DR. LEI HUANG (Orcid ID : 0000-0002-7846-9760)

Article type   : Original Article

A Quantitative Structure-Property Relationship (QSPR) for Estimating Solid Material-Air Partition Coefficients of Organic Compounds

Lei Huang[1]* and Olivier Jolliet[1]

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

*Corresponding author, huanglei@umich.edu

29 Abstract

30 The material-air partition coefficient (K_{ma}) is a key parameter to estimate the release of

31 chemicals incorporated in solid materials and resulting human exposures. Existing correlations to

32 estimate K_{ma} are applicable for a limited number of chemical-material combinations without

33 considering the effect of temperature. The present study develops a quantitative property-

34 property relationship (QSPR) to predict K_{ma} for a large number of chemical-material

35 combinations. We compiled a dataset of 991 measured K_{ma} for 179 chemicals in 22 consolidated

36 material types. A multiple linear regression model predicts K_{ma} as a function of chemical's K_{oa},

37 enthalpy of vaporization (ΔH_v), temperature and material type. The model shows good fitting of

38 the experimental dataset with adjusted R^2 of 0.93 and has been verified by internal and external

39 validations to be robust, stable and has good predicting ability ($R^2_{ext} > 0.78$). A generic QSPR is

40 also developed to predict K_{ma} from chemical properties and temperature only (adjusted $R^2 =$

41 0.84), without the need to assign a specific material type. These QSPRs provide correlation

42 methods to estimate K_{ma} for a wide range of organic chemicals and materials, which will

43 facilitate high-throughput estimates of human exposures for chemicals in solid materials,

44 particularly building materials and furniture.

45 Keywords

46 Partitioning, Indoor release, Solid materials, Organic chemicals, Consumer exposure, Correlation

47 Practical implications

48 The developed QSPRs provide a comprehensive correlation method to estimate K_{ma}, covering a

49 much wider range of organic chemicals and solid materials compared to previous studies, and

50 include a still accurate generic correlation without the need to assign a material type. Combined

51 with the QSPR estimating the internal diffusion coefficient [1], these QSPRs facilitate high-

52 throughput estimates of indoor human exposures to chemicals incorporated in solid materials.

53 This is highly relevant for multiple science-policy fields, including chemical alternatives

54 assessment (CAA), risk assessment (RA) and life cycle assessment (LCA).

55

56 1. Introduction

57 Chemicals incorporated in solid materials have been identified as a major source of passive

58 emissions to indoor air and of transfers into house dust and skin. Typical examples include

59   chemicals used as plasticizers in building materials and flame retardants in furniture. To estimate

60   the release of these chemicals from solid materials and subsequent consumer exposures, the

61   dimensionless solid material-air partition coefficient (K_{ma}), defined as the ratio of the

62   concentration in the material to the concentration in the air at equilibrium, is one of the key

63   parameters [2]. The K_{ma} is essential in determining the chemical transfer from solid material to air

64   and to house dust, as well as the chemical concentration at the material surface, which further

65   determines the inhalation, dermal and dust ingestion exposures. K_{ma} is specific to a chemical-

66   material combination and is also influenced by ambient temperature. Experimental techniques

67   such as chamber tests for building materials [3], and sorption experiments for polymer materials [4-6]

68   have enabled measurement of a limited number of K_{ma} values for building materials such as

69   vinyl flooring, gypsum board, plywood and cement, as well as polymer materials used for

70   passive samplers including polyurethane foams (PUF), polyethylene (PE), and polypropylene

71   (PP). Recently, studies have also been conducted to measure the K_{ma} for clothing and fabrics [7, 8].

72   However, since experiments are costly and time-consuming, measured K_{ma} values are only

73   available for a limited number of chemical-material combinations. Thus, quantitative

74   relationships are needed to predict this partition coefficient from known physiochemical

75   properties for chemicals without experimental data, which is especially important for high-

76   throughput approaches, for which a large number of chemical-material combinations need to be

77   evaluated.

78   Several correlation methods have been developed to estimate K_{ma} from physiochemical

79   properties of chemicals. For example, several studies have correlated K_{ma} to the chemical's

80   vapor pressure using data on volatile organic compounds (VOCs) in building materials [4, 9-11].

81   Other studies which focused on semi-volatile organic compounds (SVOCs) in passive sampling

82   devices have found correlation between K_{ma} and the octanol-air partition coefficient (K_{oa}) [5, 6, 12,

83   13]. Furthermore, Holmgren et al. estimated K_{ma} as a function of five Abraham solvation

84   parameters for six groups of materials [14], but these parameters are not readily available. For the

85   aforementioned approaches, the main limitation is that the correlations are specific to certain

86   chemical classes and materials; for example polycyclic aromatic hydrocarbons (PAHs) in low-

87   density polyethylene (LDPE), which limits their application for other chemical-material

88   combinations. Addressing this research gap to facilitate wider applicability, Guo developed a

89   method which estimates the K_{ma} as a function of the chemical's vapor pressure for all materials

90   and chemical classes [11]. However, this approach is developed based on a small dataset which

91   mainly includes VOCs in building materials which limits its applicability to also address SVOCs.

92   Another limitation of the previous studies is that the effect of temperature was not well

93   considered in the correlation. Some studies provided different correlation coefficients for certain

94   discrete temperatures [15], while others corrected the predictors for temperature [16]. However, since

95   the known physiochemical properties such as vapor pressure and K_{oa} are often only given as

96   values at 25 °C, correcting them for temperature may not always be practical as the

97   corresponding enthalpies of phase change are not available for all chemicals. Several studies did

98   establish correlations between K_{ma} and temperature, but the correlations were only verified using

99   experimental data on limited chemicals such as formaldehyde and other aldehydes [17, 18].

100   In all, the currently available correlation methods to estimate K_{ma} do not provide sufficient

101   coverage of chemicals incorporated in solid materials at different ambient temperatures. A recent

102   research hotspot in exposure sciences is to develop low tier, high-throughput methods to estimate

103   exposure to chemical in consumer products across a variety of chemical-material combinations,

104   which requires high-throughput estimates of K_{ma} for a wide range of material-chemicals

105   combinations. Thus, the present study aims to develop a more comprehensive correlation method

106   to estimate K_{ma} for a wide range of organic compounds in multiple solid materials, addressing

107   the need for high-throughput exposure assessments. More specifically, we aim to:

108   1) Carry out a comprehensive literature review to collect experimental K_{ma} data on a wide range

109   of materials and chemicals.

110   2) Use multiple linear regression techniques to establish the relationship between K_{ma} and

111   various predictor variables including physiochemical properties, material type and temperature.

112   3) Perform internal and external validations to characterize the validity and predictive power of

113   the developed correlation.

114   This QSPR provides a more advanced correlation method to estimate the K_{ma} of organic

115   compounds compared to previous studies, as it covers a wide range of solid materials and

116   chemicals, and consistently incorporates the effect of temperature. A similar QSPR has been

117   developed by our group for the internal diffusion coefficient in solid materials [1]. By providing

118   reliable estimates of the key partition and diffusion parameters for a large number of material-

119   chemical combinations, these QSPRs will facilitate high-throughput assessments of chemical

120   emissions and human exposures for chemicals incorporated in solid materials relevant for

121 various science-policy fields such as chemical alternatives assessment (CAA), risk assessment

122 and life-cycle assessment (LCA).

123

124 ## 2. Materials and Methods

125 ### 2.1 Dataset

126 #### 2.1.1 Data collection

127 Experimental material-air partition coefficient data were compiled from 43 references from the

128 peer-reviewed scientific literature (provided in Supporting Information (SI), Section S1).

129 Dimensionless partition coefficients were collected. If the partition coefficients were expressed

130 in mL/g or m^3/g, they were converted to dimensionless values by multiplying these by the

131 density of the solid material. If the partition coefficients were expressed in the unit of m, they

132 were converted to dimensionless values by dividing these by the thickness of the material. The

133 initial dataset of K_{ma} contained a total of 1008 records covering 179 unique chemicals and 75

134 distinct solid materials.

135 #### 2.1.2 Data curation

136 For the 179 unique chemicals of the initial K_{ma} dataset, molecular weight, vapor pressure, water

137 solubility and $logK_{ow}$ at 25 °C were obtained from EPISuite [19]. For these physiochemical

138 properties, experimental values were used when available, otherwise the software-estimated

139 values were used. The enthalpy of vaporization (ΔH_v, J/mol) of each chemical was obtained

140 from ChemSpider estimated values (www.chemspider.com).

141 For the octanol-air partition coefficient ($logK_{oa}$) at 25 °C, experimental values are only available

142 for part of the 179 chemicals in the dataset. To avoid inconsistency, we used the $logK_{oa}$ values

143 estimated by EPISuite [19] for all of the 179 chemicals. In EPISuite, $logK_{oa}$ is estimated by

144 subtracting $logK_{aw}$ (dimensionless log air-water partition coefficient) from $logK_{ow}$, $logK_{aw}$ and

145 $logK_{ow}$ being estimated by the HenryWin and KowWin functions, respectively [19]. Experimental

146 $logK_{oa}$ values were also collected and their impacts on the QSPR were assessed, as presented in

147 SI Section S6.

148 To avoid over-fitting of the QSPR model, the 75 original materials for K_{ma} were grouped into 22

149 consolidated material types, based on the name of the materials and the similarity of the

150 regression coefficients (see SI Section S1), ensuring a minimum of 5 data points and 3 different

151 chemicals per consolidated material type. The data points with materials that cannot be grouped

152 according to the above criteria were excluded from further analyses.

153 The final K_{ma} dataset contains 991 data points with 179 unique chemicals in 22 consolidated

154 material types. The temperature at which the K_{ma} was measured ranges from 15 to 100 °C. The

155 final dataset is provided in SI.

156

157 2.2 Modeling methods

158 2.2.1 Multiple linear regression model

159 A multiple linear regression (MLR) analysis was performed to identify and quantify the effect of

160 different parameters on the partition coefficient, with details described in our previous paper on

161 the QSPR for diffusion coefficient [1]. Briefly, the MLR model takes the following general form:

162 $$\log_{10} K_{ma} = \alpha + \beta_1 \cdot X_1 + \cdots + \beta_n \cdot X_n + b_1 \cdot M_1 + \cdots + b_m \cdot M_m \qquad (1)$$

163 where $\log_{10} K_{ma}$ is the logarithm of the dimensionless K_{ma}, α is the intercept; X_1 to X_n are

164 independent variables related to the properties of the chemical or the environment; β_1 to β_n are

165 regression coefficients for the respective independent variables X_1 to X_n. M_1 to M_m are dummy

166 variables for the packaging materials, with one dummy variable per type of material. A dummy

167 variable equals 1 for the material type it represents, and equals 0 for all other materials; for

168 example, $M_1 = 1$ for material type 1, $M_1 = 0$ for material types 2 to m. b_1 to b_m are regression

169 coefficients for the respective dummy variables M_1 to M_m. The number of m is equal to the

170 number of material types considered minus 1, since PU-ether - the material type with the highest

171 number of measured K_{ma} data - is used as the reference material type and does not require a

172 dummy available in the MLR. Regression coefficients were estimated by the least squares (LS)

173 method. All regression analyses were performed using IBM SPSS Statistics version 23 (IBM

174 corporation, Armonk, New York).

175 In previous studies, either the chemical's vapor pressure [4, 9-11] or $\log K_{oa}$ [5, 6, 12, 13] has been used as

176 predictor of the K_{ma} in a given material. Abraham solvation parameters were also used as

177 predictors by Holmgren et al. [14], but these parameters are not considered here since they are not

178 readily available. Initial regressions (SI Section S2) suggest that $\log K_{oa}$ is a better predictor of

179 K_{ma} compared to vapor pressure. Thus, the chemical's $\log K_{oa}$ at 25 °C was used as the

180 independent variable for chemical properties in Eq. 1.

181 Thus, the MLR model takes the following form:

182 $$\log_{10} K_{ma} = \alpha + \beta_{log K_{oa}} \cdot \log_{10} K_{oa} + \beta_T \cdot T\_term + b_1 \cdot M_1 + \cdots + b_{21} \cdot M_{21} \qquad (2)$$

183 where *T_term* is a term representing the effect of temperature and will be described in the next

184 section (Section 2.2.2).

185 2.2.2 Temperature dependence

186 In thermodynamics, the temperature dependence of equilibrium constant, K_{eq}, can be described

187 by the van't Hoff equation:

188 $$ln \frac{K_2}{K_1} = \frac{\Delta H_{phase\ change}}{R} \left(\frac{1}{T_2} - \frac{1}{T_1}\right) \qquad (3a)$$

189 where K_1 and K_2 are the equilibrium constants at temperature T_1 and T_2, respectively, T_1 and T_2

190 are absolute temperatures (K), R is ideal gas constant (8.314 J/(K·mol)), and $\Delta H_{phase\ change}$ is the

191 enthalpy of phase change (J/mol).

192 Since K_{ma} is an equilibrium constant by definition and the chemical's $\log_{10} K_{oa}$ at 25 °C or

193 298.15 K is used as an independent variable in the MLR model (Eq. 2), we assume that the

194 temperature dependence of K_{ma} also follows the van't Hoff equation:

195 $$T_{term} = log_{10} \frac{K_{ma\_2}}{K_{ma\_1}} = \frac{\Delta H_{ma}}{2.303 \cdot R} \left(\frac{1}{T_2} - \frac{1}{298.15}\right) \qquad (3b)$$

196 where ΔH_{ma} is the enthalpy of the partitioning between material and air (J/mol), and 2.303 is a

197 conversion factor between $\log_{10} K$ and lnK.

198 Ideally, the enthalpy ΔH_{ma} should be different for different chemical-material combinations.

199 Kamprad and Goss have determined the ΔH_{ma} values for 54 unique chemicals in PU-ether using

200 measured K_{ma} data from 15 °C to 95°C [4], so we were able to develop a linear correlation to

201 estimate ΔH_{ma} from chemical properties (results shown in Section 3.1). Since no experimental

202 ΔH_{ma} values are available for materials other than PU-ether, we use the ΔH_{ma} correlation

203 developed above across all materials. Therefore, in our regression model of K_{ma}, the ΔH_{ma} is

204 chemical-specific, but not material-specific. The final MLR model thus takes the following form:

205 $$\log_{10} K_{ma} = \alpha + \beta_{log K_{oa}} \cdot \log_{10} K_{oa} + \beta_T \cdot \frac{\Delta H_{ma}}{2.303 \cdot R} \left(\frac{1}{T} - \frac{1}{298.15}\right) + b_1 \cdot M_1 + \cdots + b_{20} \cdot M_{20} \qquad (4)$$

206

207 2.3 Model validation

208 Validation of the final MLR model (Eq. 4) was performed using the QSARINS software, version

209 2.2.1 (www.qsar.it) which is developed by Gramatica et al. [20, 21].

210   2.3.1 Internal validation

211   The MLR model's capacity to predict portions of the training dataset was evaluated in an internal

212   validation process, using two techniques in QSARINS: the leave more out (LMO) cross-

213   validation and the Y-scrambling, which have been described previously [1, 21]. 1000 iterations were

214   used for the LMO cross-validation and the percentage of the excluded elements was set as 20%,

215   and 1000 iterations for Y-scrambling.

216   2.3.2 External validation

217   We also evaluated the model's ability to provide reliable predictions on new datasets by external

218   validation, using the splitting approach, which split the existing dataset (991 data points) into one

219   training dataset and one prediction dataset. The training dataset was used to generate regression

220   coefficients of the MLR model, and then the MLR model was applied to the prediction set to

221   examine the prediction performances of the model. Three kinds of splitting were performed

222   using existing options in the QSARINS software (see SI Section S4.1 for details) by random

223   percentage, by ordered response and by structure. We introduced a fourth kind of splitting by

224   studies, where all data points from certain studies were manually selected as the training set and

225   data points from remaining studies as the prediction set. If a consolidated material type only

226   includes data points from one study, all of these data points were assigned into the training set in

227   order to ensure that the MLR model constructed using the training set includes all consolidated

228   material types. The four types of splitting yielded similar sample sizes of approximately 800 data

229   points for the training set and 200 data points for the prediction set (SI Table S3).

230

231   ## 3. Results and Discussions

232   3.1 Temperature dependence

233   As described in Section 2.2.2, the temperature dependence of K_{ma} is determined by the enthalpy

234   of the partitioning between material and air, ΔH_{ma} (J/mol). Using the measured K_{ma} data for 54

235   chemicals in PU-ether from 15 °C to 95°C [4] (data are provided in SI Section S3), we obtained the

236   following correlation to estimate ΔH_{ma}:

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

238    $N = 54, R^2 = 0.786, R^2_{adj} = 0.782, \text{SE} = 2.85, \text{RMSE} = 2.80$

239    $\text{ANOVA: F} = 191, df = 1, p < 0.0001$

240 where ΔH_v is the chemical's enthalpy of vaporization (J/mol) obtained from ChemSpider

241 (www.chemspider.com).

242 This simple linear model shows good fitting of the experimental ΔH_{ma} data, with an adjusted R-

243 squared of 0.782, and the model fit is highly significant with an ANOVA p-value < 0.0001.

244 Figure 1 shows the scatter plot of predicted vs measured ΔH_{ma} and the residual plot, which

245 indicate good agreement with the 1:1 line and random distribution of residuals throughout the

246 dataset. These results suggest that there is indeed a linear relationship between ΔH_{ma} and ΔH_v in

247 PU-ether, and Eq.5 was also used as default to estimate ΔH_{ma} for all other materials.

248

249 3.2 Final QSPR and model fitting

250 Using the full dataset (991 data points) and Eq. 4, the final MLR model for predicting the solid

251 material-air partition coefficient is as follows:

252
$$\log_{10} K_{ma} = -0.38 + 0.63 \cdot \log_{10} K_{oa} + 0.96 \cdot \frac{\Delta H_{ma}}{2.303 \cdot R}\left(\frac{1}{T} - \frac{1}{298.15}\right) + b \tag{6}$$

253
$$N = 991, R^2 = 0.934, R^2_{adj} = 0.933, \text{SE} = 0.62, \text{RMSE} = 0.62$$

254
$$\text{ANOVA: F} = 597, df = 23, \text{p} < 0.0001$$

255 where K_{ma} is the dimensionless solid material-air partition coefficient, K_{oa} is the chemical's

256 dimensionless octanol-air partition coefficient at 25 °C, ΔH_{ma} is the enthalpy of the partitioning

257 between material and air (J/mol) which is given by Eq. 5, T is absolute temperature (K), and b is

258 the material-specific coefficients presented in Table 1. This model is provided as an excel model

259 in SI to facilitate application. The standard errors for the coefficients are also presented in Table

260 1. An SE of 0.63 of the final model (Eq. 6) indicates that the 95% confidence interval (CI) of the

261 predicted $\log K_{ma}$ is the predicted value ± 1.22, indicating that most of the predicted K_{ma} are

262 within a factor of 16 from the measured K_{ma}.

263 This MLR model shows excellent fitting of the experimental data, with an adjusted R-squared of

264 0.93 and a root mean square error (RMSE) of 0.62. The model fit is highly significant with an

265 ANOVA p-value smaller than 0.0001. Figure 2A shows the scatter plot of predicted vs measured

266 $\log K_{ma}$, which aligns well with the 1:1 line. The residual plot (Figure 1B) shows that the

267 residuals are distributed evenly throughout the dataset, and most residuals have absolute values

268 smaller than 2, again indicating the good fit of the linear model for the data.

269 This MLR model assumes that the correlation between $\log K_{ma}$ and the chemical's $\log K_{oa}$ is the

270 same across material types, which seems reasonable given the excellent model fitting. Plotting

271   the $logK_{ma}$ against chemical's $logK_{oa}$ for selected material types (Figure 3) confirmed that the

272   correlation between $logK_{ma}$ and the chemical's $logK_{oa}$ (i.e., the slopes of the fitted straight lines

273   in Figure 3) is similar but with slight differences across material types, indicating that a single

274   coefficient for $logK_{oa}$, as in the present QSPR model, might not be perfect. This could have been

275   accounted for by including interaction terms between $logK_{oa}$ and material types, but this would

276   introduce 21 more terms in the model without greatly improving the model fitting (SI Section

277   S5), so the interaction terms were not retained in the final QSPR model.

278   As described in the methods, this final MLR model uses EPISuite-estimated $logK_{oa}$ values as

279   predictors, since experimental $logK_{oa}$ are not available for all chemicals in the dataset. MLR

280   models developed using mixed $logK_{oa}$ values (i.e., for a chemical experimental $logK_{oa}$ is used

281   when available, otherwise EPISuite-estimated $logK_{oa}$ is used) also yielded similar results as the

282   final MLR model (adjusted R^2 ranged from 0.930 to 0.931, for details see SI Section S6),

283   indicating that the impact of experimental $logK_{oa}$ on the model is minimal.

284

285   3.3 Impact of each predictor

286   As shown in Eq. 6, the key predictors of the solid material-air partition coefficient are the

287   chemical's $logK_{oa}$, ΔH_v, temperature, and the solid material type. The regression coefficient for

288     $logK_{ow}$ is 0.63 and is highly significant ($p < 0.0001$), indicating that the material-air partition

289   coefficient increases with increasing $logK_{oa}$, which is consistent with findings from previous

290   studies [5, 6, 13].

291   The regression coefficient of the temperature term is 0.96 and is also highly significant ($p <$

292   0.0001), indicating that the K_{ma} decreases with higher temperature. Experimental data from

293   Kamprad et al. did show reduced K_{ma} with increased temperature, and it also makes intuitive

294   sense that at higher temperature the K_{ma} is lower leading to faster chemical migration from solid

295   material to air. As discussed in Section 3.1, the effect of temperature on K_{ma} also depends on the

296   ΔH_{ma}, which increases linearly with the chemical's enthalpy of vaporization ΔH_v.

297   The 21 dummy variables for the material types reflect the material dependency of the K_{ma}. As

298   "PU-ether" (polyurethane-ether) was used as the reference material in the regression, the value of

299   its coefficient *b* is zero (Table 1). For each of the other material types, the coefficient *b*,

300   determines the difference in $logK_{ma}$ between that material type and PU-ether. Chemicals in solid

301   material types with high values of *b* are more difficult to migrate to air than in those with low

302 values of *b*. The three material types with highest *b* coefficients are ethylene vinyl acetate (EVA),

303 latex and solvent-based paint and polyether ether ketone (PEEK) which are dense materials,

304 while the three types with lowest *b* coefficients are PU-ester (polyurethane-ester), PU-ether and

305 paper which tend to be porous materials. It should be noted that the data for a given consolidated

306 material type were gathered from different studies, and the composition and properties of the

307 material type may vary between studies, so the material type coefficients in Table 1 only

308 represent an average composition and partition behavior for the specific material types.

309 The significance of the material type coefficient only indicates that the coefficient *b*s of these

310 material types are significantly different from the reference material type, PU-ether, but if

311 another material type was selected as the reference material, the regression coefficients and

312 statistical significance of all materials would change. Thus, the insignificance of the regression

313 coefficient for "paper" (Table 1) does not indicate that this material type does not have a relevant

314 influence on the K_{ma}. As a result, we keep all 21 material type dummy variables in the final

315 regression to retain as much information as possible.

316 To better illustrate the impact of each predictor on the material-air partition coefficient, we

317 varied each predictor from the minimum to the maximum value in the entire dataset (991 data

318 points) while keeping the other predictors constant, and calculated the change in $\log K_{ma}$ using

319 the regression coefficients in the final QSPR (Eq. 6). Since the chemical's ΔH_v determines the

320 ΔH_{ma} which modifies the relationship between $\log K_{ma}$ and temperature, the impact of

321 temperature was calculated as two extremes using the minimum and maximum values of ΔH_v in

322 the entire dataset. As shown in Figure 4, the chemical's $\log K_{oa}$ has the highest impact on $\log K_{ma}$

323 among predictors. The impact of temperature on $\log K_{ma}$ is very low with the lowest value of

324 ΔH_v (22.3 kJ/mol), but the impact become moderate with the highest value of ΔH_v (75.6 kJ/mol).

325 This indicates that for a chemical with low enthalpy of vaporization, the $\log K_{ma}$ only changes

326 slightly with temperature, and vice versa. The material type also has a moderate impact on the

327 $\log K_{ma}$, which is similar to the impact of temperature with the highest value of ΔH_v. Overall, the

328 impact of material type is relatively small compared to the impact of chemical's $\log K_{oa}$,

329 indicating that the variation in $\log K_{ma}$ does not strongly depend on the solid material type, which

330 suggests the possibility of developing a generic QSPR to predict $\log K_{ma}$ in absence of material-

331 specific data.

332

333   3.4 Model validation results

334   3.4.1 Internal validation

335   The correlation coefficient for the LMO cross validation, Q^2_{LMO}, averages 0.93 (range: 0.90 –

336   0.95) for the 1000 iterations, and the root mean square error for cross validation ($RMSE_{cv}$)

337   averages 0.63. Both the Q^2_{LMO} and $RMSE_{cv}$ are similar to the R^2 and RMSE computed using the

338   full dataset, which is 0.93 and 0.62, respectively, indicating that the model is internally stable.

339   For Y-scrambling, the R^2_{Yscr}, Q^2_{Yscr} and $RMSE_{Yscr}$ for the 1000 iterations average 0.023, -0.028,

340   and 2.37, respectively, which are substantially different from the R^2, Q^2_{LMO} and RMSE of the

341   original model, indicating that that no correlation exists between the scrambled responses and the

342   predictors. Thus, the internal validation overall demonstrates that the final QSPR model (Eq. 6)

343   is robust and stable, and is not a result of chance correlation.

344   3.4.2 External validation

345   As described in Section 2.3.2, four types of splitting were used for external validation, including

346   splitting by random 20%, by ordered response, by structure, and by studies. Six criteria for

347   external validation, described in detail previously [1, 22, 23], were computed and are presented in

348   Table 2. For the first three types of splitting, the R^2_{ext} are higher than 0.9, and the other five

349   criteria all pass the threshold values and are higher than 0.9, indicating good predictive ability of

350   the models constructed from training set data. This is expected because the prediction sets

351   resulted from these three types of splitting are generally well within the applicability domain

352   (described in detail below) defined by the training sets (SI, Figures S1-S6), since the data points

353   were drawn either randomly or alternately.

354   For the splitting by studies, data from 22 studies were selected as the prediction set, while data

355   from 20 studies constituted the training set. This splitting can better represent a truly "external"

356   validation, since all data from one study were either be in the training or the prediction set. the

357   prediction ability of the model constructed from the training set is apparently reduced, as the

358   R^2_{ext} of this splitting dropped to 0.79, and the values of the other five criteria are lower than

359   those for the above three types of splitting. This is reasonable since the data variability is higher

360   between studies than within studies, so the prediction set might not be well within the AD

361   defined by the training set (SI, Figures S7-S10). Nonetheless, all validation criteria for this

362   splitting still pass the thresholds, indicating acceptable prediction ability (Table 2).

363   3.4.3 Applicability domain (AD)

364   It is important to define the AD of our QSPR model, as it can provide information on the

365   reliability of the model predictions [24] for future users who would like to use the model on new

366   chemicals. If the new chemicals are inside the AD, the model predictions are interpolated and are

367   more reliable. However, if the chemicals are outside the AD, the predictions are extrapolated and

368   less reliable [24].

369   For definition of the AD, the model being evaluated is the final QSPR model presented in Eq. 6,

370   and the training dataset thus refers to the full dataset including 991 data points. Three

371   complementary methods were applied to define the AD of the K_{ma} QSPR: the range of model

372   predictors, the leverage approach, and the PCA of the model predictors, which have been

373   described in detail previously [25].

374   For the range of predictors, the model has four predictors: $\log K_{oa}$, ΔH_v, temperature and material

375   type. The $\log K_{oa}$, ΔH_v, temperature of the training dataset range from 1.4 to 14.6, from 22.3 to

376   75.6 kJ/mol, and from 15 to 100 °C, respectively, defining the AD of the model. It is noteworthy

377   that the material type is a categorical variable, and the training set contains 22 consolidated

378   materials types, so the model's AD is also restricted to these 22 material types. For the leverage

379   approach, the critical value h^* for the diagonal values of the hat (h) matrix of the model was

380   calculated to be 0.0727, and the AD is defined as the h values less than h^* [21, 25]. For the PCA

381   approach, the AD is defined as the space between the minimum and maximum values of the PC1

382   and PC2 scores of the training dataset [21, 25], which range from -4.39 to 2.04 and from -4.52 to

383   2.22, respectively. For future model users, a new chemical should be considered "inside AD" if

384   viewed inside AD by all three methods, and be considered "outside AD" if viewed out of AD by

385   all three methods, otherwise it should be considered "borderline" [25].

386

387   3.5 Generic QSPR

388   In order to predict the K_{ma} without assigning material properties, we built a generic QSPR model

389   which does not include any material-specific variables using the same dataset. This model only

390   uses the chemical properties and temperature as predictors and is as follows:

391
$$\log_{10} K_{ma} = -0.37 + 0.75 \cdot \log_{10} K_{oa} + 1.29 \cdot \frac{\Delta H_{ma}}{2.303 \cdot R}\left(\frac{1}{T} - \frac{1}{298.15}\right) \qquad (7)$$

392
$$N = 991,\ R^2 = 0.80,\ R^2_{adj} = 0.80,\ SE = 1.08,\ RMSE = 1.08$$

393
$$\text{ANOVA: F} = 1943,\ df = 2,\ p < 0.0001$$

394   This model has a still relatively high adjusted R-squared of 0.80 compared to the 0.93 of the

395   regression with material coefficient (Eq. 6), indicating a good fit of experimental data (Figure 5).

396   As discussed in Section 3.2, the impact of the solid material type on $logK_{ma}$ is relatively small

397   compared to the impact of chemical properties, so $logK_{ma}$ can be predicted with reasonably high

398   accuracy without the material type as a predictor. This generic QSPR thus provides a relatively

399   reliable method to estimate the K_{ma} for various solid materials that may be difficult to assign a

400   material type listed in Table 1, which provides a more comprehensive and flexible coverage,

401   although with a slightly lower accuracy, for different chemical-material combinations than the

402   material-specific QSPR and can therefore greatly facilitate high-throughput evaluations of a

403   large variety of chemical-material combinations. However, it should be noted that although

404   without the material type as a predictor, this generic model was still developed using the

405   experimental data of our collection of 22 material types. Thus, this generic model best applies to

406   materials listed in Table 1 and similar materials, but may cause a large error for materials with

407   special properties, e.g. in presence of strong ionic forces, or of strong pseudo-solvation such that

408   some of the target adsorbate molecules take on a different structure within the material itself,

409   either due to ionization or tautomerization.

410   3.6 Limitations and future work

411   While the coverage of 22 consolidated materials and possibly any solid material as well as

412   inclusion of the effect of temperature are major advantages, the present model has several

413   limitations. First, the model does not consider chemical ionization or interaction with other

414   chemicals within a solid material, which may affect the chemical's partitioning between the

415   material and air. Second, the present model assumes that the relationship between ΔH_{ma} and

416   chemical's ΔH_v, derived from experimental ΔH_{ma} data for one material type "PU-ether", is the

417   same across different material types. Ideally, more experimental ΔH_{ma} data for different material

418   types are needed to verify this assumption or to develop unique ΔH_{ma}-ΔH_v relationships for

419   different material types.

420   Third, since for most K_{ma} datasets the material properties are not well characterized or provided

421   in the original publications, the classification of the consolidated material types is qualitative and

422   is simply based on material names, which may result in considerable variations in material

423   properties within one consolidated material type. In addition, even with the same composition,

424   different material structure may affect the material-air partitioning. Ideally, quantitative,

425   continuous properties of the solid materials, such as descriptors of the material's composition

426   and molecular structure, could be measured and entered into the model as numerical predictors,

427   so that the model can be more accurate for particular materials and can be extrapolated to other

428   material types outside the training dataset. In addition, if quantitative variables for material types

429   are used, interaction terms between chemical's $\log K_{oa}$ and material type variables can be added

430   to the model without introducing too many additional terms, which can improve model fitting, as

431   discussed in Section 3.2.

432   Fourth, many materials that appear in indoor environments are inhomogeneous, such as plywood,

433   gypsum board, carpet, concrete, and paper, which may have layers or portions with distinctive

434   properties. Thus, the K_{ma} values measured in experiments and the QSPR built on these

435   measurements likely only represent the material properties across the experiments. As a result,

436   one needs to use caution when applying the present QSPR to predict K_{ma}, especially for highly

437   inhomogeneous materials. Another important aspect related to heterogenicity is surface

438   partitioning versus bulk partitioning. Since the partitioning between solid material and air

439   happens mainly at the material surface, the surface properties may have an unusually large

440   influence on the apparent partitioning behavior. Therefore, for materials with a surface layer of

441   distinct properties, or materials with the same composition but different surface/bulk structures,

442   the present QSPR may not give a correct estimate of the K_{ma}. The distinct surface layer may be a

443   result of oxidative aging and soiling, which may change with time, or intrinsic features that are

444   time invariant. These problems again highlight the importance of using quantitative descriptors

445   of material compositions and structures as predictors in the QSPR.

446   Finally, the functional mechanisms of other influence factors such as relative humidity are

447   unclear, so they are not included in the QSPR. The effect of relative humidity on K_{ma} is likely

448   both chemical and material dependent [4, 9], which will require more in-depth research.

449

450   ## 4. Conclusions

451   A multiple linear regression model has been developed to predict the solid material-air partition

452   coefficients (K_{ma}) of organic compounds in various solid materials. Experimental K_{ma} data

453   collected from 43 studies were used to construct the regression model. The model uses three

454   continuous variables, chemical's $\log K_{oa}$, ΔH_v, and absolute temperature, as well as one

455   categorical variable, material type, as predictors. The model has been validated internally and

456 externally to be robust and stable, and have good predicting ability. The applicability domain of

457 the model, in terms of the range of predictors, includes chemical's logKoa between 1.4 to 14.6,

458 ΔH_v from 22.3 to 75.6 kJ/mol, temperature between 15 and 100 °C, and material type belonging

459 to the 22 consolidated types.

460 The main advantage of the present model is that it is applicable for a wide range of chemical-

461 material-temperature combinations, which is more comprehensive than the correlation methods

462 developed in previous studies which were specific for one solid material and often at room

463 temperature. Moreover, a generic model is also developed which is able to give relatively

464 accurate estimates of K_{ma} without assigning a particular material type, making it suitable for

465 high-throughput assessments of the chemical releases from solid materials and subsequent

466 consumer exposures.

467

471

472

473

474

475

476

477 Tables and Figures

478 Table 1. Regression coefficients for Eq. 6

Variable	Coefficient	SE [a]	P-value
Intercept	-0.38	0.06	< 0.001
$\log_{10}K_{oa}$	0.63	0.01	< 0.001
$\Delta H_{ma}(1/T - 1/298.15)/2.303R$	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.

479

[c] Includes oriented strand board (OSB), particleboard, medium-density board and high-density board.

480

481

482 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.

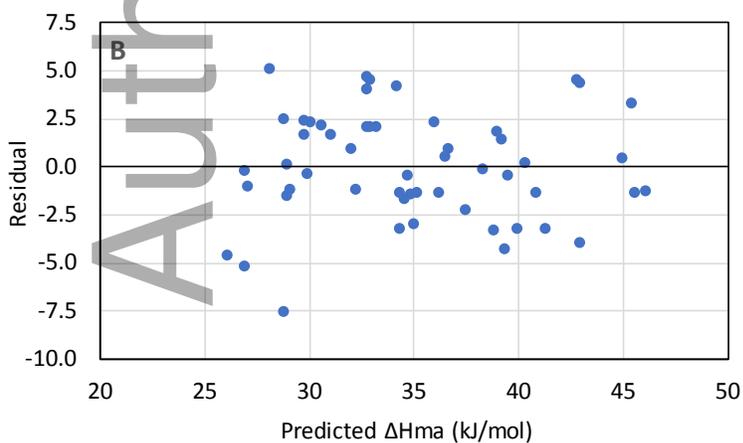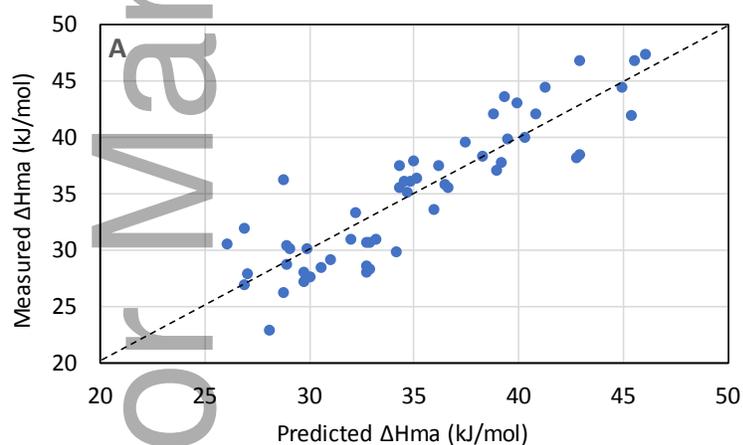Q^2_{F3}: correlation coefficient proposed by Consonni et al.

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

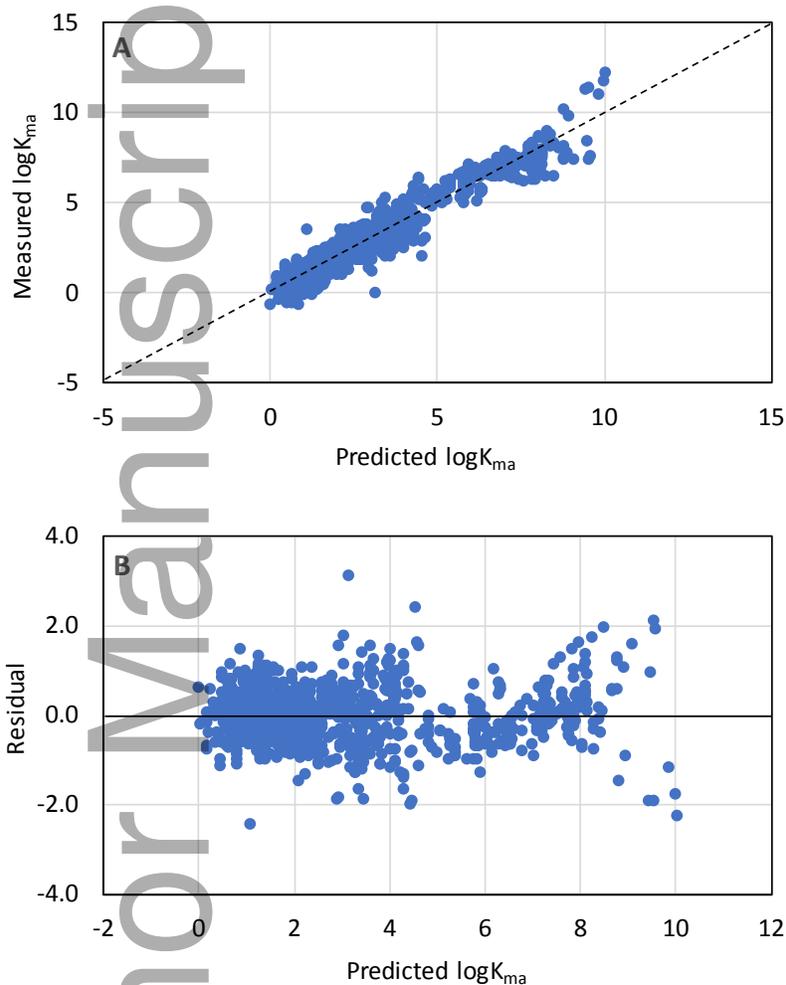CCC: concordance correlation coefficient proposed by Chirico and Gramatica.

483

484

485

486

487   Figure 1. (A) Measured Enthalpy of material-air partitioning (ΔH_{ma}) and (B) residuals as a

488   function of the (ΔH_{ma}) predicted from chemical enthalpy of vaporization (ΔH_v - Eq. 5). The

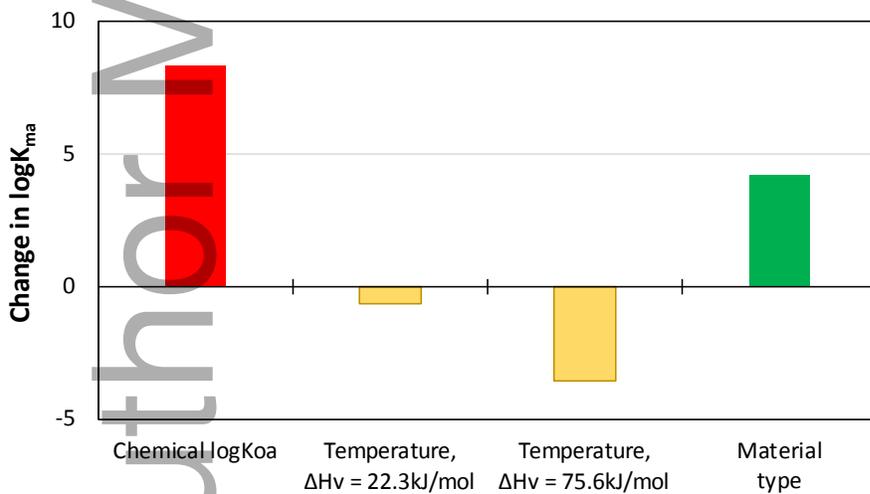489   dotted line in (A) indicates the 1:1 line.

490

491

492   Figure 2. (A) measured $logK_{ma}$ and (B) residuals as a function of $logK_{ma}$ predicted by the final

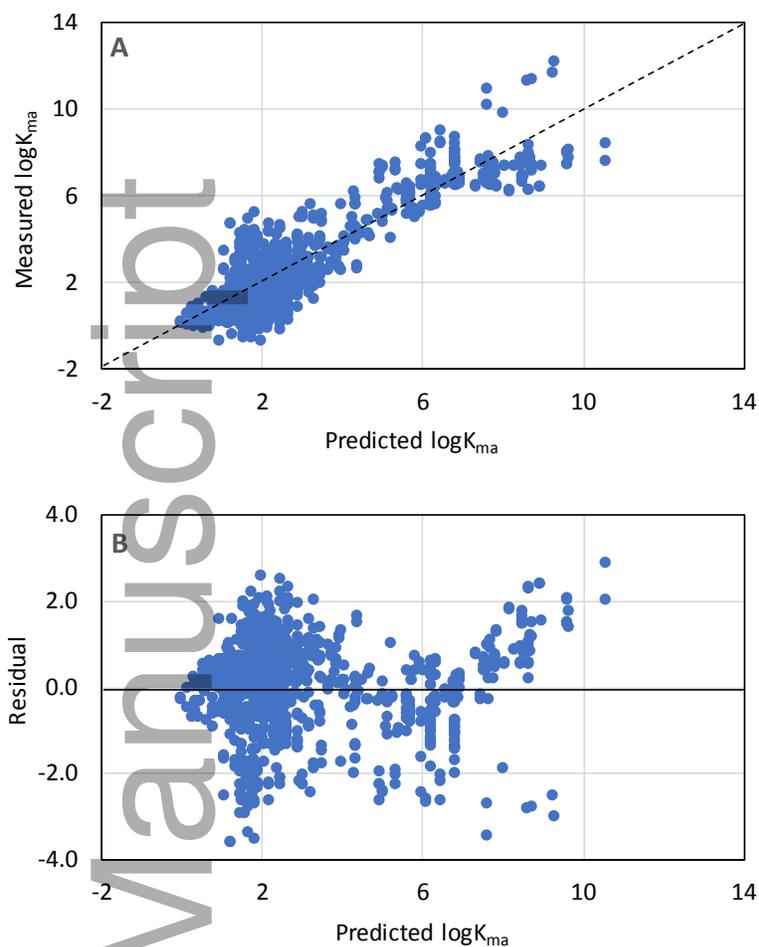493   QSPR (Eq. 6). The dotted line in (A) indicates the 1:1 line.

494

495

496
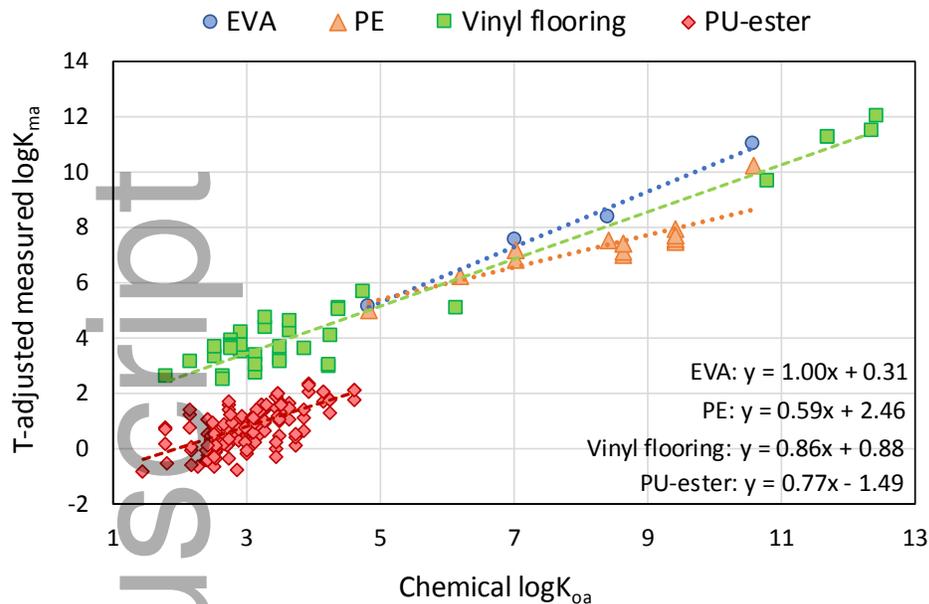
Figure 3. temperature adjusted measured $\log K_{ma}$ as a function of $\log K_{oa}$ for selected material

498   types including EVA, PE, vinyl flooring, and PU-ester.

499

500

501

502   Figure 4. Change in $\log K_{ma}$ with respect to the change in each predictor, from minimum to

503   maximum values within the entire dataset.

504

505

Figure 5. (A) measured $\log K_{ma}$ and (B) residuals as a function of $\log K_{ma}$ predicted by the generic

QSPR (Eq. 7). The dotted line in (A) indicates the 1:1 line.

508

509

510

511

512

513

514 References

515 1.   Huang L, Fantke P, Ernstoff A, et al. A Quantitative Property-Property Relationship for

516   the Internal Diffusion Coefficients of Organic Compounds in Solid Materials. *Indoor Air*.

517   2017.

518   2.   Huang L and Jolliet O. A parsimonious model for the release of chemicals encapsulated

519   in products. *Atmos Environ.* 2016; 127: p. 223-235.

520   3.   Liu Z, Ye W, Little JC. Predicting emissions of volatile and semivolatile organic

521   compounds from building materials: a review. *Build Environ.* 2013; 64: p. 7-25.

522   4.   Kamprad I and Goss K-U. Systematic investigation of the sorption properties of

523   polyurethane foams for organic vapors. *Analytical chemistry.* 2007; 79(11): p. 4222-4227.

524   5.   Bartkow ME, Hawker DW, Kennedy KE, et al. Characterizing uptake kinetics of PAHs

525   from the air using polyethylene-based passive air samplers of multiple surface area-to-

526   volume ratios. *Environ Sci Technol.* 2004; 38(9): p. 2701-2706.

527   6.   Kennedy KE, Hawker DW, Müller JF, et al. A field comparison of ethylene vinyl acetate

528   and low-density polyethylene thin films for equilibrium phase passive air sampling of

529   polycyclic aromatic hydrocarbons. *Atmos Environ.* 2007; 41(27): p. 5778-5787.

530   7.   Morrison G, Li H, Mishra S, et al. Airborne phthalate partitioning to cotton clothing.

531   *Atmos Environ.* 2015; 115: p. 149-152.

532   8.   Morrison G, Shakila N, Parker K. Accumulation of gas-phase methamphetamine on

533   clothing, toy fabrics, and skin oil. *Indoor air.* 2015; 25(4): p. 405-414.

534   9.   Zhao D, Little JC, Cox SS. Characterizing polyurethane foam as a sink for or source of

535   volatile organic compounds in indoor air. *Journal of environmental engineering.* 2004;

536   130(9): p. 983-989.

537   10.   Bodalal A, Zhang J, Plett E, et al. Correlations between the internal diffusion and

538   equilibrium partition coefficients of volatile organic compounds (VOCs) in building

539   materials and the VOC properties. *ASHRAE Trans.* 2001; 107: p. 789.

540   11.   Guo Z. Review of indoor emission source models. Part 2. Parameter estimation. *Environ*

541   *Pollut.* 2002; 120(3): p. 551-564.

542   12.   Chaemfa C, Barber JL, Gocht T, et al. Field calibration of polyurethane foam (PUF) disk

543   passive air samplers for PCBs and OC pesticides. *Environ Pollut.* 2008; 156(3): p. 1290-

544   1297.

545   13.   Shoeib M and Harner T. Characterization and comparison of three passive air samplers

546   for persistent organic pollutants. *Environ Sci Technol.* 2002; 36(19): p. 4142-4151.

547 14. Holmgren T, Persson L, Andersson PL, et al. A generic emission model to predict release
548   of organic substances from materials in consumer goods. *Science of the Total*
549   *Environment.* 2012; 437: p. 306-314.

550 15. Booij K, Hofmans HE, Fischer CV, et al. Temperature-dependent uptake rates of
551   nonpolar organic compounds by semipermeable membrane devices and low-density
552   polyethylene membranes. *Environ Sci Technol.* 2003; 37(2): p. 361-366.

553 16. Adams RG, Lohmann R, Fernandez LA, et al. Polyethylene devices: Passive samplers for
554   measuring dissolved hydrophobic organic compounds in aquatic environments. *Environ*
555   *Sci Technol.* 2007; 41(4): p. 1317-1323.

556 17. Liu Y, Zhou X, Wang D, et al. A prediction model of VOC partition coefficient in porous
557   building materials based on adsorption potential theory. *Build Environ.* 2015; 93: p. 221-
558   233.

559 18. Zhang Y, Luo X, Wang X, et al. Influence of temperature on formaldehyde emission
560   parameters of dry building materials. *Atmos Environ.* 2007; 41(15): p. 3203-3216.

561 19. USEPA, *Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11*, 2012,
562   United States Environmental Protection Agency: Washington, DC.

563 20. Gramatica P, Cassani S, Chirico N. QSARINS-chem: Insubria datasets and new
564   QSAR/QSPR models for environmental pollutants in QSARINS. *J Comp Chem.* 2014;
565   35(13): p. 1036-1044.

566 21. Gramatica P, Chirico N, Papa E, et al. QSARINS: A new software for the development,
567   analysis, and validation of QSAR MLR models. *J Comp Chem.* 2013; 34(24): p. 2121-
568   2132.

569 22. Chirico N and Gramatica P. Real external predictivity of QSAR models. Part 2. New
570   intercomparable thresholds for different validation criteria and the need for scatter plot
571   inspection. *J Chem Inf Model.* 2012; 52(8): p. 2044-2058.

572 23. Chirico N and Gramatica P. Real external predictivity of QSAR models: how to evaluate
573   it? Comparison of different validation criteria and proposal of using the concordance
574   correlation coefficient. *J Chem Inf Model.* 2011; 51(9): p. 2320-2335.

575 24. Gramatica P. Principles of QSAR models validation: internal and external. *QSAR &*
576   *combinatorial science.* 2007; 26(5): p. 694-701.

577 25. Cassani S and Gramatica P. Identification of potential PBT behavior of personal care

578   products by structural approaches. *Sustain Chem Pharm.* 2015; 1: p. 19-27.

579

580