Received: 19 March 2018 | Accepted: 26 September 2018

DOI: 10.1111/ina.12510

ORIGINAL ARTICLE

WILEY

A quantitative structure-property relationship (QSPR) for estimating solid material-air partition coefficients of organic compounds

Lei Huang | **Olivier Jolliet**

Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Abor, Michigan

Correspondence
Lei Huang, Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Abor, MI.
Email: huanglei@umich.edu

Funding information
Long Range Research Initiative of the American Chemistry Council and the US EPA, Grant/Award Number: contract EP-16-C-000070

Abstract

The material-air partition coefficient (K_{ma}) is a key parameter to estimate the release of chemicals incorporated in solid materials and resulting human exposures. Existing correlations to estimate K_{ma} are applicable for a limited number of chemical-material combinations without considering the effect of temperature. The present study develops a quantitative structure-property relationship (QSPR) to predict K_{ma} for a large number of chemical-material combinations. We compiled a dataset of 991 measured K_{ma} for 179 chemicals in 22 consolidated material types. A multiple linear regression model predicts K_{ma} as a function of chemical's K_{oa}, enthalpy of vaporization (ΔH_v), temperature, and material type. The model shows good fitting of the experimental dataset with adjusted R^2 of 0.93 and has been verified by internal and external validations to be robust, stable and has good predicting ability ($R^2_{ext} > 0.78$). A generic QSPR is also developed to predict K_{ma} from chemical properties and temperature only (adjusted $R^2 = 0.84$), without the need to assign a specific material type. These QSPRs provide correlation methods to estimate K_{ma} for a wide range of organic chemicals and materials, which will facilitate high-throughput estimates of human exposures for chemicals in solid materials, particularly building materials and furniture.

KEYWORDS
consumer exposure, correlation, indoor release, organic chemicals, partitioning, solid materials

1 | INTRODUCTION

Chemicals incorporated in solid materials have been identified as a major source of passive emissions to indoor air and of transfers into house dust and skin. Typical examples include chemicals used as plasticizers in building materials and flame retardants in furniture. To estimate the release of these chemicals from solid materials and subsequent consumer exposures, the dimensionless solid material-air partition coefficient (K_{ma}), defined as the ratio of the concentration in the material to the concentration in the air at equilibrium, is one of the key parameters.[2] The K_{ma} is essential in determining the chemical transfer from solid material to air and to house dust, as well as the chemical concentration at the material surface, which further determines the inhalation, dermal and dust ingestion exposures. K_{ma} is specific to a chemical-material combination and is also influenced by ambient temperature. Experimental techniques such as chamber tests for building materials,[3] and sorption experiments for polymer materials[4-6] have enabled measurement of a limited number of K_{ma} values for building materials such as vinyl flooring, gypsum board, plywood and cement, as well as polymer materials used for passive samplers including polyurethane foams (PUF), polyethylene (PE), and polypropylene (PP). Recently, studies have also been conducted to measure the K_{ma} for clothing and fabrics.[7,8] However, since experiments are costly and time-consuming, measured K_{ma} values are only available for a limited number of chemical-material

combinations. Thus, quantitative relationships are needed to predict this partition coefficient from known physiochemical properties for chemicals without experimental data, which is especially important for high-throughput approaches, for which a large number of chemical-material combinations need to be evaluated.

Several correlation methods have been developed to estimate K_{ma} from physiochemical properties of chemicals. For example, several studies have correlated K_{ma} to the chemical's vapor pressure using data on volatile organic compounds (VOCs) in building materials.[4,9-11] Other studies which focused on semi-volatile organic compounds (SVOCs) in passive sampling devices have found correlation between K_{ma} and the octanol-air partition coefficient (K_{oa}).[5,6,12,13] Furthermore, Holmgren et al[14] estimated K_{ma} as a function of five Abraham solvation parameters for six groups of materials, but these parameters are not readily available. For the aforementioned approaches, the main limitation is that the correlations are specific to certain chemical classes and materials; for example polycyclic aromatic hydrocarbons (PAHs) in low-density polyethylene (LDPE), which limits their application for other chemical-material combinations. Addressing this research gap to facilitate wider applicability, Guo developed a method which estimates the K_{ma} as a function of the chemical's vapor pressure for all materials and chemical classes.[11] However, this approach is developed based on a small dataset which mainly includes VOCs in building materials limiting its applicability to also address SVOCs. Another limitation of the previous studies is that the effect of temperature was not well considered in the correlation. Some studies provided different correlation coefficients for certain discrete temperatures,[15] while others corrected the predictors for temperature.[16] However, since the known physiochemical properties such as vapor pressure and K_{oa} are often only given as values at 25°C, correcting them for temperature may not always be practical as the corresponding enthalpies of phase change are not available for all chemicals. Several studies did establish correlations between K_{ma} and temperature, but the correlations were only verified using experimental data on limited chemicals such as formaldehyde and other aldehydes.[17,18]

In all, the currently available correlation methods to estimate K_{ma} do not provide sufficient coverage of chemicals incorporated in solid materials at different ambient temperatures. A recent research hotspot in exposure sciences is to develop low tier, high-throughput methods to estimate exposure to chemical in consumer products across a variety of chemical-material combinations, which requires high-throughput estimates of K_{ma} for a wide range of material-chemical combinations. Thus, the present study aims to develop a more comprehensive correlation method to estimate K_{ma} for a wide range of organic compounds in multiple solid materials, addressing the need for high-throughput exposure assessments. More specifically, we aim to:

(1). Carry out a comprehensive literature review to collect experimental K_{ma} data on a wide range of materials and chemicals.
(2). Use multiple linear regression techniques to establish the relationship between K_{ma} and various predictor variables including physiochemical properties, material type, and temperature.
(3). Perform internal and external validations to characterize the validity and predictive power of the developed correlation.

Practical implications

- The developed QSPRs provide a comprehensive correlation method to estimate K_{ma}, covering a much wider range of organic chemicals and solid materials compared to previous studies.
- A still accurate generic correlation without the need to assign a material type is also included.
- Combined with the QSPR estimating the internal diffusion coefficient,[1] these QSPRs facilitate high-throughput estimates of indoor human exposures to chemicals incorporated in solid materials.
- This is highly relevant for multiple science-policy fields, including chemical alternatives assessment (CAA), risk assessment (RA), and life cycle assessment (LCA).

This QSPR provides a more advanced correlation method to estimate the K_{ma} of organic compounds compared to previous studies, as it covers a wide range of solid materials and chemicals, and consistently incorporates the effect of temperature. A similar QSPR has been developed by our group for the internal diffusion coefficient in solid materials.[1] By providing reliable estimates of the key partition and diffusion parameters for a large number of material-chemical combinations, these QSPRs will facilitate high-throughput assessments of chemical emissions and human exposures for chemicals incorporated in solid materials relevant for various science-policy fields such as chemical alternatives assessment (CAA), risk assessment, and life cycle assessment (LCA).

2 | MATERIALS AND METHODS

2.1 | Dataset

2.1.1 | Data collection

Experimental material-air partition coefficient data were compiled from 43 references from the peer-reviewed scientific literature (provided in Section S1). Dimensionless partition coefficients were collected. If the partition coefficients were expressed in mL/g or m^3/g, they were converted to dimensionless values by multiplying these by the density of the solid material. If the partition coefficients were expressed in the unit of m, they were converted to dimensionless values by dividing these by the thickness of the material. The initial dataset of K_{ma} contained a total of 1008 records covering 179 unique chemicals and 75 distinct solid materials.

2.1.2 | Data curation

For the 179 unique chemicals of the initial K_{ma} dataset, molecular weight, vapor pressure, water solubility, and logK_{ow} at 25°C were obtained from EPISuite.[19] For these physiochemical properties, experimental values were used when available, otherwise the software-estimated values were used. The enthalpy of vaporization (ΔH_v, J/mol) of each chemical was obtained from ChemSpider estimated values (www.chemspider.com).

For the octanol-air partition coefficient (logK_{oa}) at 25°C, experimental values are only available for part of the 179 chemicals in the dataset. To avoid inconsistency, we used the logK_{oa} values estimated by EPISuite[19] for all of the 179 chemicals. In EPISuite, logK_{oa} is estimated by subtracting logK_{aw} (dimensionless log air-water partition coefficient) from logK_{ow}, logK_{aw} and logK_{ow} being estimated by the HenryWin and K_{ow}Win functions, respectively.[19] Experimental logK_{oa} values were also collected and their impacts on the QSPR were assessed, as presented in Section S6.

To avoid over-fitting of the QSPR model, the 75 original materials for K_{ma} were grouped into 22 consolidated material types, based on the name of the materials and the similarity of the regression coefficients (see Section S1), ensuring a minimum of five data points and three different chemicals per consolidated material type. The data points with materials that cannot be grouped according to the above criteria were excluded from further analyses.

The final K_{ma} dataset contains 991 data points with 179 unique chemicals in 22 consolidated material types. The temperature at which the K_{ma} was measured ranges from 15 to 100°C. The final dataset is provided in Supporting information.

2.2 | Modeling methods

2.2.1 | Multiple linear regression model

A multiple linear regression (MLR) analysis was performed to identify and quantify the effect of different parameters on the partition coefficient, with details described in our previous paper on the QSPR for diffusion coefficient.[1] Briefly, the MLR model takes the following general form:

$$\log_{10} K_{ma} = \alpha + \beta_1 \cdot X_1 + \cdots + \beta_n \cdot X_n + b_1 \cdot M_1 + \cdots + b_m \cdot M_m \quad (1)$$

where $\log_{10}K_{ma}$ is the logarithm of the dimensionless K_{ma}, α is the intercept; X_1 to X_n are independent variables related to the properties of the chemical or the environment; β_1 to β_n are regression coefficients for the respective independent variables X_1 to X_n. M_1 to M_m are dummy variables for the packaging materials, with one dummy variable per type of material. A dummy variable equals 1 for the material type it represents, and equals 0 for all other materials; for example, M_1 = 1 for material type 1, M_1 = 0 for material types 2 to m. b_1 to b_m are regression coefficients for the respective dummy variables M_1 to M_m. The number of m is equal to the number of material types considered minus one, since PU-ether—the material type with the highest number of measured K_{ma} data—is used as the reference material type and does not require a dummy available in the MLR. Regression coefficients were estimated by the least squares (LS) method. All regression analyses were performed using IBM SPSS Statistics version 23 (IBM corporation, Armonk, New York).

In previous studies, either the chemical's vapor pressure[4,9-11] or logK_{oa}[5,6,12,13] has been used as predictor of the K_{ma} in a given material. Abraham solvation parameters were also used as predictors by Holmgren et al,[14] but these parameters are not considered here since they are not readily available. Initial regressions (Section S2) suggest that logK_{oa} is a better predictor of K_{ma} compared to vapor pressure. Thus, the chemical's logK_{oa} at 25°C was used as the independent variable for chemical properties in Equation (1).

Thus, the MLR model takes the following form:

$$\log_{10} K_{ma} = \alpha + \beta_{\log K_{oa}} \cdot \log_{10} K_{oa} + \beta_T \cdot T\_term + b_1 \cdot M_1 + \cdots + b_{21} \cdot M_{21} \quad (2)$$

where T_term is a term representing the effect of temperature and will be described in the next section (Section 2.2.2).

2.2.2 | Temperature dependence

In thermodynamics, the temperature dependence of equilibrium constant, K_{eq}, can be described by the van't Hoff equation:

$$\ln \frac{K_2}{K_1} = \frac{\Delta H_{phasechange}}{R} \left(\frac{1}{T_2} - \frac{1}{T_1} \right) \quad (3a)$$

where K_1 and K_2 are the equilibrium constants at temperature T_1 and T_2, respectively, T_1 and T_2 are absolute temperatures (K), R is ideal gas constant (8.314 J/(K·mol)), and $\Delta H_{phase\ change}$ is the enthalpy of phase change (J/mol).

Since K_{ma} is an equilibrium constant by definition and the chemical's $\log_{10}K_{oa}$ at 25°C or 298.15 K is used as an independent variable in the MLR model (Equation 2), we assume that the temperature dependence of K_{ma} also follows the van't Hoff equation:

$$T_{term} = \log_{10} \frac{K_{ma\_2}}{K_{ma\_1}} = \frac{\Delta H_{ma}}{2.303 \cdot R} \left(\frac{1}{T_2} - \frac{1}{298.15} \right) \quad (3b)$$

where ΔH_{ma} is the enthalpy of the partitioning between material and air (J/mol), and 2.303 is a conversion factor between $\log_{10}$K and lnK.

Ideally, the enthalpy ΔH_{ma} should be different for different chemical-material combinations. Kamprad and Goss have determined the ΔH_{ma} values for 54 unique chemicals in PU-ether using measured K_{ma} data from 15 to 95°C,[4] so we were able to develop a linear correlation to estimate ΔH_{ma} from chemical properties (results

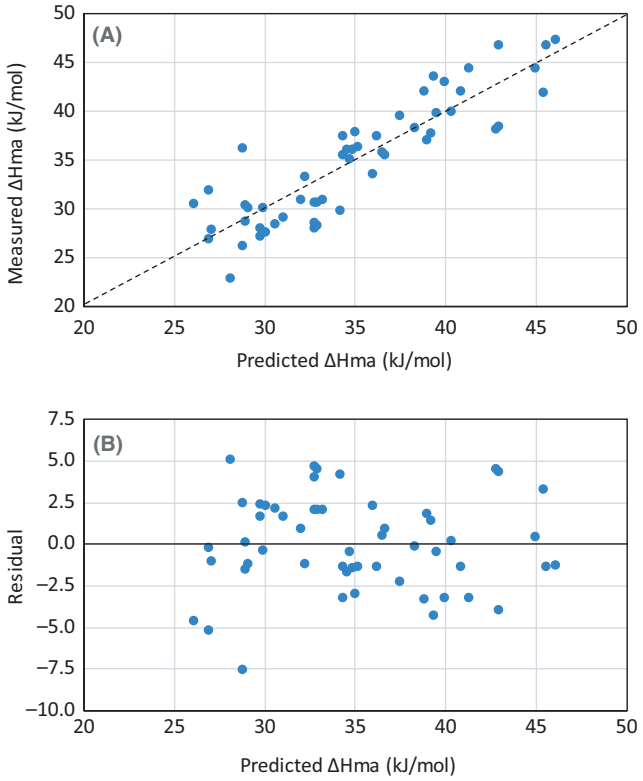

FIGURE 1 A, Measured Enthalpy of material-air partitioning (ΔH_{ma}) and B, residuals as a function of the (ΔH_{ma}) predicted from chemical enthalpy of vaporization (ΔH_v - Equation 5). The dotted line in (A) indicates the 1:1 line

shown in Section 3.1). Since no experimental ΔH_{ma} values are available for materials other than PU-ether, we use the ΔH_{ma} correlation developed above across all materials. Therefore, in our regression model of K_{ma}, the ΔH_{ma} is chemical-specific, but not material-specific. The final MLR model thus takes the following form:

$$\log_{10} K_{ma} = \alpha + \beta_{\log K_{oa}} \cdot \log_{10} K_{oa} + \beta_T \frac{\Delta H_{ma}}{2.303 \cdot R}\left(\frac{1}{T} - \frac{1}{298.15}\right) + b_1 \cdot M_1 + \cdots + b_{21} \cdot M_{21} \quad (4)$$

2.3 | Model validation

Validation of the final MLR model (Equation 4) was performed using the QSARINS software, version 2.2.1 (www.qsar.it) which is developed by Gramatica et al.[20,21]

2.3.1 | Internal validation

The MLR model's capacity to predict portions of the training dataset was evaluated in an internal validation process, using two techniques in QSARINS: the leave more out (LMO) cross-validation and the Y-scrambling, which have been described previously.[1,21] 1000 iterations were used for the LMO cross-validation, and the percentage of the excluded elements was set as 20%, and 1000 iterations for Y-scrambling.

2.3.2 | External validation

We also evaluated the model's ability to provide reliable predictions on new datasets by external validation, using the splitting approach, which split the existing dataset (991 data points) into one training dataset and one prediction dataset. The training dataset was used to generate regression coefficients of the MLR model, and then the MLR model was applied to the prediction set to examine the prediction performances of the model. Three kinds of splitting were performed using existing options in the QSARINS software (see Section S4.1 for details) by random percentage, by ordered response and by structure. We introduced a fourth kind of splitting by studies, where all data points from certain studies were manually selected as the training set and data points from remaining studies as the prediction set. If a consolidated material type only includes data points from one study, all of these data points were assigned into the training set in order to ensure that the MLR model constructed using the training set includes all consolidated material types. The four types of splitting yielded similar sample sizes of approximately 800 data points for the training set and 200 data points for the prediction set (Table S3).

3 | RESULTS AND DISCUSSIONS

3.1 | Temperature dependence

As described in Section 2.2.2, the temperature dependence of K_{ma} is determined by the enthalpy of the partitioning between material and air, ΔH_{ma} (J/mol). Using the measured K_{ma} data for 54 chemicals in PU-ether from 15 to 95°C[4] (data are provided in Section S3), we obtained the following correlation to estimate ΔH_{ma}:

$$\Delta H_{ma} = 1.37 \cdot \Delta H_v - 14.0 \quad (5)$$

N = 54, R^2 = 0.786, R^2_{adj} = 0.782, SE = 2.85, RMSE = 2.80
ANOVA: F = 191, df = 1, P < 0.0001

where ΔH_v is the chemical's enthalpy of vaporization (J/mol) obtained from ChemSpider (www.chemspider.com).

This simple linear model shows good fitting of the experimental ΔH_{ma} data, with an adjusted R-squared of 0.782, and the model fit is highly significant with an ANOVA P-value < 0.0001. Figure 1 shows the scatter plot of predicted vs measured ΔH_{ma} and the residual plot, which indicate good agreement with the 1:1 line and random distribution of residuals throughout the dataset. These results suggest that there is indeed a linear relationship between ΔH_{ma} and ΔH_v in PU-ether, and Equation (5) was also used as default to estimate ΔH_{ma} for all other materials.

3.2 | Final QSPR and model fitting

Using the full dataset (991 data points) and Equation (4), the final MLR model for predicting the solid material-air partition coefficient is as follows:

TABLE 1 Regression coefficients for Equation (6)

Variable	Coefficient	SE[a]	*P*-value
Intercept	−0.38	0.06	<0.001
$\log_{10}K_{oa}$	0.63	0.01	<0.001
ΔH_{ma} (1/T-1/298.15)/2.303R	0.96	0.04	<0.001
Consolidated material types (coefficient b)			
Carpet	1.97	0.14	<0.001
Cellulose fabric (cotton, linen)	0.72	0.12	<0.001
Cement, Calcium silicate	1.11	0.10	<0.001
Concrete	2.20	0.29	<0.001
Ethylene Vinyl Acetate (EVA)	3.50	0.32	<0.001
Glass	1.11	0.29	<0.001
Gypsum board	1.28	0.18	<0.001
Latex and solvent-based paint	2.92	0.19	<0.001
Paper	0.14	0.10	0.16
Plywood	1.36	0.18	<0.001
Polyester fabric	0.60	0.14	<0.001
Polyether ether ketone (PEEK)	2.73	0.29	<0.001
Polyethylene (PE)	2.45	0.17	<0.001
Polypropylene (PP)	2.06	0.29	<0.001
Polytetrafluoroethylene (PTFE)	2.08	0.29	<0.001
PU-ester	−0.72	0.07	<0.001
PU-ether[b]	0.00	0.19	n/a
PUF-undefined	1.06	0.15	<0.001
Rayon fabric	0.97	0.18	<0.001
Stainless steel	2.07	0.29	<0.001
Vinyl flooring	2.26	0.11	<0.001
Wooden boards[c]	2.01	0.09	<0.001

[a]Standard error.
[b]Reference material.
[c]Includes oriented strand board (OSB), particleboard, medium-density board and high-density board.

$$\log_{10} K_{ma} = -0.38 + 0.63 \cdot \log_{10} K_{oa} + 0.96 \cdot \frac{\Delta H_{ma}}{2.303 \cdot R}\left(\frac{1}{T} - \frac{1}{298.15}\right) + b \qquad (6)$$

$N = 991$, $R^2 = 0.934$, $R^2_{adj} = 0.933$, SE = 0.62, RMSE = 0.62
ANOVA: F = 597, $df = 23$, $P < 0.0001$

where K_{ma} is the dimensionless solid material-air partition coefficient, K_{oa} is the chemical's dimensionless octanol-air partition coefficient at 25°C, ΔH_{ma} is the enthalpy of the partitioning between material and air (J/mol) which is given by Equation (5), T is absolute temperature (K), and b is the material-specific coefficients presented in Table 1. This model is provided as an excel model in Supporting Information to facilitate application. The standard errors for the coefficients are also presented in Table 1. An SE of 0.63 of the final model (Equation 6) indicates that the 95% confidence interval (CI) of the predicted $\log K_{ma}$ is the predicted value ± 1.22, indicating that most of the predicted K_{ma} are within a factor of 16 from the measured K_{ma}.

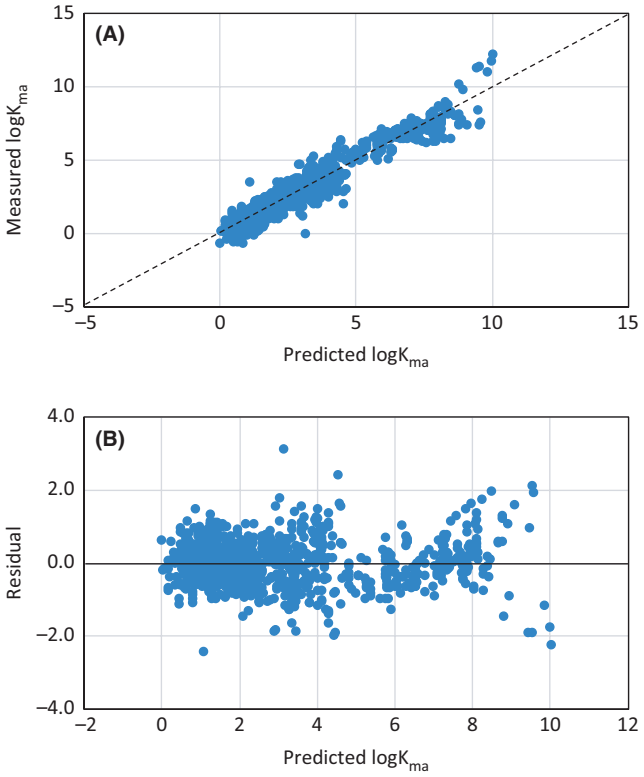

FIGURE 2 A, measured $\log K_{ma}$ and B, residuals as a function of $\log K_{ma}$ predicted by the final QSPR (Equation 6). The dotted line in (A) indicates the 1:1 line

This MLR model shows excellent fitting of the experimental data, with an adjusted R-squared of 0.93 and a root mean square error (RMSE) of 0.62. The model fit is highly significant with an ANOVA *P*-value smaller than 0.0001. Figure 2A shows the scatter plot of predicted vs measured $\log K_{ma}$, which aligns well with the 1:1 line. The residual plot (Figure 1B) shows that the residuals are distributed evenly throughout the dataset, and most residuals have absolute values smaller than 2, again indicating the good fit of the linear model for the data.

This MLR model assumes that the correlation between $\log K_{ma}$ and the chemical's $\log K_{oa}$ is the same across material types, which seems reasonable given the excellent model fitting. Plotting the $\log K_{ma}$ against chemical's $\log K_{oa}$ for selected material types (Figure 3) confirmed that the correlation between $\log K_{ma}$ and the chemical's $\log K_{oa}$ (ie, the slopes of the fitted straight lines in Figure 3) is similar but with slight differences across material types, indicating that a single coefficient for $\log K_{oa}$, as in the present QSPR model, might not be perfect. This could have been accounted for by including interaction terms between $\log K_{oa}$ and material types, but this would introduce 21 more terms in the model without greatly improving the model fitting (Section S5), so the interaction terms were not retained in the final QSPR model.

As described in the methods, this final MLR model uses EPISuite-estimated logK_{oa} values as predictors, since experimental logK_{oa} are not available for all chemicals in the dataset. MLR models developed using mixed logK_{oa} values (ie, for a chemical experimental logK_{oa} is used when available, otherwise EPISuite-estimated logK_{oa} is used) also yielded similar results as the final MLR model (adjusted R^2 ranged from 0.930 to 0.931, for details see Section S6), indicating that the impact of experimental logK_{oa} on the model is minimal.

3.3 | Impact of each predictor

As shown in Equation (6), the key predictors of the solid material-air partition coefficient are the chemical's logK_{oa}, ΔH_v, temperature, and the solid material type. The regression coefficient for logK_{ow} is 0.63 and is highly significant ($P < 0.0001$), indicating that the material-air partition coefficient increases with increasing logK_{oa}, which is consistent with findings from previous studies.[5,6,13]

The regression coefficient of the temperature term is 0.96 and is also highly significant ($P < 0.0001$), indicating that the K_{ma} decreases with higher temperature. Experimental data from Kamprad et al did show reduced K_{ma} with increased temperature, and it also makes intuitive sense that at higher temperature the K_{ma} is lower leading to faster chemical migration from solid material to air. As discussed in Section 3.1, the effect of temperature on K_{ma} also depends on the ΔH_{ma}, which increases linearly with the chemical's enthalpy of vaporization ΔH_v.

The 21 dummy variables for the material types reflect the material dependency of the K_{ma}. As "PU-ether" (polyurethane-ether) was used as the reference material in the regression, the value of its coefficient b is zero (Table 1). For each of the other material types, the coefficient b determines the difference in logK_{ma} between that material type and PU-ether. Chemicals in solid material types with high values of b are more difficult to migrate to air than in those with low values of b. The three material types with highest b coefficients are ethylene vinyl acetate (EVA), latex and solvent-based paint and polyether ether ketone (PEEK) which are dense materials, while the three types with lowest b coefficients are PU-ester (polyurethane-ester), PU-ether and paper which tend to be porous materials. It should be noted that the data for a given consolidated material type were gathered from different studies, and the composition and properties of the material type may vary between studies, so the material type coefficients in Table 1 only represent an average composition and partition behavior for the specific material types.

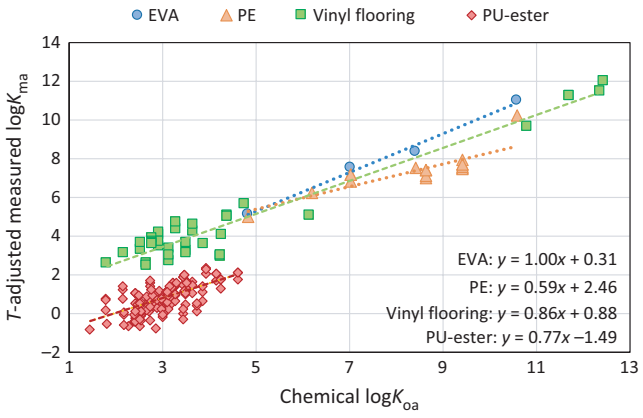

FIGURE 3 Temperature adjusted measured logK_{ma} as a function of logK_{oa} for selected material types including EVA, PE, vinyl flooring, and PU-ester

The significance of the material type coefficient only indicates that the coefficient bs of these material types are significantly different from the reference material type, PU-ether, but if another material type was selected as the reference material, the regression coefficients and statistical significance of all materials would change. Thus, the insignificance of the regression coefficient for "paper" (Table 1) does not indicate that this material type does not have a relevant influence on the K_{ma}. As a result, we keep all 21 material type dummy variables in the final regression to retain as much information as possible.

To better illustrate the impact of each predictor on the material-air partition coefficient, we varied each predictor from the minimum to the maximum value in the entire dataset (991 data points) while keeping the other predictors constant, and calculated the change in logK_{ma} using the regression coefficients in the final QSPR (Equation 6). Since the chemical's ΔH_v determines the ΔH_{ma} which modifies the relationship between logK_{ma} and temperature, the impact of temperature was calculated as two extremes using the minimum and maximum values of ΔH_v in the entire dataset. As shown in Figure 4, the chemical's logK_{oa} has the highest impact on logK_{ma} among predictors. The impact of temperature on logK_{ma} is very low with the lowest value of ΔH_v (22.3 kJ/mol), but the impact become moderate with the highest value of ΔH_v (75.6 kJ/mol). This indicates that for a chemical with low enthalpy of vaporization, the logK_{ma} only changes slightly with temperature, and vice versa. The material type also has a moderate impact on the logK_{ma}, which is similar to the impact of temperature with the highest value of ΔH_v. Overall, the impact of material type is relatively small compared to the impact of chemical's logK_{oa}, indicating that the variation in logK_{ma} does not strongly depend on the solid material type, which suggests the possibility of developing a generic QSPR to predict logK_{ma} in absence of material-specific data.

3.4 | Model validation results

3.4.1 | Internal validation

The correlation coefficient for the LMO cross-validation, Q^2_{LMO}, averages 0.93 (range: 0.90-0.95) for the 1000 iterations, and the root mean square error for cross-validation ($RMSE_{cv}$) averages 0.63. Both the Q^2_{LMO} and $RMSE_{cv}$ are similar to the R^2 and RMSE computed using the full dataset, which is 0.93 and 0.62, respectively, indicating that the model is internally stable.

For Y-scrambling, the R^2_{Yscr}, Q^2_{Yscr} and $RMSE_{Yscr}$ for the 1000 iterations average 0.023, −0.028, and 2.37, respectively, which are

substantially different from the R^2, Q^2_{LMO} and RMSE of the original model, indicating that that no correlation exists between the scrambled responses and the predictors. Thus, the internal validation overall demonstrates that the final QSPR model (Equation 6) is robust and stable, and is not a result of chance correlation.

3.4.2 | External validation

As described in Section 2.3.2, four types of splitting were used for external validation, including splitting by random 20%, by ordered response, by structure, and by studies. Six criteria for external validation, described in detail previously,[1,22,23] were computed and are presented in Table 2. For the first three types of splitting, the R^2_{ext} are higher than 0.9, and the other five criteria all pass the threshold values and are higher than 0.9, indicating good predictive ability of the models constructed from training set data. This is expected because the prediction sets resulted from these three types of splitting are generally well within the applicability domain (described in detail below) defined by the training sets (Figures S1-S6), since the data points were drawn either randomly or alternately.

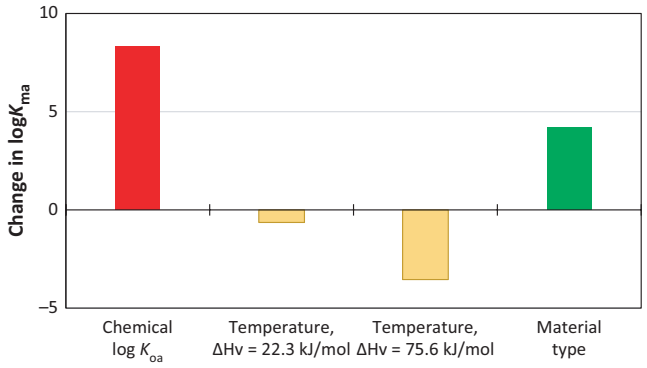

FIGURE 4 Change in logK_{ma} with respect to the change in each predictor, from minimum to maximum values within the entire dataset

For the splitting by studies, data from 22 studies were selected as the prediction set, while data from 20 studies constituted the training set. This splitting can better represent a truly "external" validation, since all data from one study were either be in the training set or be in the prediction set. The prediction ability of the model constructed from the training set is apparently reduced, as the R^2_{ext} of this splitting dropped to 0.79, and the values of the other five criteria are lower than those for the above three types of splitting. This is reasonable since the data variability is higher between studies than within studies, so the prediction set might not be well within the AD defined by the training set (Figures S7-S10). Nonetheless, all validation criteria for this splitting still pass the thresholds, indicating acceptable prediction ability (Table 2).

3.4.3 | Applicability domain

It is important to define the Applicability domain (AD) of our QSPR model, as it can provide information on the reliability of the model predictions[24] for future users who would like to use the model on new chemicals. If the new chemicals are inside the AD, the model predictions are interpolated and are more reliable. However, if the chemicals are outside the AD, the predictions are extrapolated and less reliable.[24]

For definition of the AD, the model being evaluated is the final QSPR model presented in Equation (6), and the training dataset thus refers to the full dataset including 991 data points. Three complementary methods were applied to define the AD of the K_{ma} QSPR: the range of model predictors, the leverage approach, and the PCA of the model predictors, which have been described in detail previously.[25]

For the range of predictors, the model has four predictors: logK_{oa}, ΔH_v, temperature, and material type. The logK_{oa}, ΔH_v, temperature of the training dataset range from 1.4 to 14.6, from 22.3 to 75.6 kJ/mol, and from 15 to 100°C, respectively, defining the AD of the model. It is noteworthy that the material type is a categorical variable, and the training set contains 22 consolidated materials types,

TABLE 2 External validation results

External validation criteria	R^2_{ext}	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	$\overline{r^2_m}$	CCC
Threshold		>0.70	>0.70	>0.70	>0.65	>0.85
Splitting by random percentage	0.93	0.93	0.93	0.92	0.90	0.96
Splitting by ordered response	0.93	0.93	0.93	0.93	0.90	0.96
Splitting by ordered structure	0.94	0.94	0.94	0.94	0.91	0.97
Splitting by studies	0.79	0.86	0.78	0.86	0.71	0.89

R^2_{ext}: determination coefficient of the prediction set external data.
Q^2_{F1}: correlation coefficient proposed by Shi et al.
Q^2_{F2}: correlation coefficient proposed by Schuurmann et al.
Q^2_{F3}: correlation coefficient proposed by Consonni et al.
$\overline{r^2_m}$: determination coefficient proposed by Ojha et al.
CCC: concordance correlation coefficient proposed by Chirico and Gramatica.

so the model's AD is also restricted to these 22 material types. For the leverage approach, the critical value h^* for the diagonal values of the hat (h) matrix of the model was calculated to be 0.0727, and the AD is defined as the h values less than h^*.[21,25] For the PCA approach, the AD is defined as the space between the minimum and maximum values of the PC1 and PC2 scores of the training dataset,[21,25] which range from −4.39 to 2.04 and from −4.52 to 2.22, respectively. For future model users, a new chemical should be considered "inside AD" if viewed inside AD by all three methods, and be considered "outside AD" if viewed out of AD by all three methods, otherwise it should be considered "borderline".[25]

3.5 | Generic QSPR

In order to predict the K_{ma} without assigning material properties, we built a generic QSPR model which does not include any material-specific variables using the same dataset. This model only uses the chemical properties and temperature as predictors and is as follows:

$$\log_{10} K_{ma} = -0.37 + 0.75 \cdot \log_{10} K_{oa} + 1.29 \cdot \frac{\Delta H_{ma}}{2.303 \cdot R}\left(\frac{1}{T} - \frac{1}{298.15}\right) \quad (7)$$

N = 991, R^2 = 0.80, R^2_{adj} = 0.80, SE = 1.08, RMSE = 1.08

ANOVA: F = 1943, df = 2, P < 0.0001.

This model has a still relatively high adjusted R-squared of 0.80 compared to the 0.93 of the regression with material coefficient (Equation 6), indicating a good fit of experimental data (Figure 5). As discussed in Section 3.2, the impact of the solid material type on logK_{ma} is relatively small compared to the impact of chemical properties, so logK_{ma} can be predicted with reasonably high accuracy without the material type as a predictor. This generic QSPR thus provides a relatively reliable method to estimate the K_{ma} for various solid materials that may be difficult to assign a material type listed in Table 1, which provides a more comprehensive and flexible coverage, although with a slightly lower accuracy, for different chemical-material combinations than the material-specific QSPR and can therefore greatly facilitate high-throughput evaluations of a large variety of chemical-material combinations. However, it should be noted that although without the material type as a predictor, this generic model was still developed using the experimental data of our collection of 22 material types. Thus, this generic model best applies to materials listed in Table 1 and similar materials, but may cause a large error for materials with special properties, for example in presence of strong ionic forces, or of strong pseudo-solvation such that some of the target adsorbate molecules take on a different structure within the material itself, either due to ionization or tautomerization.

3.6 | Limitations and future work

While the coverage of 22 consolidated materials and possibly any solid material as well as inclusion of the effect of temperature are

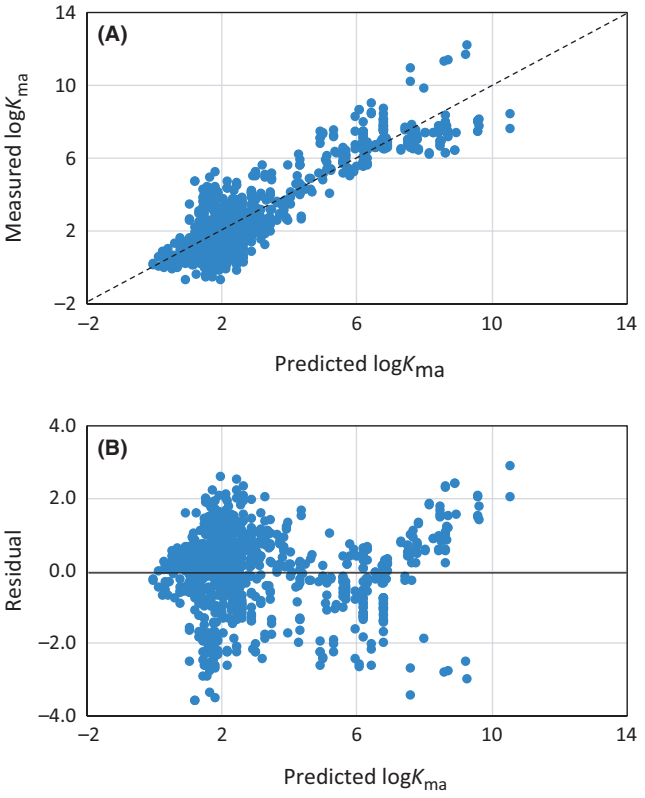

FIGURE 5 A, measured logK_{ma} and B, residuals as a function of logK_{ma} predicted by the generic QSPR (Equation 7). The dotted line in (A) indicates the 1:1 line

major advantages, the present model has several limitations. First, the model does not consider chemical ionization or interaction with other chemicals within a solid material, which may affect the chemical's partitioning between the material and air. Second, the present model assumes that the relationship between ΔH_{ma} and chemical's ΔH_v, derived from experimental ΔH_{ma} data for one material type "PU-ether", is the same across different material types. Ideally, more experimental ΔH_{ma} data for different material types are needed to verify this assumption or to develop unique ΔH_{ma}-ΔH_v relationships for different material types.

Third, since for most K_{ma} datasets the material properties are not well characterized or provided in the original publications, the classification of the consolidated material types is qualitative and is simply based on material names, which may result in considerable variations in material properties within one consolidated material type. In addition, even with the same composition, different material structure may affect the material-air partitioning. Ideally, quantitative, continuous properties of the solid materials, such as descriptors of the material's composition and molecular structure, could be measured and entered into the model as numerical predictors, so that the model can be more accurate for particular materials and can be extrapolated to other material types outside the training dataset. In addition, if quantitative variables for material types are used, interaction terms between chemical's logK_{oa} and material type variables

can be added to the model without introducing too many additional terms, which can improve model fitting, as discussed in Section 3.2.

Fourth, many materials that appear in indoor environments are inhomogeneous, such as plywood, gypsum board, carpet, concrete, and paper, which may have layers or portions with distinctive properties. Thus, the K_{ma} values measured in experiments and the QSPR built on these measurements likely only represent the material properties across the experiments. As a result, one needs to use caution when applying the present QSPR to predict K_{ma}, especially for highly inhomogeneous materials. Another important aspect related to heterogenicity is surface partitioning versus bulk partitioning. Since the partitioning between solid material and air happens mainly at the material surface, the surface properties may have an unusually large influence on the apparent partitioning behavior. Therefore, for materials with a surface layer of distinct properties, or materials with the same composition but different surface/bulk structures, the present QSPR may not give a correct estimate of the K_{ma}. The distinct surface layer may be a result of oxidative aging and soiling, which may change with time, or intrinsic features that are time invariant. These problems again highlight the importance of using quantitative descriptors of material compositions and structures as predictors in the QSPR.

Finally, the functional mechanisms of other influence factors such as relative humidity are unclear, so they are not included in the QSPR. The effect of relative humidity on K_{ma} is likely both chemical and material dependent,[4,9] which will require more in-depth research.

4 | CONCLUSIONS

A multiple linear regression model has been developed to predict the solid material-air partition coefficients (K_{ma}) of organic compounds in various solid materials. Experimental K_{ma} data collected from 43 studies were used to construct the regression model. The model uses three continuous variables, chemical's logK_{oa}, ΔH_v, and absolute temperature, as well as one categorical variable, material type, as predictors. The model has been validated internally and externally to be robust and stable, and have good predicting ability. The applicability domain of the model, in terms of the range of predictors, includes chemical's logK_{oa} between 1.4 and 14.6, ΔH_v from 22.3 to 75.6 kJ/mol, temperature between 15 and 100°C, and material type belonging to the 22 consolidated types.

The main advantage of the present model is that it is applicable for a wide range of chemical-material-temperature combinations, which is more comprehensive than the correlation methods developed in previous studies which were specific for one solid material and often at room temperature. Moreover, a generic model is also developed which is able to give relatively accurate estimates of K_{ma} without assigning a particular material type, making it suitable for high-throughput assessments of the chemical releases from solid materials and subsequent consumer exposures.

ACKNOWLEDGEMENTS

This research partly benefited from the support of the Long Range Research Initiative of the American Chemistry Council and the US EPA (contract EP-16-C-000070).

ORCID

Lei Huang http://orcid.org/0000-0002-7846-9760

REFERENCES

1. Huang L, Fantke P, Ernstoff A, et al. A quantitative property-property relationship for the internal diffusion coefficients of organic compounds in solid materials. *Indoor Air*. 2017;27(6):1128-1140.
2. Huang L, Jolliet O. A parsimonious model for the release of chemicals encapsulated in products. *Atmos Environ*. 2016;127:223-235.
3. Liu Z, Ye W, Little JC. Predicting emissions of volatile and semivolatile organic compounds from building materials: a review. *Build Environ*. 2013;64:7-25.
4. Kamprad I, Goss K-U. Systematic investigation of the sorption properties of polyurethane foams for organic vapors. *Anal Chem*. 2007;79(11):4222-4227.
5. Bartkow ME, Hawker DW, Kennedy KE, et al. Characterizing uptake kinetics of PAHs from the air using polyethylene-based passive air samplers of multiple surface area-to-volume ratios. *Environ Sci Technol*. 2004;38(9):2701-2706.
6. Kennedy KE, Hawker DW, Müller JF, et al. A field comparison of ethylene vinyl acetate and low-density polyethylene thin films for equilibrium phase passive air sampling of polycyclic aromatic hydrocarbons. *Atmos Environ*. 2007;41(27):5778-5787.
7. Morrison G, Li H, Mishra S, et al. Airborne phthalate partitioning to cotton clothing. *Atmos Environ*. 2015;115:149-152.
8. Morrison G, Shakila N, Parker K. Accumulation of gas-phase methamphetamine on clothing, toy fabrics, and skin oil. *Indoor Air*. 2015;25(4):405-414.
9. Zhao D, Little JC, Cox SS. Characterizing polyurethane foam as a sink for or source of volatile organic compounds in indoor air. *J Environ Eng*. 2004;130(9):983-989.
10. Bodalal A, Zhang J, Plett E, et al. Correlations between the internal diffusion and equilibrium partition coefficients of volatile organic compounds (VOCs) in building materials and the VOC properties. *ASHRAE Trans*. 2001;107:789.
11. Guo Z. Review of indoor emission source models. Part 2. Parameter estimation. *Environ Pollut* 2002;120(3):551-564.
12. Chaemfa C, Barber JL, Gocht T, et al. Field calibration of polyurethane foam (PUF) disk passive air samplers for PCBs and OC pesticides. *Environ Pollut*. 2008;156(3):1290-1297.
13. Shoeib M, Harner T. Characterization and comparison of three passive air samplers for persistent organic pollutants. *Environ Sci Technol*. 2002;36(19):4142-4151.
14. Holmgren T, Persson L, Andersson PL, et al. A generic emission model to predict release of organic substances from materials in consumer goods. *Sci Total Environ*. 2012;437:306-314.
15. Booij K, Hofmans HE, Fischer CV, et al. Temperature-dependent uptake rates of nonpolar organic compounds by semipermeable membrane devices and low-density polyethylene membranes. *Environ Sci Technol*. 2003;37(2):361-366.
16. Adams RG, Lohmann R, Fernandez LA, et al. Polyethylene devices: Passive samplers for measuring dissolved hydrophobic

organic compounds in aquatic environments. *Environ Sci Technol*. 2007;41(4):1317-1323.
17. Liu Y, Zhou X, Wang D, et al. A prediction model of VOC partition coefficient in porous building materials based on adsorption potential theory. *Build Environ*. 2015;93:221-233.
18. Zhang Y, Luo X, Wang X, et al. Influence of temperature on formaldehyde emission parameters of dry building materials. *Atmos Environ*. 2007;41(15):3203-3216.
19. USEPA. *Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11*. Washington, DC: United States Environmental Protection Agency; 2012.
20. Gramatica P, Cassani S, Chirico N. QSARINS-chem: insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. *J Comp Chem*. 2014;35(13):1036-1044.
21. Gramatica P, Chirico N, Papa E, et al. QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. *J Comp Chem*. 2013;34(24):2121-2132.
22. Chirico N, Gramatica P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J Chem Inf Model*. 2012;52(8):2044-2058.
23. Chirico N, Gramatica P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J Chem Inf Model*. 2011;51(9):2320-2335.
24. Gramatica P. Principles of QSAR models validation: internal and external. *QSAR Comb Sci*. 2007;26(5):694-701.
25. Cassani S, Gramatica P. Identification of potential PBT behavior of personal care products by structural approaches. *Sustain Chem Pharm*. 2015;1:19-27.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Huang L, Jolliet O. A quantitative structure-property relationship (QSPR) for estimating solid material-air partition coefficients of organic compounds. *Indoor Air*. 2019;29:79-88. https://doi.org/10.1111/ina.12510