DR. LEI HUANG (Orcid ID : 0000-0002-7846-9760)
DR. PETER FANTKE (Orcid ID : 0000-0001-7148-6982)

Article type : Original Article

A Quantitative Property-Property Relationship for the Internal Diffusion Coefficients of Organic Compounds in Solid Materials

Lei Huang[1]*, Peter Fantke[2], Alexi Ernstoff[2] and Olivier Jolliet[1]

[1]Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Abor, MI, USA

[2]Division for Quantitative Sustainability Assessment, Department of Management Engineering, Technical University of Denmark, 2200 Kgs. Lyngby, Denmark

*Corresponding author, huanglei@umich.edu

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/ina.12395

Abstract

Indoor releases of organic chemicals encapsulated in solid materials are major contributors to human exposures and are directly related to the internal diffusion coefficient in solid materials. Existing correlations to estimate the diffusion coefficient are only valid for a limited number of chemical-material combinations. This paper develops and evaluates a quantitative property-property relationship (QPPR) to predict diffusion coefficients for a wide range of organic chemicals and materials. We first compiled a training dataset of 1103 measured diffusion coefficients for 158 chemicals in 32 consolidated material types. Following a detailed analysis of the temperature influence, we developed a multiple linear regression model to predict diffusion coefficients as a function of chemical molecular weight (MW), temperature, and material type (adjusted R^2 of 0.93). The internal validations showed the model to be robust, stable and not a result of chance correlation. The external validation against two separate prediction datasets demonstrated the model has good predicting ability within its applicability domain ($R^2_{ext} > 0.8$), namely MW between 30 and 1178 g/mol and temperature between 4 and 180 °C. By covering a much wider range of organic chemicals and materials, this QPPR facilitates high-throughput estimates of human exposures for chemicals encapsulated in solid materials.

Keywords

Diffusion, Solid materials, Consumer products, Indoor release, Organic chemicals, Correlation

Practical implications

The quantitative property-property relationship developed by the present study provides a more comprehensive correlation method to estimate the diffusion coefficients, as it covers a wide range of organic chemicals and solid materials, and also considers the effect of temperature. This model provides the basis for facilitating high-throughput estimates of indoor human exposures for chemicals encapsulated in solid materials relevant for several science-policy fields, such as chemical alternatives assessment (CAA), risk assessment (RA) and life cycle assessment (LCA).

1. Introduction

Chemicals encapsulated in solid materials have been identified as a major source of passive emissions to indoor air [1-3] and of transfers into food [4] and onto skin [5]. Typical examples include

chemicals used as flame retardants in furniture and plasticizers in food contact materials. To estimate the release of these chemicals from solid materials, and eventually consumer exposures, the diffusion coefficient, D (m^2/s), for chemicals encapsulated in solid materials, is essential information. D describes the transport of a molecule through a material, which is specific for a chemical-material combination and is also influenced by ambient temperature. Experimental techniques such as chamber tests for building materials [6, 7], and sorption/desorption experiments for polymer materials [8-10] have enabled measurement of a limited number of chemical diffusion coefficients for building materials such as vinyl flooring, gypsum board, particle board, plywood, carpet and cement [11-14], as well as polymer materials including polyethylene (PE), polystyrene (PS), polypropylene (PP), and polyvinyl chloride (PVC) [4, 15, 16]. However, given the limited number of chemical-material combinations with measured Ds, and the costly and time-consuming nature of experiments, quantitative relationships are needed to complement existing measurements by predicting the diffusion coefficients from known physiochemical properties for chemicals without experimental data. This is especially important for high-throughput approaches where a large number of chemical-material combinations need to be evaluated and for which it is unrealistic to perform experiments on all relevant combinations.

Several correlation methods have been developed to estimate the diffusion coefficients from physicochemical properties of chemicals [8, 12, 17-19]. For example, Berens and Hopfenberg correlated the D to the mean molecular diameter of the diffusing molecule, using data on more than 20 chemicals in 3 glassy materials including PVC, PS and polymethyl methacrylate (PMMA) [8]. Zhao et al. found a correlation between D and vapor pressure for water and 8 aromatic hydrocarbons in polyurethane foam (PUF) [19]. Furthermore, both Bodalal et al. and Cox et al. estimated the D as a function of molecular weight [12, 18]. The former study considered measured D data on 5 aromatics and 5 aldehydes in several building materials [12], while the latter study considered data on 4 alkanes in vinyl flooring [18]. For each of these aforementioned approaches, the main limitation is that the correlations are specific to certain chemical classes and materials; for example aldehydes in plywood, which limits their application for other materials and chemical classes. Addressing this research gap to facilitate wider applicability, Guo developed a method which estimates the diffusion coefficient as a function of the chemical's molar volume for mixed chemical classes [17]. However, this approach is limited to 6

building materials and are developed based on a small dataset of limited chemical classes (≤ 3 chemical classes for 5 of the 6 building materials).

The aforementioned correlation methods consider experiments for building materials at room temperature, and therefore temperature is not relevant and thus not considered in the correlation model. For other exposure scenarios, such as transfer of chemicals from food contact materials (FCMs) into food, ambient temperature is highly relevant because FCMs can be heated, refrigerated, or frozen. Accordingly, Begley et al. presented a correlation method to estimate the diffusion coefficient in 9 polymer materials as a function of molecular weight and temperature [4], which is not applicable beyond the considered polymers.

In all, the currently available correlation methods to estimate *D* do not provide sufficient coverage of chemicals encapsulated in consumer products in different use scenarios (i.e. ambient temperatures). Developing low-tier, high-throughput methods to estimate exposure to chemical in consumer products across a variety of chemical-material combinations is a recent focus in various science-policy fields such as computational exposure science and life cycle assessment (LCA) [20-25]. Addressing the lack of methods to estimate *D* for a variety of chemical-product scenarios, the present study aims to develop a more comprehensive correlation method to estimate *D* for wide range of organic compounds in multiple solid materials. More specifically, we aim to:

1) Carry out a comprehensive and extensive literature review to collect experimental diffusion coefficient data on a wide range of materials and chemicals.

2) Use multiple linear regression techniques to establish the relationship between the diffusion coefficient and various predictor variables including physiochemical properties, material properties and environmental characteristics.

3) Perform internal and external validations to characterize the validity and predictive power of the developed correlation.

Since the material type is a categorical property variable and is not related to the chemical's molecular structure, we call this correlation a quantitative property-property relationship (QPPR) instead of a quantitative structure-activity relationship (QSAR). This QPPR provides a more advanced correlation method to estimate the diffusion coefficients of organic compounds compared to previous studies, as it covers a wide range of solid materials and physiochemical properties, and also considers the effect of temperature. By providing reliable estimates of this

key diffusion parameter for a large number of chemicals, this method will facilitate high-throughput assessments of chemical emissions and human exposures for chemicals encapsulated in solid materials relevant for chemical alternatives assessment (CAA), risk assessment and LCA.

2. Materials and methods

2.1 Dataset

Experimental diffusion coefficient data were compiled from 68 references from the peer-reviewed scientific literature. The initial dataset contained a total of 1124 records covering 161 unique chemicals and 88 distinct solid materials (provided in Supporting Info). Experimental data expressed in cm^2/s were converted to m^2/s. There are different types of diffusion coefficients reported in the literature, so harmonization of these data was performed to develop a consistent dataset. For diffusion coefficients measured in liquid sorption experiments, the 'intrinsic' diffusion coefficients, corrected for the swelling of materials were collected [10]. Sorption of the liquid molecules inside the solid material may cause swelling of the material, which would lead to decreased observed diffusion coefficients and thus need to be corrected [10]. For porous materials consisting of pore space and solid material, two types of models can be used to describe the chemical transport through these materials. The one-phase model considers the porous material as an assumed homogeneously mixed material, so an 'apparent' diffusion coefficient is used to describe the chemical diffusion through such imaginary material [7]. In contrast, the multi-phase model considers the material as a mixture of pores and solid parts, and the chemical diffuses mainly through the pores if the pores are interconnected, or through the pores and solid parts alternately if the pores are isolated from each other. The gas-phase diffusion through the pores, which can be described by an 'effective' diffusion coefficient, is assumed to be much faster than the diffusion through the solid parts [7]. Haghighat et al., [7] has demonstrated that the 'apparent' diffusion coefficient is equivalent to the 'effective' diffusion coefficient (D_e) divided by the material phase-gas phase partition coefficient (K_{ma}). Thus, for porous materials the 'apparent' diffusion coefficients reported in studies were collected [26]. For studies where only the D_e and K_{ma} were reported [27-29], they were converted to 'apparent' diffusion coefficients using the aforementioned method. Data were excluded for studies where only the 'effective' diffusion coefficients were reported.

From the initial dataset, 21 records were excluded from further analyses because they involve chemicals that are inorganic, chemicals for which no CAS number could be identified, or

chemicals that are polymer chains with varying molecular weights. The final considered dataset thus includes 1103 records for 158 unique chemicals and 87 materials.

2.2 Modeling methods

2.2.1 Multiple linear regression

A multiple linear regression (MLR) analysis was performed to identify and quantify the effect of different parameters on the diffusion coefficient. The MLR model takes the following general form:

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

where $\log_{10} D$ is the logarithm of the diffusion coefficient (m^2/s), α is the intercept; X_1 to X_n are independent variables related to physiochemical properties, such as molecular weight, molar volume, and vapor pressure, and/or environmental characteristics like temperature; β_1 to β_n are regression coefficients for the respective independent variables X_1 to X_n; and M_1 to M_m are dummy variables for the solid 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 1, since the material type with the highest number of measured D data is used as the reference material type and does not require a dummy available in the MLR. Note that the MLR model gives one coefficient for each material type, while a material type can represent a single pure substance such as calcium silicate, a composite material such as vinyl flooring and gypsum board, or a group of similar materials such as wooden boards. Details of the material types will be discussed later. This regression equation also implies that the material coefficients (b_1 to b_m) and the physiochemical property coefficients (β_1 to β_n) are independent of each other, which if corroborated by internal and external validations (Section 2.3), allow for the maximum prediction coverage in terms of chemical-material combinations. All 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).

2.2.2 Grouping of materials and initial regressions

To reduce the number of dummy variables, to avoid over-fitting of the MLR model, and to have a minimum of 10 records and 3 different chemicals per material type to ensure enough variability,

the 87 original materials were grouped into 32 consolidated material types, based on the similarity of the regression coefficients and the material types (see Supporting Information (SI), Section S1). Thus m = 31 in Eq. 1, with PET as the 32nd and reference material, since it is the material with most reported diffusion coefficients.

In previous studies, either the chemical's molecular weight (*MW*), molar volume (*MV*) or vapor pressure (*VP*) has been used as predictor of the diffusion coefficient in a given material [12, 17-19]. Begley et al. [4] also suggested that the logarithm of the diffusion coefficient varies linearly with the inverse of the absolute temperature (*1/T*). Thus, the initial regression was performed to identify which of the above variables (*MW*, *MV*, *VP* and *1/T*) are best predictors of the diffusion coefficients of compounds encapsulated in the 32 material types, i.e., to identify X_1 to X_n in Eq. (1). Details of the initial regression process are presented in SI, Section S2. Results of the initial regression model suggest that the log-molecular weight and the inverse of the absolute temperature are the most important predictors, and therefore the employed MLR model takes the following form:

$$\log_{10} D = \alpha + \beta_{\log_{10} MW} \cdot \log_{10} MW + \beta_{1/T} \cdot \frac{1}{T} + b_1 \cdot M_1 + \cdots + b_m \cdot M_m \quad (2)$$

where *MW* is the chemical's molecular weight (g/mol) and *T* is the absolute temperature (K). The model performance of using log-molecular weight and molecular weight as predictors were very close when using the training dataset (1103 records, m=31), but the model using log-molecular weight as predictor was finally selected since it performs better for high-molecular-weight chemicals (Section 3.3.3).

2.2.3 Temperature dependence

Studies have shown that the activation energy of diffusion is a contributor to the temperature dependence of the diffusion coefficient and varies as function of both the material and the chemical properties [4, 30, 31]. Thus, ideally a specific temperature correction coefficient should be used for each chemical-material combination. Since data availability is not sufficient to determine chemical-specific temperature coefficients for each of the 32 materials, and since chemical properties seem to have limited influence on the activation energy [4, 30], we followed the strategy of Begley et al. [4], differentiating temperature coefficients for a limited number of material groups, applying one generic temperature coefficient for all chemicals within each material group. Begley et al. [4] have introduced a variable τ to adjust the temperature coefficient for two groups of materials, where τ equals 0 or 1577 for 9 different polymers, which

corresponds to activation energy of 86.9 kJ/mol for e.g. LDPE or 100 kJ/mol for e.g. HDPE. To analyze the temperature dependency of the diffusion coefficients in our dataset, we first plotted $log_{10}D$ against $1/T$ for each of the 32 material types (SI Section S3). The plots generally show as expected [4] an inverse relationship in which $log_{10}D$ is decreasing with increasing $1/T$, different materials exhibiting different slopes. Since variability in diffusion coefficient is higher between than within given studies, we first determined a temperature coefficient for each chemical-material-study combination, and then calculated an average temperature coefficient for each material type by averaging all temperature coefficients belonging to the same material type. The analysis of the material-specific temperature coefficients showed that the materials can be grouped into three categories: (1) high-, (2) medium- and (3) low-coefficient categories, with three corresponding values for the temperature coefficient adjustment factor τ, which are given in Section 3.1. Details are presented in SI Section S3.3. The adjusted MLR model takes the following form accordingly:

$$\log_{10}D = \alpha + \beta_{\log MW} \cdot \log_{10}MW + \frac{\beta_{1/T}+\tau}{T} + b_1 \cdot M_1 + \cdots + b_m \cdot M_m, \qquad (3)$$

2.2.4 Final regression

To avoid multicollinearity problems in the MLR model and to avoid the influence of the material type "Limited-data material group" on the temperature coefficients, we fixed the temperature coefficients determined using Eq. 3 and thus the final regression takes the following form:

$$\log_{10}D - \frac{\beta_{1/T}+\tau}{T} = \alpha + \beta_{\log MW} \cdot \log_{10}MW + b_1 \cdot M_1 + \cdots + b_m \cdot M_m, \qquad (4)$$

where the dependent variable is $\log_{10}D - (\beta_{1/T}+\tau)/T$ instead of $\log_{10}D$, with the values of $\beta_{1/T}$ and τ obtained from Eq. 3 and presented later in Section 3.1. In this final regression, all 1103 records of measured D data were utilized including the material type "Limited-data material group", leading to m=31 material types, plus one reference material type, PET, with $b_{PET} = 0$.

2.3 Model validation

Validation of the final MLR model (Eq. 4) was performed using the QSARINS software, version 2.2.1 (www.qsar.it) which is developed by Gramatica et al. [32, 33].

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 for internal validation in QSARINS. The first one is the

leave more out (LMO) cross-validation technique, which iteratively and randomly exclude a certain percentage of the measured diffusion coefficient data, and then computes the regression coefficients with the remaining data and uses those coefficients to make predictions for the excluded ones [33]. We used 1000 iterations and the percentage of the excluded elements was set as 20%.

The second technique for internal validation is the Y-scrambling procedure, which demonstrates that the model is not the result of chance correlation. In this procedure, the experimental responses (in our study, the temperature-adjusted diffusion coefficients) are shuffled at random and used with the original predictors to establish an MLR model. If the original MLR model is internally valid, the performances of the scrambled models should be much worse than the original model[33]. We used 1000 iterations for the Y-scrambling.

2.3.2 External validation

We also evaluated the model ability to provide reliable predictions on new datasets in a so-called external validation process, using the following two approaches.

The first approach was to split the existing dataset (1103 records) into one training dataset and one prediction datasets. 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 SI, Section S5.1 for details) by random percentage (20% of the entire dataset randomly selected as the prediction set, 80% rest to the training set), by response and by structure (data first ordered by responses of the temperature-adjusted diffusion coefficient, or by the first axis of principal component analysis (PCA) of the descriptors, respectively). We introduced a fourth kind of splitting by studies, since variability across studies for a given material is in general larger than variability within a given study, yielding similar sample sizes of approximately 880 data for the training set and 220 data for the prediction set (SI, Table S3).

The second approach of external validation was to use the entire collected dataset (1103 records) as the training set and to use an entirely separate dataset as the prediction set. For the prediction set, two datasets were used. The first one is a database of diffusion coefficients from the United States Food and Drug Administration (FDA), which is a “database available upon request" for guidance for industry (http://www.fda.gov/Food/ucm081818.htm), and includes non-peer

reviewed diffusion coefficient data reported by industry. This dataset includes 191 records of experimental diffusion coefficients of 46 chemicals in 22 materials which are mainly polymers used for food contact materials (see SI, Section S5.1 for details). The quality and reliability of these data are not characterized by FDA. The second prediction dataset is constructed from several studies conducted before 1982 [34-36], referenced in[37]. This dataset, designated as "Data by 1982", includes 281 records of measured diffusion coefficients of 92 chemicals in 8 polymer materials, also including self-diffusion (see SI, Section S5.1 for details). Data for both prediction sets are provided in Supporting Info.

2.3.3 Applicability domain (AD)

The analysis and definition of the applicability domain (AD) of models is a fundamental issue that must be addressed in QSAR and QPPR studies. The study of AD can provide information on the reliability of the model predictions, i.e., if the chemicals are inside the AD, the predictions are interpolated and are more reliable; if the chemicals are outside the AD, the predictions are extrapolated and less reliable, because effects can occur outside the AD that do not exist within the AD [38]. Three complementary methods were applied to define the AD of the diffusion coefficient QPPR: the range of model predictors, the leverage approach, and the PCA of the model predictors [39]. More explanation of these methods is provided in SI, Section S4. In our analysis, chemicals are considered inside the AD if they are viewed inside AD by all three methods, whereas chemicals are considered outside AD if they are viewed outside AD by all three methods, and finally chemicals that fall inside the AD for only one or two methods are considered as 'borderline.'

3. Results and discussion

3.1 Temperature dependence of the diffusion coefficient

The compiled dataset of 1103 records including 158 chemicals and 32 material types shows that the diffusion coefficient in solid materials decreases with decreasing temperature, as demonstrated by the highly significant negative regression coefficient for the variable *1/T*, with $\beta_{1/T} = -4440\ (K)$ with a standard error (SE) of 164 (K) and $p < 0.001$ in Eq. 2 (SI, Section S3.1). This is in agreement with previous studies [4, 30, 31]. This general tendency of decreasing diffusion with increasing *1/T* is well illustrated by the example of PET, the material with the most data available (Figure 1A – see SI, Figure S1 for other materials). To further refine the

coefficient for the temperature variable into specific materials groups, Figure 1B illustrates well for methyl methacrylate (MMA) homopolymer the importance of first determining a temperature coefficient for each separate study and material-chemical combination (Section 2.2.3) and then averaging the temperature coefficients across studies. The molecular weight-normalized diffusion coefficients show a negative linear relationship with *1/T* within each of the three experimental studies of Figure 1B[40-42], with similar regression coefficients of -4530 (K), -5704 (K), -3415 (K), averaging -4550 (K) with an SE of 305 (K) . However, since the absolute $\log_{10}MW$-normalized diffusion coefficients reported by Hennebert et al. [42] are much higher than those reported by the other two studies, doing one regression with all data from the three studies would result in a non-significant temperature coefficient (p-value of 0.19), thus demonstrating the importance to first perform temperature regressions using data from the same study and for the same chemical.

Table 1 presents the average temperature coefficients and their standard errors for each of the 32 consolidated material types. Based on the values of the temperature coefficients (unit in K), the 32 material types can be grouped into three categories: (1) high-coefficient category with relatively high (absolute value) temperature coefficients (< -5000), i.e., materials in which diffusion coefficients are highly sensitive to the change in temperature, (2) medium-coefficient category with temperature coefficients in between (-5000 < $(\beta_{1/T} + \tau)$ < -3000), and (3) low-coefficient category with relatively low (absolute value) temperature coefficients (> -3000), i.e., materials in which diffusion coefficients are least sensitive to the change in temperature. Details for the grouping of temperature coefficients can be found in SI, Section S3.3.

The temperature coefficients $\beta_{1/T}$ and τ used in Eq. 4 for each of the three temperature-dependency material categories are obtained from the regression using the MLR model of Eq. S3-2 (SI, Section S3.3), yielding values of $\beta_{1/T} = -3486 \pm 299\ (K)$ and $\tau_{\text{high}} = -2391 \pm 356\ (K), \tau_{\text{medium}} = 0\ (K)$ and $\tau_{\text{low}} = +1676 \pm 510 (K)$. Thus, for the High-, Medium- and Low-coefficient categories, the final temperature coefficients ($\beta_{1/T}$ + τ) are -5877 (K), -3486 (K), and -1810 (K), corresponding to activation energy of 113, 66.7 and 34.7 (kJ/mol), respectively. Begley et al. [4] also aggregated 9 types of polymer materials into two temperature categories, with activation energy of 100 and 86.9 (kJ/mol), which have similar values with the high- and medium-coefficient categories in the present paper, to which these 9 polymer materials are assigned. These results indicate that the categorization of the temperature coefficient in the

present paper is consistent with previous studies, while extending the QPPR to a wider range of materials.

3.2 Final QPPR and model fitting

Using the full dataset (1103 records) and Eq. 4, the final MLR model for predicting the diffusion coefficient in solid materials is as follows:

$$\log_{10}D - \frac{\tau-3486}{T} = 6.39 - 2.49 \cdot \log_{10}MW + b \quad (5)$$

$$N = 1103,\ R^2 = 0.932,\ R^2_{adj} = 0.930,\ SE = 1.17,\ RMSE = 1.15$$

$$ANOVA:\ F = 457,\ df = 32,\ p < 0.0001$$

where D is the diffusion coefficient (m^2/s), MW is molecular weight (g/mol), T is absolute temperature (K), b and τ (K) are the material-specific coefficients presented in Table 2. This model is provided as an excel model in Supporting Info to facilitate application. The standard errors for the intercept (6.39) and the coefficient of $\log_{10}MW$ (-2.49) are 0.29 and 0.13, respectively. An SE of 1.17 of the final model (Eq. 5) indicates that the 95% confidence interval (CI) of the predicted response, $\log_{10}D$-$(\tau$-$3486)/T$, is the predicted value ± 2.30. The 95% CI of the $\log_{10}D$ cannot be directly calculated, but the average absolute difference between predicted and measured $\log_{10}D$ is 0.83 across the whole dataset (1103 records), and 95% of this absolute difference is below 2.54.

This MLR model shows excellent fitting of the experimental data, with an adjusted R-square of 0.932 and a root mean square error (RMSE) of 1.15. The model fit is highly significant with an ANOVA p-value smaller than 0.0001. Figure 2 shows the scatter plot of experimental versus predicted responses, which aligns well with the 1:1 line. In this MLR model, the response (dependent variable) is the temperature-adjusted log diffusion coefficient, i.e., $\log_{10}D$-$(\tau$-3486$)/T$, instead of $\log_{10}D$, in order to fix the temperature coefficients and to avoid multicollinearity problems, as mentioned in Section 2.2.4. The residual plot (Figure 3) shows that the residuals are distributed evenly throughout the dataset, again indicating the good fit of the linear model for the data.

The key predictors other than temperature in the MLR model are the material type and the molecular weight of the diffusing chemical. The regression coefficient when considering log-molecular weight is equal to -2.49, indicating that the diffusion coefficient decreases with increasing molecular weight. This implies that larger molecules diffuse more slowly compared to smaller molecules in solid materials, which is intuitive and consistent with findings from

previous studies [4, 12, 17, 18]. However, although the molecular weight is a highly significant predictor ($p < 0.0001$), it explains less than 10% of the total variance of the diffusion coefficient (SI, Section S4).

The 31 dummy variables for the material types reflect the material dependency and account for most of the total variance of the diffusion coefficient, indicating that the diffusion coefficient in solid materials is strongly dependent on the material type. Since "Polyethylene terephthalate (PET)" was used as the reference material in the regression, the value of its coefficient *b* is zero (Table 2). For each of the other material types, the coefficient *b*, combined with the temperature coefficient τ, i.e. $b+(\tau+2391)/T$, determines the difference in log-diffusion coefficient between that material type and PET, since PET has a temperature coefficient τ of -2391 (K) (Table 2, last column). Chemicals in material types with high values of $b+(\tau+2391)/T$ diffuse quicker than in material types with low values. Therefore, under room temperature (T = 298.15 K), the values of $b+(\tau+2391)/T$ and the corresponding diffusion coefficients tend to be lower in dense, rigid materials such as glass, stainless steel, methyl methacrylate (MMA) polymers, polyethylene naphthalate (PEN), and rigid polymers including polyether ether ketone (PEEK), rigid PVC, polytetrafluoroethylene (PTFE), and polycarbonate (Table 2). In contrast, the values of $b+(\tau+2391)/T$ and the corresponding diffusion coefficients can be up to 13 orders of magnitude higher in flexible or porous materials, such as gypsum, wood, rubber, and polyurethane foam-based materials (Table 2). It should be noted that the composition and properties of a given material type may vary considerably depending on the intended use, as well as over time as material substitutions are made and production procedures differ. Thus, the material type coefficients in Table 2 actually represent an average composition and diffusion behavior for the specific material types.

The significance of the material type coefficient only indicates that the coefficients *b*s of these material types are significantly different from the reference material type, PET, 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 coefficients for material type variables does not indicate that those material types do not have a relevant influence on the diffusion coefficient. As a result, we keep all 31 material type dummy variables in the final regression to retain as much information as possible.

The MLR model given in Eq. 5 contains material-specific variables, so it is only valid for the 32 material types presented in Table 2. For materials that do not belong to those 32 types, we built another generic QPPR to predict the diffusion coefficients, which is presented in SI, Section S4, which should be used with caution because of higher uncertainties.

3.3 Model validation results

3.3.1 Internal validation

For the 20% leave-more-out (LMO) cross validation, the correlation coefficient, Q^2_{LMO} for the 1000 iterations ranges from 0.89 to 0.95, with an average of 0.93, and a root mean square error for cross validation ($RMSE_{cv}$) average of 1.19. 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 1.15, respectively. These results indicate that when fitted to a random 80% of the dataset the model is still able to predict the remaining 20% of the dataset, meaning that the model is internally stable.

For the Y-scrambling, the average R^2_{Yscr} and Q^2_{Yscr} for the 1000 iterations are 0.029 and -0.033, respectively, which are much smaller than the R^2 and Q^2_{LMO} of the original model. The RMSE for Y-scrambling, $RMSE_{Yscr}$, is 4.36 which is much higher than the RMSE and $RMSE_{cv}$ of the original model. These results demonstrate that no correlation exists between the scrambled responses and the predictors. Thus, chance correlation for the original model can be ruled out. Overall, the internal validation demonstrates that the MLR model represented by Eq. 5 is robust and stable, and is not a result of chance correlation.

3.3.2 External validation

As described in Section 2.3.2, the first method of external validation was to split the full dataset (1103 records) into training set and prediction set, and four types of splitting were performed, including splitting by a random 20%, by ordered response, by ordered structure, and by studies. Six criteria for external validation were computed and are presented in Table 3. The R^2_{ext} is the determination coefficient of the prediction set data using the model calculated using the training set data. The other five criteria, Q^2_{F1} [43], Q^2_{F2} [44], Q^2_{F3} [45], r_m^2 [46], and CCC [47], are external validation criteria proposed by different studies, which evaluate various aspects of the model's external prediction ability. These criteria are usually in accordance with each other but can sometimes give contradictory results [47], so they need to be evaluated together. Chirico and Gramatica have proposed threshold values for these different criteria [48], which are presented in Table 3. For the first three types of splitting (by random 20%, by ordered response, and by

ordered structure), the R^2_{ext} are higher than 0.9, and all of the other five criteria pass the threshold values and are also higher than 0.9, indicating good prediction ability of the model calculated using only the training set data. In these three types of splitting, the data were assigned to the training and prediction data sets either randomly or alternately (by ordered response or structure), so it is likely that a portion of the data from each study was assigned to the training set while the remaining portion of the data was assigned to the prediction set. As the result, the prediction set is well within the applicability domain (AD) defined by the training set (SI, Figures S2-S7), so it is expected that the model calculated using the training set can well predict the prediction set.

For the fourth type of splitting, splitting by studies, data from 30 studies were selected as the prediction set, while data from the remaining 48 studies constituted the training set. Thus, all data from one study and for one particular material will be either in the training or in the prediction set, so the validation using this splitting is close to a truly "external" validation. Most of the prediction set is inside the AD defined by the training set except for two data points (SI, Figures S8-S9). As a result, the R^2_{ext} dropped to 0.85, and the values of the other five validation criteria are apparently lower than those for the above three types of splitting, reflecting that variability is higher between than within studies. The five validation criteria nevertheless all pass the threshold values (Table 3), indicating that the model calculated using the training set has good prediction ability.

As a second method of external validation, the 1103 data points from the 68 studies were used as the training set, and additional data from an FDA database and from studies before 1982 were used as two separate prediction sets. As presented in Table 3, when using FDA dataset as the prediction set, the R^2_{ext} is reduced to 0.80 which is lower than the R^2_{ext} for the above four types of splitting. Four of the five validation criteria pass the threshold values, while Q^2_{F3} does not pass the threshold. In contrast, when using data by 1982 as the prediction set, the R^2_{ext} is 0.93, which is very close to the R^2 of the training dataset (Section 3.2). The absolute difference between predicted and measured $\log_{10}D$ averages 2.20 (95^{th} percentile of 5.53) for the FDA dataset, and averages 1.08 (95^{th} percentile of 2.68) for the data by 1982. Figure 3 presents the comparison between model predicted and experimental responses for these two prediction sets. Data from both prediction sets are generally distributed close to the 1:1 line, but the FDA data are more dispersed compared to the training set data while the data by 1982 are almost as

compact as the training set data. The FDA data lack documentation of experimental details, so their quality may not be as good as the data reported in peer-reviewed literature. Also, when the FDA polymer types were linked to our consolidated material types, mismatches may have occurred due to lack of description of the polymers in the FDA dataset, which may lead to inaccuracies in model predictions. Overall, however, our QPPR performs reasonably well on these two fully external datasets, demonstrating its good predictive ability.

3.3.3 Applicability domain (AD)

We performed the analysis of the model's applicability domain (AD) using the three approaches explained in Section 2.3.3. The model being evaluated is the final MLR model presented in Eq. 5, which was calculated using the training set of 1103 data points collected from 68 studies obtained from the peer-reviewed literature. For the analysis of AD, we focus on the two external prediction datasets: the FDA dataset (189 data points) and the data by 1982 (239 data points). Detailed results of the AD analysis are presented in SI, Section S6.1.

Combining the three methods, none of the data points in both prediction sets fell out of the AD. For the FDA dataset, the majority of the data points were inside the AD, while 15 data points were on borderline of AD. Similarly, only 35 data points from the data by 1982 were on borderline of AD. Thus, it is valid to use the present QPPR to make reliable estimates of diffusion coefficients for all data points in the two prediction sets. The physiochemical property space covered by the QPPR is mainly determined by the chemical's molecular weight, which ranges from 30 to 1178 g/mol. The vapor pressure at 25 °C may also be a relevant property, which ranges from $9.8 \cdot 10^{-29}$ to $5.2 \cdot 10^{5}$ Pa. The range of $\log_{10} D$ covered by the QPPR ranges from -22.1 to -5.2 where D is measured in m^2/s.

As mentioned in Section 2.2.2, the model performances of using log-molecular weight and molecular weight as predictors were very close to each other when using the training dataset. However, residual analysis and external validation showed that $\log_{10} MW$ is a more stable predictor than MW when handling high-molecular-weight chemicals, which becomes prominent for the FDA dataset which includes certain chemicals with molecular weight higher than 1500 g/mol. While none of the data points in the FDA dataset fell out of the AD using the $\log_{10} MW$ model, 11 data points would be outside AD using the MW model. Details are presented in SI, Section S6.2. Thus, $\log_{10} MW$ instead of MW was selected as a predictor in the final QPPR (Eq. 5).

Schwope et al. [37] suggested that the linear relationship between $\log_{10}D$ and $\log_{10}MW$ may only be valid for a certain range of molecular weight, and there may be a saturation of diffusion coefficients for small molecular weights, i.e., for a given material and a given temperature, the diffusion coefficient does not continue to increase for chemicals with molecular weight lower than a certain value, which is likely determined by the material type. To further examine the effect of molecular weight on model applicability, we analyzed the model residuals versus the log of molecular weight for the training dataset and the two prediction sets (Figure 4). For the three datasets, the residuals are distributed evenly on both sides of zero in the MW range of the training dataset of 30 and 1178 g/mol ($\log_{10}MW$ of 1.48 to 3.07). For methane (MW=16 g/mol), most of the predictions overestimate diffusivity, suggesting that diffusivity may indeed not further decrease below MW 30 g/mol. Since methane was the only chemical with data available for MW lower than 30 g/mol, data for additional chemicals and materials are therefore needed to further test this hypothesis of saturation at low MW. Similarly, additional data are needed to provide more accurate estimates for chemicals with very high molecular weights.

Overall, the performance of the final model (Eq. 5) in this external validation indicates that it has the ability to provide reliable predictions, as long as the considered chemicals are within the model's applicability domain. With the log-molecular weight as a predictor, our model is able to make reliable extrapolations on chemicals with molecular weights up to about 2500 g/mol, but caution still needs to be taken when applying the model on extremely-high-molecular-weight chemicals. Ideally, the model should be applied to predict diffusion coefficients for chemicals with molecular weights lower than 1178 g/mol which is the maximum within the training dataset. Caution also needs to be taken when applying the model on very-low-molecular-weight chemicals due to the possible saturation effect. Both the FDA dataset and the data by 1982 were used for the external validation but not combined with the original training dataset to calculate a more comprehensive MLR model, because these data are somewhat outdated; the FDA data are not published in literature, so there is a lack of experimental details, making these undocumented data less reliable than the data collected from peer-reviewed literature.

3.4 Limitations and future work

While the extension to 32 different consolidated material types is a major progress, the present model is still not fully comprehensive. First, the model may not be valid for very high or very

low molecular weight (MW) chemicals. It may not be valid for ionizing organic chemicals either, since ionizing chemicals such as acids, alcohols/phenols and amines are not well represented in the training dataset, as they only account for less than 10% of the data points, and the model does not consider chemical ionization or interaction within a material, which may make the chemical's diffusivity lower than that predicted by the model. Second, the present model is not applicable for materials types other than the 32 types in the training set, e.g. for material such as resin and textiles, due to the lack of experimental data. Although a more general MLR model (SI, Section S4) was developed which does not require material type as the predictor, it gives much less accurate predictions of the diffusion coefficient. Third, the present model does not consider any interaction between MW and material type, i.e., it assumes the effect of MW is the same across different materials. Although model validations show that this assumption may be reasonable for the existing data, ideally it needs to be further verified using data spanning the whole MW range (30 to 1178 g/mol) for each material. Therefore, more experimental diffusion coefficient data need to be obtained, or more advanced experimental methods to measure diffusion coefficients need to be developed, for other material types and chemical sizes and classes to make the model more comprehensive.

There are also large variations in the experimental diffusion coefficients between some of different studies for three material types, namely "MMA homopolymer", "Natural rubber" and "Rigid polymers", even after correcting for molecular weight and temperature, as shown in Figure 1 and SI, Figure S1. This means that the regression coefficients b and τ for these material types should be taken with care. The variations could be due to three causes. First, experimental variation; for example, Franz et al. [40] used desorption experiments to measure the diffusion coefficients in MMA homopolymer, while Hennebert et al. [42] used sorption experiments. Second, the swelling of polymers during liquid sorption experiments, which generally occurs for crosslinked polymers in low-molecular weight solvents [49], may not always be accounted for, and can lower the diffusion coefficients by orders of magnitude [10]. Third, the properties of the same material can vary between studies depending on how it was made and which additives were used. This may also be the case for some other materials such as vinyl flooring, carpet, synthetic rubber, etc., for which the material type coefficients in Eq. 5 can only represent some sort of average composition and diffusion behavior for the specific materials. Ideally, quantitative, continuous properties of the solid materials, such as density, porosity and crystalline state of the

material as well as other descriptors of the material's composition and molecular structure, instead of qualitative material types could be measured and entered into the model as predictors, so that the model can be more accurate and can be extrapolated to various material types outside the training dataset.

4. Conclusions

A multiple linear regression model has been developed to predict the internal diffusion coefficients of organic compounds in various solid materials (excel model provided in SI). Experimental diffusion coefficient data collected from 68 studies of the peer-reviewed literature were used as the training set for the regression. The model uses two continuous variables, molecular weight and inversed absolute temperature, and one categorical variable, material type, as predictors. The model has been internally validated to be robust, stable and not a result of chance correlation. External validation using two prediction sets demonstrates that the model predictions are most reliable within the model's applicability domain, namely molecular weight between 30 and 1178 g/mol temperature between 4 and 180 °C, and material type belonging to the 32 consolidated types.

The main advantage of the present model is that it is applicable for chemicals with a wide range of molecular weights (but only up to about 16 to 2500 g/mol, with special treatment for molecular weight lower than 30 g/mol) in various materials. This is advantageous compared to the correlation methods developed in previous studies often specific for certain chemical classes or materials. The present model is able to provide reliable estimates of diffusion coefficients for a large number of chemical-material combinations, making it suitable for high-throughput assessments of the releases and human exposures to chemicals encapsulated in solid materials, particularly building materials and food contact materials. To make the model comprehensive, more experimental diffusion coefficient data need to be obtained for other material types, or quantitative and continuous parametrization of various solid materials needs to be further developed.

Acknowledgements

The authors thank Prof. Ester Papa, Dr Allessandro Sangion, and Prof. Paola Gramatica from the University of Insubria, Italy for advice on MLR modeling and validation, as well as support for

the QSARINS software. Funding for this research was provided by US EPA contract EP-16-C-000070 and by the Long Range Research Initiative of the American Chemistry Council. P. Fantke was supported by the Marie Curie project Quan-Tox (GA No. 631910) funded by the European Commission under the Seventh Framework Programme.

Tables and Figures

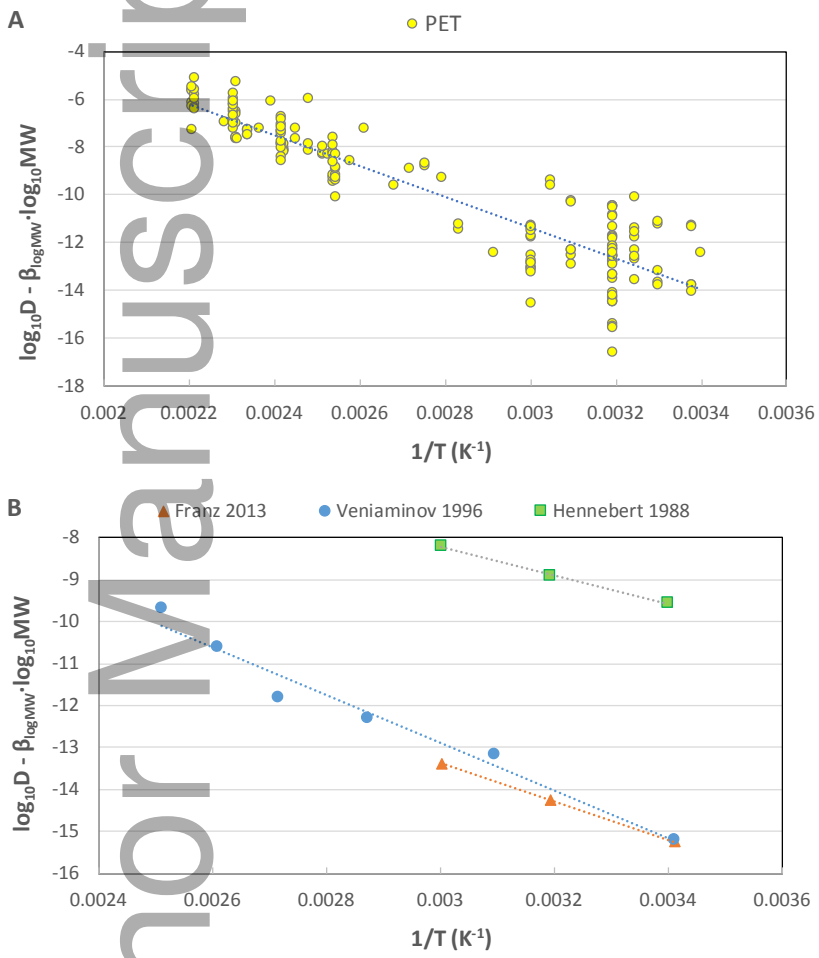

Figure 1. Relationship between the diffusion coefficient *D* (corrected for $\log_{10}MW$) and the inverse of temperature for (A) PET, and (B) methyl methacrylate (MMA) homopolymer. The units of *D* and *MW* are m^2/s and g/mol, respectively.

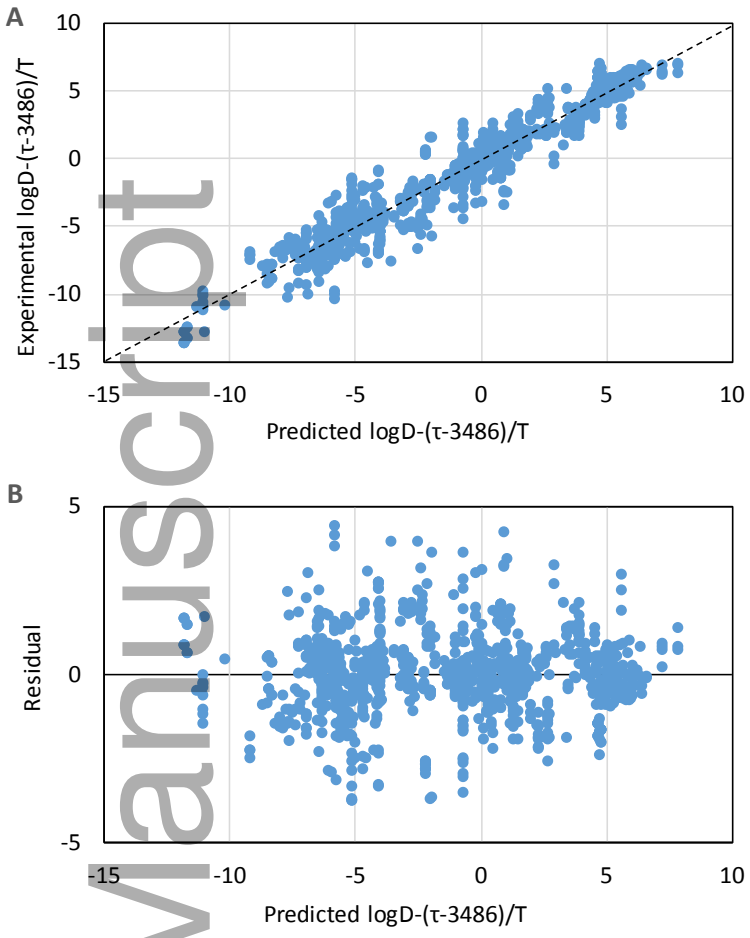

Figure 2. Values of $\log_{10}D\text{-}(\tau\text{-}3486)/T$ predicted by the final QPPR (Eq. 5) vs. (A) experimental values, and (B) residuals. The dotted line in (A) indicates the 1:1 line. The units of D and T are m^2/s and K, respectively.

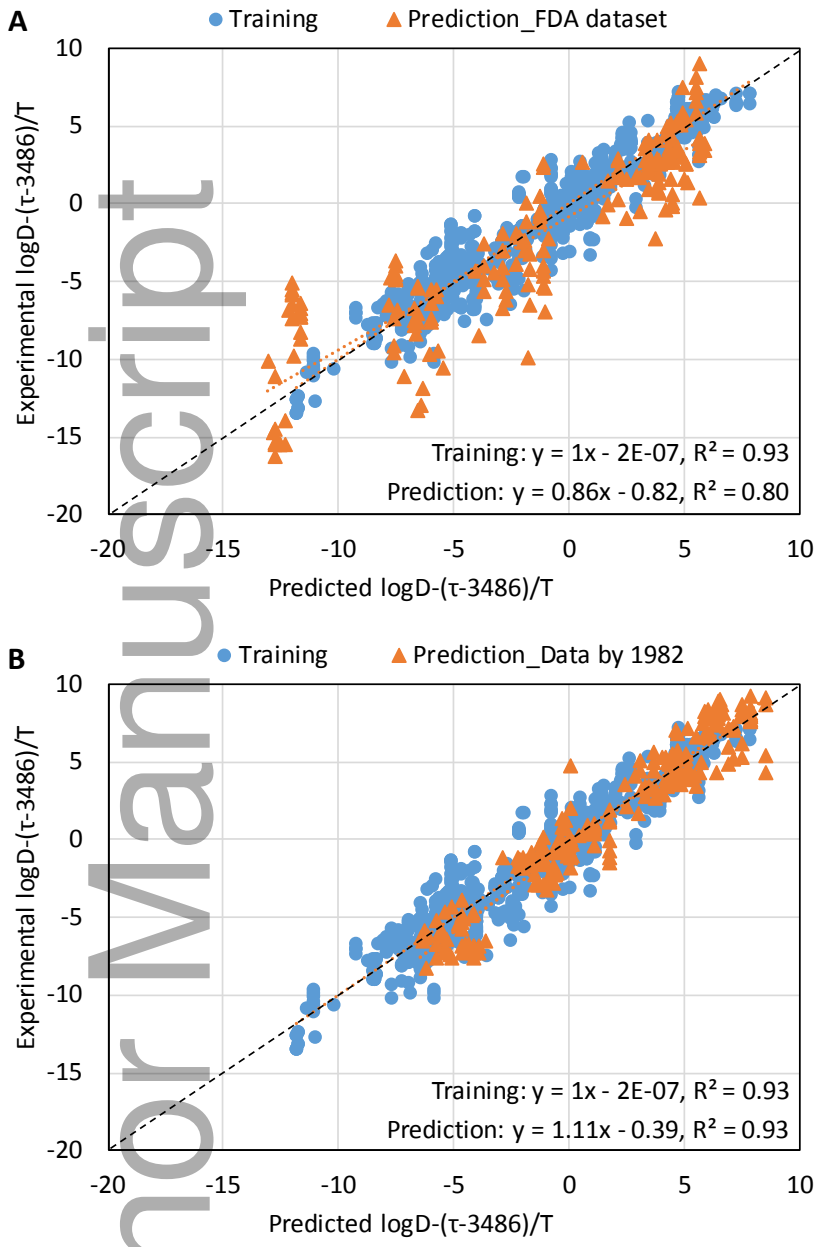

Figure 3. Values of $\log_{10}D\text{-}(\tau\text{-}3486)/T$ predicted by the final QPPR (Eq. 5) vs. experimental values when using (A) FDA dataset and (B) Data by 1982 as the prediction sets. The black dotted line indicates the 1:1 line. The units of D and T are m^2/s and K, respectively.

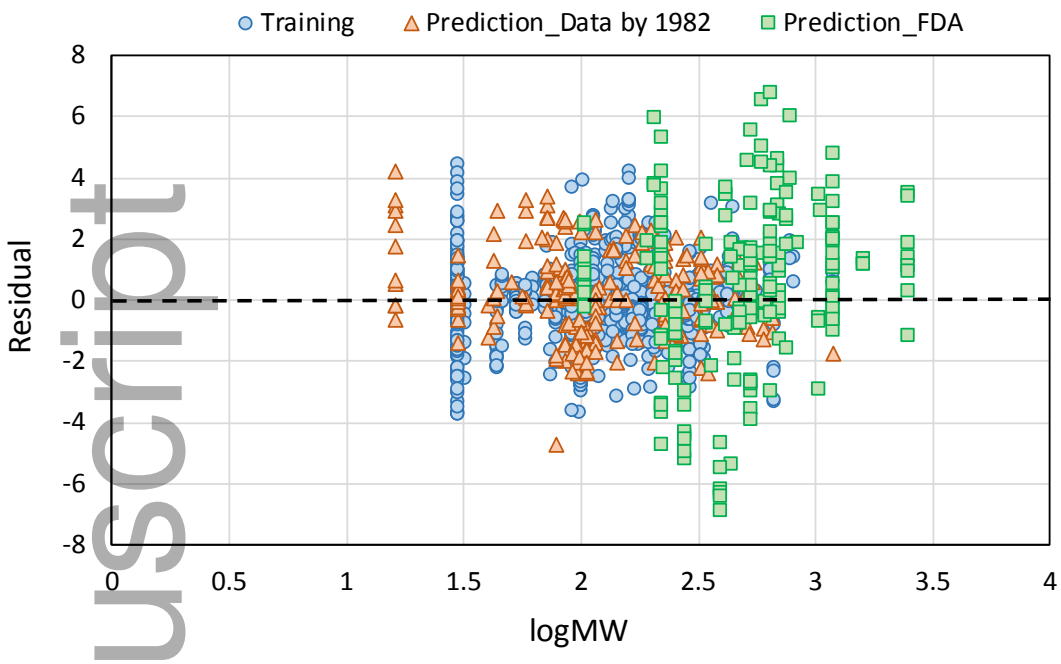

Figure 4. Residual between the present QPPR and observed data as a function of $\log_{10}$MW for the training dataset, the FDA dataset, and the data by 1982 set. The unit of MW is g/mol.

Table 1. Temperature dependence of diffusion coefficient in the 32 consolidated material types (all numbers are in the unit of K)

Category	Material	Mean coefficient of 1/T	SD between studies	Coefficient value for Eq. 5		
				$\beta_{1/T}$	τ	$\beta_{1/T}+\tau$
High-coefficient category	PP homopolymer	-6665	2354	-3486	-2391	-5877
	Polyethylene terephthalate (PET)	-6567	2399			
	General polystyrene (PS)	-5713	3560			
	Polyethylene naphthalate (PEN)	-5449	1940			
	PP copolymer	-5384	1194			
	High-density polyethylene (HDPE)	-5294	1124			
Medium-coefficient category	MMA homopolymer	-4549	1145	-3486	0	-3486
	ABS, EVOH	-4222	n/a			
	High-impact polystyrene (HIPS)	-4215	n/a			
	Polyamide (PA)	-4179	1854			
	MMA copolymer-medium or low density	-4056	1272			
	Polyethylene (PE, LDPE, LLDPE)	-3713	536			
	Limited-data material group	n/a	n/a			
	Calcium silicate	n/a	n/a			
	Carpet	n/a	n/a			
	Glass, Stainless steel	n/a	n/a			
	Vinyl acetate-based polymers	n/a	n/a			
	Cement	n/a	n/a			
Low-coefficient category	Gypsum board	n/a	n/a	-3486	1676	-1810
	Plywood	n/a	n/a			
	Flexible PVC	-2917	2618			
	Other wooden boards	-2411	888			
	Polychloroprene (CR)	-2127	286			
	Vinyl flooring	-1951	n/a			
	Polystyrene foam (XPS, EPS)	-1806	n/a			
	Polyurethane foam-based materials*	-1705	699			
	Synthetic rubber	-1326	205			
	Ethylene-propylene rubbers	-1145	300			
	Natural rubber (NR)	-939	337			
	Rigid polymers	-510	1552			
	Paper	-312	n/a			
	Gypsum and cellulose ceiling tile	331	294			

*This material type refers to low-density polyurethane foams with a density of 0.005 to 0.03 g/cm^3.

Table 2. Material-specific coefficients for Eq. 5

Material	Coefficient *b*				*b+(τ+2391.15)/T*
	Coefficient	SE[f]	p-value	τ (K)	at 25 °C
Calcium silicate	1.17	0.29	< 0.0001	0	9.19
Carpet	-1.23	0.28	< 0.0001	0	6.79
Cement	0.330	0.226	0.15	0	8.35
Ethylene-propylene rubbers	-6.32	0.29	< 0.0001	1676	7.32
Flexible PVC	-8.51	0.31	< 0.0001	1676	5.13
General polystyrene (PS)	2.04	0.30	< 0.0001	-2391	2.04
Glass, Stainless steel	-8.57	0.38	< 0.0001	0	-0.550
Gypsum and cellulose ceiling tile	-1.24	0.31	< 0.0001	1676	12.4
Gypsum board	-5.77	0.30	< 0.0001	1676	7.87
High density polyethylene (HDPE)	5.11	0.20	< 0.0001	-2391	5.11
High-impact polystyrene (HIPS)	-7.11	0.27	< 0.0001	0	0.907
Methyl methacrylate (MMA) copolymer-medium or low density	-7.73	0.21	< 0.0001	0	0.294
Methyl methacrylate (MMA) homopolymer[h]	-7.84	0.31	< 0.0001	0	0.175
Natural rubber (NR)[h]	-3.60	0.27	< 0.0001	1676	10.0
Other wooden boards[a]	-6.72	0.21	< 0.0001	1676	6.92
Paper	-8.53	0.34	< 0.0001	1676	5.11
Plywood	-5.61	0.34	< 0.0001	1676	8.03
Polyamide (PA)	-5.40	0.16	< 0.0001	0	2.62
Poly acrylnitrile butadiene styrene (ABS), Ethylene vinyl alcohol (EVOH)	-4.97	0.23	< 0.0001	0	3.05
Polychloroprene (CR)	-6.31	0.35	< 0.0001	1676	7.33
Polyethylene (PE, LDPE, LLDPE)	-1.65	0.16	< 0.0001	0	6.37
Polyethylene naphthalate (PEN)	-1.16	0.28	< 0.0001	-2391	-1.16
Polyethylene terephthalate (PET) [g]	**0.00**	**0.15**	**n/a**	-2391	0.00
Polystyrene foam (XPS, EPS)	-8.32	0.29	< 0.0001	1676	5.32
Polyurethane foam-based materials[b]	-7.35	0.25	< 0.0001	1676	6.30
PP copolymer	4.79	0.28	< 0.0001	-2391	4.79
PP homopolymer	4.53	0.15	< 0.0001	-2391	4.53
Rigid polymers[c, h]	-11.9	0.25	< 0.0001	1676	1.70
Synthetic rubber	-5.93	0.32	< 0.0001	1676	7.71
Vinyl acetate-based polymers[d]	-0.459	0.326	0.16	0	7.56
Vinyl flooring	-6.77	0.21	< 0.0001	1676	6.87
Limited-data material group[e]	see footnotes				

[a] Includes Particleboard, Oriented strand board (OSB), Medium-density fiberboard (MDF), High-density board, and Wood chamber wall.

[b] This material type refers to low-density polyurethane foams with a density of 0.005 to 0.03 g/cm^3.

[c] Includes Polyether ether ketone (PEEK), Rigid PVC, Polytetrafluoroethylene (PTFE), and Polycarbonate.

[d] Includes Ethyl vinyl acetate (EVA), Polyvinyl acetate (PVA), and Polyvinyl acetate polyacrylic acid copolymer.

[e] The coefficient *b* for this group is -2.26 with an SE of 0.18, and the coefficient τ is 0. "Limited-data material group" includes data from 20 different materials, so the accuracy of the coefficients is low and they are not recommended for use in predicting diffusion coefficients. This group includes Alginate film, Balance, Decorative and Overlay layers of wooden flooring, Cellulose, Epichlorhydrin-dimethylamine polymer (EDP), Epoxy/acrylic copolymer, latex, MMA/Butyl methacrylic (BMA) copolymer -very low density, Nanocomposite polyamide, Paint, Pectin film, Pectin/Alginate composite film, Polydimethylsiloxane (PDMS) membrane, Polyisoprene (PI) membrane, Polyoctenamer (PO) membrane, Polyoxymethylene, Polytrimethylene terephthalate (PTT), Polyvinylidene chloride (PVDC), and Silicone.

[f] Standard error.

[g] Reference material.

[h] Coefficients should be taken with care due to large variations between studies.

Table 3. External validation results

External validation criteria	R^2_{ext}	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	$\overline{r_m^2}$	CCC
Threshold		> 0.70	> 0.70	> 0.70	> 0.65	> 0.85
Splitting by random percentage	0.92	0.92	0.92	0.92	0.90	0.96
Splitting by ordered response	0.94	0.94	0.94	0.95	0.93	0.97
Splitting by ordered structure	0.94	0.94	0.94	0.94	0.91	0.97
Splitting by studies	0.85	0.85	0.84	0.85	0.78	0.92
FDA dataset as prediction set	0.80	0.77	0.77	0.60	0.71	0.89
Data by 1982 as prediction set	0.93	0.93	0.92	0.90	0.85	0.95

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.

$\overline{r_m^2}$: determination coefficient proposed by Ojha et al.

CCC: concordance correlation coefficient proposed by Chirico and Gramatica.

References

1. Little JC, Weschler CJ, Nazaroff WW, et al. Rapid methods to estimate potential exposure to semivolatile organic compounds in the indoor environment. *Environ Sci Technol.* 2012; 46(20): p. 11171-11178.
2. Xu Y, Cohen Hubal EA, Clausen PA, et al. Predicting residential exposure to phthalate plasticizer emitted from vinyl flooring: a mechanistic analysis. *Environ Sci Technol.* 2009; 43(7): p. 2374-2380.
3. Guo Z. Review of indoor emission source models. Part 1. Overview. *Environ Pollut.* 2002; 120(3): p. 533-549.
4. Begley T, Castle L, Feigenbaum A, et al. Evaluation of migration models that might be used in support of regulations for food-contact plastics. *Food Addit Contam.* 2005; 22(1): p. 73-90.
5. Xie M, Wu Y, Little JC, et al. Phthalates and alternative plasticizers and potential for contact exposure from children's backpacks and toys. *J Expo Sci Env Epid.* 2016; (26): p. 119-124.
6. Liu Z, Ye W, Little JC. Predicting emissions of volatile and semivolatile organic compounds from building materials: a review. *Build Environ.* 2013; 64: p. 7-25.
7. Haghighat F, Huang H, Lee C-S. Modeling approaches for indoor air VOC emissions from dry building materials–a review. *ASHRAE Trans.* 2005; 111(1): p. 635-645.

8. Berens A and Hopfenberg H. Diffusion of organic vapors at low concentrations in glassy PVC, polystyrene, and PMMA. *J Membrane Sci.* 1982; 10(2-3): p. 283-303.
9. Hickey AS and Peppas NA. Solute diffusion in poly (vinyl alcohol)/poly (acrylic acid) composite membranes prepared by freezing/thawing techniques. *Polymer.* 1997; 38(24): p. 5931-5936.
10. John J, Kunchandy S, Kumar A, et al. Transport of methyl methacrylate monomer through natural rubber. *J Mater Sci.* 2010; 45(2): p. 409-417.
11. Luo R and Niu J. Determining diffusion and partition coefficients of VOCs in cement using one FLEC. *Build Environ.* 2006; 41(9): p. 1148-1160.
12. 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: p. 789.
13. Bodalal A, Zhang J, Plett E. A method for measuring internal diffusion and equilibrium partition coefficients of volatile organic compounds for building materials. *Build Environ.* 2000; 35(2): p. 101-110.
14. Little JC, Hodgson AT, Gadgil AJ. Modeling emissions of volatile organic compounds from new carpets. *Atmos Environ.* 1994; 28(2): p. 227-234.
15. Dole P, Feigenbaum AE, Cruz CDL, et al. Typical diffusion behaviour in packaging polymers–application to functional barriers. *Food Addit Contam.* 2006; 23(2): p. 202-211.
16. Reynier A, Dole P, Humbel S, et al. Diffusion coefficients of additives in polymers. I. Correlation with geometric parameters. *J Appl Polym Sci.* 2001; 82(10): p. 2422-2433.
17. Guo Z. Review of indoor emission source models. Part 2. Parameter estimation. *Environ Pollut.* 2002; 120(3): p. 551-564.
18. Cox SS, Zhao D, Little JC. Measuring partition and diffusion coefficients for volatile organic compounds in vinyl flooring. *Atmos Environ.* 2001; 35(22): p. 3823-3830.
19. Zhao D, Cox S, Little J. *Source/sink characterization of diffusion controlled building materials*. in *Proceedings of the 8th International Conference on Indoor Air Quality and Climate-Indoor Air*. 1999.
20. Jolliet O, Ernstoff AS, Csiszar SA, et al. Defining Product Intake Fraction to Quantify and Compare Exposure to Consumer Products. *Environ Sci Technol.* 2015; 49: p. 8924-8931.
21. Shin H-M, Ernstoff A, Arnot JA, et al. Risk-based high-throughput chemical screening and prioritization using exposure models and in vitro bioactivity assays. *Environ Sci Technol.* 2015; 49(11): p. 6760-6771.

22. Shin H-M, McKone TE, Bennett DH. Intake fraction for the indoor environment: a tool for prioritizing indoor chemical sources. *Environ Sci Technol.* 2012; 46(18): p. 10063-10072.
23. Ernstoff AS, Fantke P, Csiszar SA, et al. Multi-pathway exposure modelling of chemicals in cosmetics with application to shampoo. *Environ Int.* 2016; 92-93: p. 87-96.
24. Csiszar SA, Ernstoff AS, Fantke P, et al. Stochastic modeling of near-field exposure to parabens in personal care products. *J Expo Sci Env Epid.* 2017; (27): p. 152-159.
25. Egeghy PP, Sheldon LS, Isaacs KK, et al. Computational exposure science: An emerging discipline to support 21st-century risk assessment. *Environ Health Persp.* 2016; 124(6): p. 697.
26. Deng Q, Yang X, Zhang J. Study on a new correlation between diffusion coefficient and temperature in porous building materials. *Atmos Environ.* 2009; 43(12): p. 2080-2083.
27. Xu J and Zhang JS. An experimental study of relative humidity effect on VOCs' effective diffusion coefficient and partition coefficient in a porous medium. *Build Environ.* 2011; 46(9): p. 1785-1796.
28. Xu J, Zhang JS, Liu X, et al. Determination of partition and diffusion coefficients of formaldehyde in selected building materials and impact of relative humidity. *J Air Waste Ma.* 2012; 62(6): p. 671-679.
29. Park J-S, Little JC, Kim S-D, et al. The Determination of Diffusion and Partition Coefficients of PUF. *J Korean Soc Atmos Environ.* 2010; 26(1): p. 77-84.
30. Welle F and Franz R. Diffusion coefficients and activation energies of diffusion of low molecular weight migrants in Poly(ethylene terephthalate) bottles. *Polym Test.* 2012; 31(1): p. 93-101.
31. Ewender J and Welle F. Determination of the activation energies of diffusion of organic molecules in poly (ethylene terephthalate). *J Appl Polym Sci.* 2013; 128(6): p. 3885-3892.
32. 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): p. 1036-1044.
33. 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): p. 2121-2132.
34. Flynn JH. A collection of kinetic data for the diffusion of organic compounds in polyolefins. *Polymer.* 1982; 23(9): p. 1325-1344.
35. Park G. The diffusion of some organic substances in polystyrene. *Transactions of the Faraday Society.* 1951; 47: p. 1007-1013.
36. Park G. The diffusion of some halo-methanes in polystyrene. *Transactions of the Faraday Society.* 1950; 46: p. 684-697.

37. Schwope A, Goydan R, Reid R, *Methods for assessing exposure to chemical substances Volume 11: Methodology for Estimating the Migration of Additives and Impurities from Polymeric Materials* 1990, U.S.EPA: Washington, D.C.
38. Gramatica P. Principles of QSAR models validation: internal and external. *QSAR & combinatorial science.* 2007; 26(5): p. 694-701.
39. Cassani S and Gramatica P. Identification of potential PBT behavior of personal care products by structural approaches. *Sustain Chem Pharm.* 2015; 1: p. 19-27.
40. Franz R and Brandsch R. Migration of acrylic monomers from methacrylate polymers–establishing parameters for migration modelling. *Packag Technol Sci.* 2013; 26(8): p. 435-451.
41. Veniaminov A and Sedunov YN. Diffusion of phenanthrenequinone in poly (methyl methacrylate): holographic measurements. *Polym Sci Ser A.* 1996; 38: p. 59-63.
42. Hennebert P. Solubility and diffusion coefficients of gaseous formaldehyde in polymers. *Biomaterials.* 1988; 9(2): p. 162-167.
43. Shi LM, Fang H, Tong W, et al. QSAR models using a large diverse set of estrogens. *J Chem Inf Comp Sci.* 2001; 41(1): p. 186-195.
44. Schüürmann G, Ebert R-U, Chen J, et al. External validation and prediction employing the predictive squared correlation coefficient - Test set activity mean vs training set activity mean. *J Chem Inf Model.* 2008; 48(11): p. 2140-2145.
45. Consonni V, Ballabio D, Todeschini R. Comments on the definition of the Q^2 parameter for QSAR validation. *J Chem Inf Model.* 2009; 49(7): p. 1669-1678.
46. Ojha PK, Mitra I, Das RN, et al. Further exploring r_m^2 metrics for validation of QSPR models. *Chemometr Intell Lab.* 2011; 107(1): p. 194-205.
47. Chirico N and 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): p. 2320-2335.
48. Chirico N and 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): p. 2044-2058.
49. Nandi S and Winter HH. Swelling behavior of partially cross-linked polymers: a ternary system. *Macromolecules.* 2005; 38(10): p. 4447-4455.