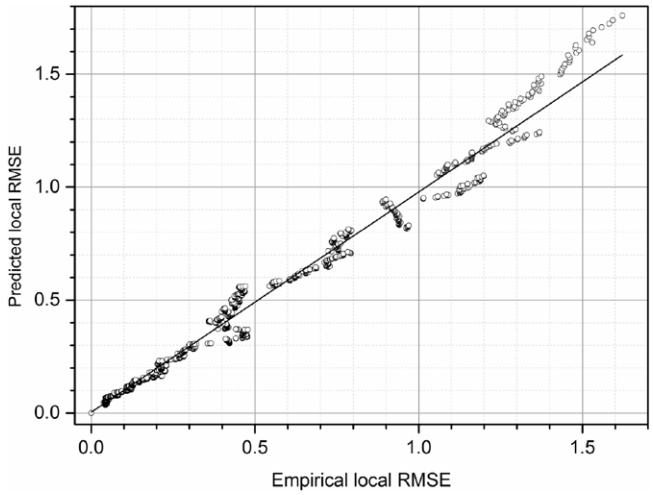

Figure 2. Predicted local RMSE versus empirical local RMSE in $\log_{10}$H, both calculate using the moving window Ansatz.

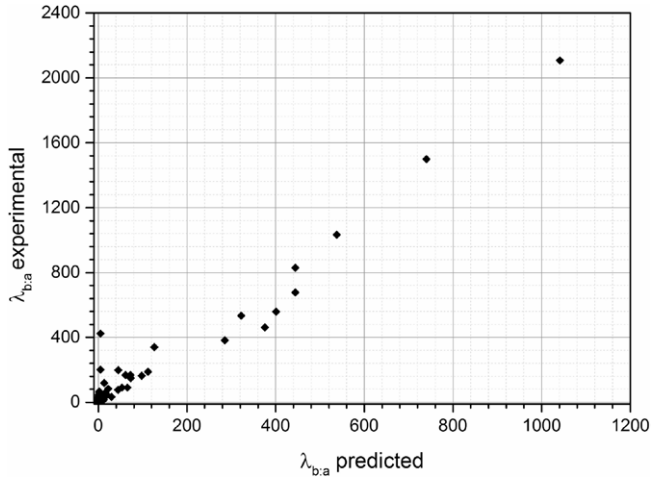

Figure 3. Plot of estimated versus experimentally determined blood:air partition coefficients ($\lambda_{b:a}$).

prediction interval. It means, that there is a 68% probability that the experimental value is within the predicted value +/− the standard deviation. Large standard deviations indicate that the model predictions are rather unsafe, and there is probably no way around measuring the Henry's constant. In contrast, small standard deviations indicate that the predicted values are highly reliable, whereas the individual usage of the predicted values dictates whether the reliability is sufficient.

3.3. Blood:air and fat:air partition coefficients

The predicted values of blood:air and fat:air partition coefficients for selected compounds of interest are presented in table 1. These values were used to calculate the fat:blood partition coefficients ($\lambda_{f:b}$) using the formula $\lambda_{f:b}\lambda_{f:a}/\lambda_{b:a}$. The values of estimated blood:air and fat:air partition coefficients were compared to the experimentally determined values of species of interest. An extensive literature survey resulted in collecting the experimental data for 90 VOCs in case of $\lambda_{b:a}$ and 29 VOCs for $\lambda_{f:a}$ from the list of 1678 species [29, 33, 40, 60–71]. Only values obtained for human blood and adipose tissue were taken into consideration. In the case of the majority of species only one single literature value determined for small populations, was available. This fact may lower the reliability of the experimental data as reference values. A plot of estimated versus measured blood:air partition coefficients is shown in figure 3. The predicted blood:air partition coefficients correlate reasonably well ($R^2 = 0.95$) with the experimentally determined values of this parameter. For instance, the predicted value of $\lambda_{b:a}$ for acetone amounts to 126 and agrees well with the available experimental data (340 [72], 245 [61], 186 [29]). In case of isoprene the predicted and experimentally determined values amounted to 0.253 and 0.95 [26], respectively.

4. Conclusions

The main goal of this work was to create a database of blood:air and fat:air partition coefficients for 1764 VOCs reported to occur in the human volatilome. Since experimental data of these parameters are relatively rare (e.g. measured $\lambda_{b:a}$ have been found for only

Table 1. Estimated/predicted values of Henry's constants ($\lambda_{w:a}$), blood:air partition coefficients ($\lambda_{b:a}$), fat:air partition coefficients ($\lambda_{f:a}$) and fat:blood partition coefficients ($\lambda_{f:b}$) for 1678 volatile organic compounds reviewed by de Lacy Costello *et al* [13]. The table presents also occurrence of these species in human body (after [13]). Only a part of this table is presented here; for the full version, see the online supplementary material.

		Occurrence							Predicted partition coefficients			
VOC	CAS	Faeces	Urine	Breath	Skin	Milk	Blood	Saliva	$\lambda_{w:a}$	$\lambda_{b:a}$	$\lambda_{f:a}$	$\lambda_{f:b}$
Carbon disulfide	75-15-0	F	U	Br	Sk	M			4.70	3.61	13.9	3.85
Furan	110-00-9		U	Br		M			10.2	6.84	35.4	5.18
1,1,1-Trichloroethane	71-55-6					M	Bl		2.15	1.89	4.40	2.49
Vinylchloride	75-01-4			Br					1.15	1.13	1.23	1.1
3,7,7-Trimethyl-bicyclo[4.1.0]hept-3-ene	13466-78-9	F		Br					0.41	0.470	0.0391	0.0832
Tetrachloroethane	79-34-5		U				Bl		21.9	13.0	496	38.2
DL-Limonene	138-86-3	F	U	Br		M		Sa	0.75	0.783	0.331	0.423
2,3,4-Trimethylpentane	565-75-3			Br					5.65×10^{-3}	0.013	8.18×10^{-9}	6.3×10^{-7}
Methanol	67-56-1	F		Br		M	Bl		4840	1040	0.0337	3.24×10^{-5}
1,3-Butadiene	106-99-0			Br					0.342	0.411	0.163	0.396

90 species), their values were estimated using a simple predictive approach proposed by Poulin & Krishnan [13]. For the purpose of this approach, a QSPR model was generated to predict the Henry's constant values. The QSPR model was built based on 2619 Henry's constant values assembled from the literature. It uses standard QSPR methodology, the random forest machine learning algorithm, and MOE2D descriptors. In addition, a separate error model was generated to estimate individual model uncertainty, an approach that has recently been pioneered by Sheridan [54]. Compared to other QSPR models, the model has an excellent performance ($R^2 = 0.97$) and a highly reliable error estimation, which allows judging the predicted values for each compound individually. The blood:air partition coefficients were calculated for 1678 species from the list of 1764 volatiles reported by de Lacy Costello *et al* according to the method of Poulin & Krishnan using the modelled Henry's constant values. These values agree reasonably well with the available experimental data with $R^2 = 0.95$.

Nevertheless, the limitations of the study should be indicated. First, the values of the estimated partition coefficients can be affected by the uncertainties of the other parameters used for their prediction (e.g. octanol:water partition coefficient, enthalpy of vaporization). Secondly, additional factors influencing the compound's solubility (e.g. blood proteins binding, differences in blood composition) were not taken into consideration. Considering these variations, it becomes clear that the predictions do not provide precise values of partition coefficients under scrutiny. Thus the values reported within this manuscript should be considered as an approximation of real values. Consequently, careful use of the database is needed. Thus, it is recommended in the first place to use experimentally determined values of blood:air and fat:air partition coefficients. If the human experimental data are not available in the literature, experimental values of partition coefficients determined for animals (e.g. rat-$\lambda_{b:a}$) can be applied as a reasonable surrogate. Finally, in case of the absence of measured values of $\lambda_{b:a}$ and $\lambda_{f:a}$ estimated within this study partition coefficients should be applied. To sum up, it is expected that partition coefficients data provided by this study will assist future investigations in this exciting field.

Acknowledgments

PM and KU gratefully acknowledge support from the Austrian Science Fund (FWF) under Grant No. P24736-B23. We thank the government of Vorarlberg (Austria) for its generous support. This work received funding from the European Union's Horizon 2020 Programme for research, technological development and demonstration under grant agreement No 644031. We also appreciate funding from the Austrian Agency for International Cooperation in Education and Research (OeAD-GmbH, project SPA 05/202-FEM-BREATH)

References

[1] Amann A and Smith D 2005 *Breath Analysis for Clinical Diagnosis and Therapeutic Monitoring* (New Jersey: World Scientific)
[2] Amann A and Smith D 2013 *Volatile Biomarkers Non-Invasive Diagnosis in Physiology and Medicine* ed A Amann and D Smith (Amsterdam: Elsevier)
[3] Horvath I and de Jongste J E 2010 *European Respiratory Monograph, Number 49: Exhaled Biomarkers* (Sheffield: European Respiratory Society)
[4] Hakim M, Broza Y Y, Barash O, Peled N, Phillips M, Amann A and Haick H 2012 Volatile organic compounds of lung cancer and possible biochemical pathways *Chem. Rev.* **112** 5949–66
[5] Barash O, Peled N, Tisch U, Ionescu R, Ilouze M, Mattei J, Bunn P, Hirsch F and Haick H 2012 Volatile fingerprints of lung cancer specific genetic mutations *J. Thorac. Oncol.* **7** S39
[6] Haick H, Peled N, Hakim M, Barash O, Mettei J, Tisch U, Bunn P and Hirsch F R 2011 Early detection and screening of lung cancer via volatile biomarkers *J. Thorac. Oncol.* **6** S85–7
[7] Bajtarevic A *et al* 2009 Noninvasive detection of lung cancer by analysis of exhaled breath *BMC Cancer* **9** 348
[8] Phillips M *et al* 2007 Prediction of lung cancer using volatile biomarkers in breath *Cancer Biomark.* **3** 95–109 (PMID: 17522431)
[9] Phillips M *et al* 2008 Detection of lung cancer using weighted digital analysis of breath biomarkers *Clin. Chim. Acta* **393** 76–84
[10] Leja M, Amal H, Funka K, Lasina I, Skapars R, Ancans G, Liepniece-Karele I and Haick H 2013 Volatile organic compound in the breath to differentiate between gastric cancer and benign conditions *Helicobacter* **18** 119
[11] Amal H, Funka K, Liepniece-Karele I, Skapars R, Leja M and Haick H 2013 Volatile markers can discriminate between gastric cancer and benign conditions *Gastroenterology* **144** S353
[12] Phillips M, Cataneo R N, Saunders C, Hope P, Schmitt P and Wai J 2010 Volatile biomarkers in the breath of women with breast cancer *J. Breath Res.* **4** 026003
[13] Costello B D, Amann A, Al-Kateb H, Flynn C, Filipiak W, Khalid T, Osborne D and Ratcliffe N M 2014 A review of the volatiles from the healthy human body *J Breath Res.* **8**
[14] King J, Koc H, Unterkofler K, Mochalski P, Kupferthaler A, Teschl G, Teschl S, Hinterhuber H and Amann A 2010 Physiological modeling of isoprene dynamics in exhaled breath *J. Theor. Biol.* **267** 626–37
[15] King J, Unterkofler K, Teschl G, Teschl S, Koc H, Hinterhuber H and Amann A 2011 A mathematical model for breath gas analysis of volatile organic compounds with special emphasis on acetone *J. Math. Biol.* **63** 959–99
[16] King J, Unterkofler K, Teschl G, Teschl S, Mochalski P, Koc H, Hinterhuber H and Amann A 2012 A modeling-based evaluation of isothermal rebreathing for breath gas analyses of highly soluble volatile organic compounds *J. Breath Res.* **6** 016005
[17] Unterkofler K, King J, Mochalski P, Jandacka M, Koc H, Teschl S, Amann A and Teschl G 2015 Modeling-based determination of physiological parameters of systemic VOCs by breath gas analysis: a pilot study *J. Breath Res.* **9** 036002
[18] King J, Mochalski P, Unterkofler K, Teschl G, Klieber M, Stein M, Amann A and Baumann M 2012 Breath isoprene: Muscle dystrophy patients support the concept of a pool of isoprene in the periphery of the human body *Biochem. Biophys. Res. Commun.* **423** 526–30
[19] Koc H, King J, Teschl G, Unterkofler K, Teschl S, Mochalski P, Hinterhuber H and Amann A 2011 The role of mathematical modeling in VOC analysis using isoprene as a prototypic example *J. Breath Res.* **5** 037102
[20] Amann A, Mochalski P, Ruzsanyi V, Broza Y Y and Haick H 2014 Assessment of the exhalation kinetics of volatile cancer

biomarkers based on their physicochemical properties *J. Breath Res.* **8** 016003
[21] Haick H, Broza Y Y, Mochalski P, Ruzsanyi V and Amann A 2014 Assessment, origin, and implementation of breath volatile cancer markers *Chem Soc Rev* **43** 1423–49
[22] Mochalski P, King J, Klieber M, Unterkofler K, Hinterhuber H, Baumann M and Amann A 2013 Blood and breath levels of selected volatile organic compounds in healthy volunteers *Analyst* **138** 2134–45
[23] Mochalski P, King J, Unterkofler K, Hinterhuber H and Amann A 2014 Emission rates of selected volatile organic compounds from skin of healthy volunteers *J Chromatogr B Analyt. Technol. Biomed. Life Sci.* **959** 62–70
[24] Schwarz K *et al* 2009 Breath acetone-aspects of normal physiology related to age and gender as determined in a PTR-MS study *J. Breath Res.* **3** 027003
[25] Kushch I *et al* 2008 Breath isoprene—aspects of normal physiology related to age, gender and cholesterol profile as determined in a proton transfer reaction mass spectrometry study *Clin. Chem. Lab. Med.* **46** 1011–8
[26] Mochalski P, King J, Kupferthaler A, Unterkofler K, Hinterhuber H and Amann A 2011 Measurement of isoprene solubility in water, human blood and plasma by multiple headspace extraction gas chromatography coupled with solid phase microextraction *J. Breath Res.* **5** 046010
[27] Filser J G, Csanady G A, Denk B, Hartmann M, Kauffmann A, Kessler W, Kreuzer P E, Putz C, Shen J H and Stei P 1996 Toxicokinetics of isoprene in rodents and humans *Toxicology* **113** 278–87
[28] Schrikker A C M, Devries W R, Zwart A and Luijendijk S C M 1985 Uptake of highly soluble gases in the epithelium of the conducting airways *Pflugers Arch.* **405** 389–94
[29] Fiserova-Bergerova V and Diaz M L 1986 Determination and prediction of tissue-gas partition coefficients *Int. Arch. Occup. Environ. Health* **58** 75–87
[30] Amann A, Costello B, Miekisch W, Schubert J, Buszewski B, Pleil J, Ratcliffe N and Risby T 2014 The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva *J. Breath. Res.* **8** 034001
[31] King J, Mochalski P, Kupferthaler A, Unterkofler K, Koc H, Filipiak W, Teschl S, Hinterhuber H and Amann A 2010 Dynamic profiles of volatile organic compounds in exhaled breath as determined by a coupled PTR-MS/GC-MS study *Physiol. Meas.* **31** 1169–84
[32] Abraham M H and Ibrahim A 2006 Air to fat and blood to fat distribution of volatile organic compounds and drugs: linear free energy analyses *Eur. J. Med. Chem.* **41** 1430–8
[33] Meulenberg C J and Vijverberg H P 2000 Empirical relations predicting human and rat tissue:air partition coefficients of volatile organic compounds *Toxicol. Appl. Pharmacol.* **165** 206–16
[34] Abraham M H, Ibrahim A and Acree W E 2005 Air to blood distribution of volatile organic compounds: a linear free energy analysis *Chem. Res. Toxicol.* **18** 904–11
[35] Beliveau M and Krishnan K 2000 Estimation of rat blood:air partition coefficients of volatile organic chemicals using reconstituted mixtures of blood components. *Toxicol. Lett.* **116** 183–8
[36] Paterson S and Mackay D 1989 Correlation of tissue, blood, and air partition coefficients of volatile organic chemicals *Br. J. Ind. Med.* **46** 321–8
[37] Basak S C, Mills D, El-Masri H A, Mumtaz M M and Hawkins D M 2004 Predicting blood:air partition coefficients using theoretical molecular descriptors *Environ. Toxicol. Pharmacol.* **16** 45–55
[38] Peyret T, Poulin P and Krishnan K 2010 A unified algorithm for predicting partition coefficients for PBPK modeling of drugs and environmental chemicals *Toxicol. Appl. Pharmacol.* **249** 197–207
[39] Yun Y E and Edginton A N 2013 Correlation-based prediction of tissue-to-plasma partition coefficients using readily available input parameters *Xenobiotica* **43** 839–52
[40] Mochalski P, King J, Kupferthaler A, Unterkofler K, Hinterhuber H and Amann A 2012 Human blood and plasma partition coefficients for C4-C8 n-alkanes, isoalkanes, and 1-alkenes *Int. J. Toxicol.* **31** 267–75
[41] Poulin P and Krishnan K 1996 Molecular structure-based prediction of the partition coefficients of organic chemicals for physiological pharmacokinetic models *Toxicol. Methods* **6** 117–37
[42] Sander R 2015 Compilation of Henry's law constants (version 4.0) for water as solvent *Atmos. Chem. Phys.* **15** 4399–981
[43] Katritzky A R, Wang Y L, Sild S, Tamm T and Karelson M 1998 QSPR studies on vapor pressure, aqueous solubility, and the prediction of water-air partition coefficients *J. Chem. Inf. Comput. Sci.* **38** 720–5
[44] Modarresi H, Modarress H and Dearden J C 2007 QSPR model of Henry's law constant for a diverse set of organic chemicals based on genetic algorithm-radial basis function network approach *Chemosphere* **66** 2067–76
[45] Gharagheizi F, Abbasi R and Tirandazi B 2010 Prediction of Henry's Law constant of organic compounds in water from a new group-contribution-based model *Ind. Eng. Chem. Res.* **49** 10149–52
[46] Pence H E and Williams A 2010 ChemSpider: an online chemical information resource *J. Chem. Edu.* **87** 1123–4
[47] Heller S, McNaught A, Stein S, Tchekhovskoi Dand Pletnev I 2013 InChI—the worldwide chemical structure identifier standard *J. Cheminform.* **5** 7
[48] O'Boyle N M, Banck M, James C A, Morley C, Vandermeersch T and Hutchison G R 2011 Open babel: an open chemical toolbox *J. Cheminform.* **3** 33
[49] Labute P 2000 A widely applicable set of descriptors *J. Mol. Graph. Model.* **18** 464–77
[50] Landrum G 2014 RDKit Open-Source Cheminformatics Toolkit. Release 2014.03.1
[51] Kramer C, Beck B and Clark T 2010 A surface-integral model for log P-OW *J. Chem. Inf. Model.* **50** 429–36
[52] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
[53] Svetnik V, Liaw A, Tong C, Culberson J C, Sheridan R P and Feuston B P 2003 Random forest: a classification and regression tool for compound classification and QSAR modeling *J. Chem. Inf. Comput. Sci.* **43** 1947–58
[54] Sheridan R P 2013 Using random forest to model the domain applicability of another random forest model *J. Chem. Inf. Model.* **53** 2837–50
[55] Wold H 2004 Partial least squares *Encyclopedia of Statistical Sciences* (New York: Wiley)
[56] Kramer C, Tautermann C S, Livingstone D J, Salt D W, Whitley D C, Beck B and Clark T 2009 Sharpening the toolbox of computational chemistry: a new approximation of critical F-values for multiple linear regression *J. Chem. Inf. Model.* **49** 28–34
[57] Sheridan R P 2012 Three useful dimensions for domain applicability in QSAR models using random forest *J. Chem. Inf. Model.* **52** 814–23
[58] Sheridan R P 2015 The relative importance of domain applicability metrics for estimating prediction errors in QSAR varies with training set diversity *J. Chem. Inf. Model.* **55** 1098–107
[59] Liaw A and Wiener M 2002 Classification and regression by random forest *R. News* **2** 18–22
[60] Gargas M L, Burgess R J, Voisard D E, Cason G H and Andersen M E 1989 Partition-coefficients of low-molecular-weight volatile chemicals in various liquids and tissues *Toxicol. Appl. Pharmacol.* **98** 87–99
[61] Sato A and Nakajima T 1979 Partition-coefficients of some aromatic-hydrocarbons and ketones in water, blood and oil *Br. J. Ind. Med.* **36** 231–4
[62] Abraham M H, Chadha H S and Mitchell R C 1994 Hydrogen-bonding .33. factors that influence the distribution of solutes between blood and brain *J. Pharm. Sci.* **83** 1257–68
[63] Falk A, Gullstrand E, Lof A and Wigaeushjelm E 1990 Liquid air partition-coefficients of 4 terpenes *Br. J. Ind. Med.* **47** 62–4
[64] Poyart C, Bursaux E, Freminet A and Bertin M 1976 Interactions of short chain aliphatic-hydrocarbons with human-blood and hemoglobin-a solutions *Biomed. Express* **25** 224–7

[65] Kaneko T, Wang P Y and Sato A 1994 Partition-coefficients of some acetate esters and alcohols in water, blood, olive oil, and rat-tissues *Occup. Environ. Med.* **51** 68–72
[66] Perbellini L, Brugnone F, Caretta D and Maranelli G 1985 Partition-coefficients of some industrial aliphatic-hydrocarbons (C5-C7) in blood and human-tissues *Br. J. Ind. Med.* **42** 162–7
[67] Nihlen A, Lof A and Johanson G 1995 Liquid/air partition coefficients of methyl and ethyl t-butyl ethers, t-amyl methyl ether, and t-butyl alcohol *J. Expo. Anal. Environ. Epidemiol.* **5** 573–82
[68] Kaneko T, Wang P Y and Sato A 2000 Partition coefficients for gasoline additives and their metabolites *J. Occup. Health* **42** 86–7
[69] Jarnberg J and Johanson G 1995 Liquid/air partition-coefficients of the trimethylbenzenes *Toxicol. Ind. Health* **11** 81–8
[70] Stowell A R, Lindros K O and Salaspuro M P 1980 Breath and blood-acetaldehyde concentrations and their correlation during normal and calcium carbimide-modified ethanol oxidation in man *Biochem. Pharmacol.* **29** 783–7
[71] Luan F, Liu H T, Ma W P and Fan B T 2008 QSPR analysis of air-to-blood distribution of volatile organic compounds *Ecotoxicol. Environ. Saf.* **71** 731–9
[72] Anderson J C, Lamm W J and Hlastala M P 2006 Measuring airway exchange of endogenous acetone using a single-exhalation breathing maneuver *J. Appl. Physiol.* **100** 880–9