
SAR and QSAR in Environmental Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gsar20

# Predicting skin permeability from complex chemical mixtures: incorporation of an expanded QSAR model

G. Xu [a], J.M. Hughes-Oliver [b], J.D. Brooks [c] & R.E. Baynes [c]

[a] Department of Statistics, North Carolina State University, Raleigh, USA

[b] Department of Statistics, Volgenau School of Engineering, George Mason University, Fairfax, USA

[c] Center for Chemical Toxicology Research and Pharmacokinetics, North Carolina State University College of Veterinary Medicine, Raleigh, USA

Published online: 14 Jun 2013.

To cite this article: G. Xu, J.M. Hughes-Oliver, J.D. Brooks & R.E. Baynes (2013) Predicting skin permeability from complex chemical mixtures: incorporation of an expanded QSAR model, SAR and QSAR in Environmental Research, 24:9, 711-731, DOI: 10.1080/1062936X.2013.792875

To link to this article: http://dx.doi.org/10.1080/1062936X.2013.792875

*SAR and QSAR in Environmental Research*, 2013
Vol. 24, No. 9, 711–731, http://dx.doi.org/10.1080/1062936X.2013.792875

Taylor & Francis
Taylor & Francis Group

# Predicting skin permeability from complex chemical mixtures: incorporation of an expanded QSAR model

G. Xu[a], J.M. Hughes-Oliver[b], J.D. Brooks[c] and R.E. Baynes[c]*

[a]*Department of Statistics, North Carolina State University, Raleigh, USA;* [b]*Department of Statistics, Volgenau School of Engineering, George Mason University, Fairfax, USA;* [c]*Center for Chemical Toxicology Research and Pharmacokinetics, North Carolina State University College of Veterinary Medicine, Raleigh, USA*

(*Received 1 February 2013; in final form 23 March 2013*)

Quantitative structure–activity relationship (QSAR) models have been widely used to study the permeability of chemicals or solutes through skin. Among the various QSAR models, Abraham's linear free-energy relationship (LFER) model is often employed. However, when the experimental conditions are complex, it is not always appropriate to use Abraham's LFER model with a single set of regression coefficients. In this paper, we propose an expanded model in which one set of partial slopes is defined for each experimental condition, where conditions are defined according to solvent: water, synthetic oil, semi-synthetic oil, or soluble oil. This model not only accounts for experimental conditions but also improves the ability to conduct rigorous hypothesis testing. To more adequately evaluate the predictive power of the QSAR model, we modified the usual leave-one-out internal validation strategy to employ a leave-one-solute-out strategy and accordingly adjust the $Q^2_{LOO}$ statistic. Skin permeability was shown to have the rank order: water > synthetic > semi-synthetic > soluble oil. In addition, fitted relationships between permeability and solute characteristics differ according to solvents. We demonstrated that the expanded model ($r^2$ = 0.70) improved both the model fit and the predictive power when compared with the simple model ($r^2$ = 0.21).

**Keywords:** Abraham's LFER model; adjusted $Q^2_{LOO}$; expanded LFER model; leave-one-solute-out; QSAR



## 1. Introduction

Quantitative structure–activity relationship (QSAR) modelling is a popular method used in the chemical and biological sciences, and in industry to relate biological activity with the chemical structures or properties [1,2]. These chemical properties can often be identified by molecular descriptors [2] obtained through certain computer software programs such as the commercially available ADME Boxes from ACD/Labs [3]. In fact, one popular category of QSAR modelling is to use a regression model with these molecular descriptors as the predictors and the biological activity of interest as the response variable [2,4]. Among the various application fields of QSAR models, skin permeability is an area where QSAR models are widely utilized. For example, a QSAR model can be adopted to study the permeability of chemical solutes through skin in the metalworking and agricultural industries [5,6]. Such a QSAR model is of great importance in the risk assessment of skin absorption.

*Corresponding author. Email: rebaynes@ncsu.edu

Abraham's linear free-energy relationship (Abraham's LFER) model [4] is a particular type of QSAR model that is widely used in skin permeability studies to quantify the relationship between skin permeability and molecular descriptors that characterize the solvation process. Abraham's LFER model is easy to use and interpret; however, when the experimental conditions are complex, a single model is insufficient. By insufficient, we mean that the fitted model may not be able to account for the effects of different levels of factors (e.g. the solvent type, solvent concentration, etc.) in the experiment, and this leads to a model with poor fit statistics and low predictive power. To account for heterogeneity introduced by the experimental factors, we propose an expanded Abraham's model in which one set of partial slopes is defined for each experimental condition. There are other techniques of QSAR modelling, such as nonlinear models, nearest-neighbour methods, artificial neural networks etc. [7]. These advanced approaches may result in more sophisticated models. We restrict our paper to Abraham's LFER model due to its popularity in modelling dermal permeability and also its ease of implementation and interpretation.

To assess the predictive power of a QSAR model, it is recommended that validation, such as leave-one-out cross-validation, be utilized [8]. The leave-one-out idea is simple, that is, leave one observation out at a time and fit a regression model with the remainder of the data to predict the left-out observation. This process is repeated until each observation is left out and predicted by the remainder of the data. The residuals, the differences between the observed values and the predicted values, are used to measure the predictive power. However, this leave-one-out strategy has an obvious disadvantage for a dataset where there are several pseudo or real replicates for each solute. A pseudo replicate occurs when the same solute is treated under multiple experimental conditions, while a true replicate occurs when the solute is treated multiple times under the same experimental condition. Leave-one-out validation assumes that the observations are independent, while in a dataset the replicates for the same solute are often correlated. Thus, a new internal validation method that we call "leave-one-solute-out" is proposed to adjust for the dependence.

The primary aim of this study was to develop a robust QSAR model within a LFER framework that adequately estimates solute permeability in skin in four formulations that are representative of occupational exposure in the metal fabrication industry. Except for water, these industrial formulations are complex mixtures with multiple solute–formulation interactions within any one formulation. Our objective is not to identify each of these interactions but to test the feasibility of simple and expanded LFER models to estimate skin permeability in four very diverse exposure conditions relevant to occupational exposure, and ultimately to glean some mechanistic interpretation within the construct of a LFER.



## 2. Materials and methods

### 2.1 *Experiments and data*

The experiments used flow-through (FT) diffusion cells with porcine skin obtained from weanling female Yorkshire pigs. Experimental details of this type of experiment can be found in Baynes et al. [9] and Riviere and Brooks [10]. In total, 23 solutes (see Table 1) were selected for experimental study from a large collection of 4534 solutes using the two-phase strategy of Xu et al. [6] to select a diverse and representative training set. Each solute was dissolved in four different metalworking fluid mixtures (referred to as MWF formulations):

Table 1. Solute list and descriptor values.

| *Solute ID* | *Solute* | *E* | *S* | *A* | *B* | *V* |
|---|---|---|---|---|---|---|
| 1 | Heptan-2-ol | 0.19 | 0.36 | 0.33 | 0.56 | 1.1536 |
| 2 | 2-Ethylthiophene | 0.69 | 0.58 | 0 | 0.13 | 0.9229 |
| 3 | Ethyl-4-methylpentanoate | 0.07 | 0.58 | 0 | 0.40 | 1.3102 |
| 4 | 2-Methyl-3-methoxypyrazine | 0.65 | 0.81 | 0 | 0.60 | 0.9747 |
| 5 | 3-Butylpyridine | 0.63 | 0.73 | 0 | 0.45 | 1.2389 |
| 6 | 1-Fluoronaphthalene | 1.14 | 0.97 | 0 | 0.13 | 1.1030 |
| 7 | 2,3,5-Trimethylphenol | 0.86 | 0.84 | 0.52 | 0.41 | 1.1978 |
| 8 | 3,4-Dimethylbenzyl-alcohol | 0.83 | 0.90 | 0.39 | 0.61 | 1.1978 |
| 9 | 4-Chloro-3-methylphenol | 0.92 | 1.02 | 0.67 | 0.22 | 1.0384 |
| 10 | Methyl-2-chlorobenzoate | 0.82 | 0.99 | 0 | 0.47 | 1.1950 |
| 11 | Dimethyl-pimelate | 0.15 | 1.00 | 0 | 0.77 | 1.5255 |
| 12 | 1-Methylisoquinoline | 1.28 | 1.03 | 0 | 0.51 | 1.1852 |
| 13 | 6-Methylindole | 1.20 | 1.09 | 0.44 | 0.49 | 1.0873 |
| 14 | 3-Nitrobenzonitrile | 1.02 | 1.60 | 0 | 0.47 | 1.0453 |
| 15 | Methyl-4-hydroxybenzoate | 0.90 | 1.37 | 0.69 | 0.45 | 1.1313 |
| 16 | Dimethyl-phthalate | 0.78 | 1.40 | 0 | 0.84 | 1.4288 |
| 17 | 2-Acetylnaphthalene | 1.50 | 1.40 | 0 | 0.55 | 1.3829 |
| 18 | 1-Naphthaleneethanol | 1.54 | 1.31 | 0.3 | 0.70 | 1.4259 |
| 19 | Pirimicarb | 1.18 | 1.34 | 0 | 1.34 | 1.8945 |
| 20 | Diphenylsulfone | 1.57 | 2.15 | 0 | 0.70 | 1.6051 |
| 21 | 4,4'-Isopropylidenediphenol | 1.61 | 1.56 | 0.99 | 0.91 | 1.8643 |
| **22 | Dibutylamine | 0.11 | 0.30 | 0.08 | 0.69 | 1.3356 |
| **23 | Methyl-trichloroacetate | 0.41 | 0.81 | 0 | 0.42 | 0.9729 |

**Not absorbed by fibres, no data available.

(1) 100% Water (Water)
(2) 95% Water + 5% Synthetic oil (SYN)
(3) 95% Water + 5% Semi-synthetic oil (SSYN)
(4) 95% Water + 5% Soluble oil (SO)

Generic metalworking fluids, which include soluble oil, semi-synthetic oil, and synthetic oil are generally alkaline (pH∼9) solutions diluted with water before use [11]. Soluble oil concentrates contain 60–90% mineral oil, semi-synthetic oil concentrates contain 2–30% mineral oil, and synthetic oil concentrates contain no mineral oil, but contain various carboxylic acid salts, ethanolamines, ethyleneglycols, and plant seed oils [12].

The solutions were applied topically to the skin in the diffusion cells and the solutes that permeated the pig skin were collected in the receptor fluid of the FT cells over time. The absorbed amount of each solute was denoted as absorption (μg). The steady state flux (Js) was calculated as the slope of the cumulative absorption/area vs. time curve. Because our doses were unsaturated infinite doses, this is not a maximal slope as described by Zhang et al. [13] where Js corresponds to Jmax for saturated infinite doses. The permeability coefficient (*Kp* in cm/hr) was calculated as the ratio between the steady state flux and the initial drug concentration applied to the pig skin surface (μg/ml). There were seven replicates per solute per MWF formulation, but several of the solutes were absorbed at levels below the limit of detection of our gas chromatograph. The gas chromatograph assays targeted only parent compounds, while levels of metabolites and degradation products of the test compounds were not assessed. As a result, 387 permeability coefficients were obtained as the training set,

in which 119 observations were from water, 92 from synthetic oil, 84 from semi-synthetic oil and 92 from soluble oil.

### 2.2 *Abraham's LFER model*

Among the thousands of molecular descriptors of each solute, it has been found that the five solvatochromic descriptors ($E$, $S$, $A$, $B$, $V$) are most relevant to the solvation process during skin permeation [2,4]. These descriptors represent different characteristics of solutes involved in the solvation process, specified as follows. $E$ is the solute excess molar refraction, $S$ is the solute dipolarity/polarizability, $A$ is the overall hydrogen bond acidity, $B$ is the overall hydrogen bond basicity, and $V$ is the McGowan characteristic volume. For most solutes, $V$ can be calculated directly, $E$ can be obtained experimentally or calculated, while $A$, $B$, and $S$ are experimentally derived. (We, however, simply obtained them from software; see below.)

Abraham and Martins [4] proposed the following general LFER model to study dermal absorption:

$$SP = \beta_0 + \beta_1 E + \beta_2 S + \beta_3 A + \beta_4 B + \beta_5 V,$$

where SP is the property of interest for the solutes. In this paper, logarithm of the permeability coefficient, log $Kp$, is the property of interest. Abraham's LFER model is shown in Equation (1).

$$\log Kp = \beta_0 + \beta_1 E + \beta_2 S + \beta_3 A + \beta_4 B + \beta_5 V \quad (1)$$

ADME Boxes 4.95, a commercial software package from ACD/Labs (www.acdlabs.com) [3] was used to identify the $E$, $S$, $A$, $B$, and $V$ descriptors for all 23 solutes used in the experiment. Solute names along with their descriptor values are shown in Table 1. Summary statistics of the response variable log $Kp$ and the predictors ($E$, $S$, $A$, $B$, and $V$) are shown in Table 2. The complete data are available as supplementary information, under the 'Supplemental' tab on the article's web page: http://dx.doi.org/10.1080/1062936X.2013.792875.

### 2.3 *Insufficiency of Abraham's model*

While Abraham's LFER model is simple and relatively easy to interpret, it is not always sufficient, especially for large datasets under complex experimental conditions. As is clear in Equation (1), this model makes no allowances for differences due to experimental conditions;

Table 2. Summary statistics for response variable and descriptors.

| *Variable* | *Minimum* | *Lower Quartile* | *Mean* | *Median* | *Upper Quartile* | *Maximum* | *Std Dev* | *N* |
|---|---|---|---|---|---|---|---|---|
| log*Kp* | −4.269 | −2.750 | −2.354 | −2.242 | −1.930 | −1.172 | 0.574 | 387 |
| E | 0.070 | 0.690 | 0.990 | 0.920 | 1.280 | 1.570 | 0.379 | 387 |
| S | 0.360 | 0.810 | 1.080 | 1.020 | 1.370 | 2.150 | 0.425 | 387 |
| A | 0 | 0 | 0.193 | 0 | 0.39 | 0.69 | 0.244 | 387 |
| B | 0.130 | 0.410 | 0.501 | 0.510 | 0.610 | 1.340 | 0.233 | 387 |
| V | 0.923 | 1.045 | 1.209 | 1.185 | 1.383 | 1.895 | 0.211 | 387 |

predictions of log *Kp* remain the same, irrespective of experimental conditions. This insufficiency can be illustrated by our dataset.

Figure 1 shows boxplots of the response variable log *Kp* grouped by MWF formulations (i.e. experimental conditions). Each boxplot has five lines representing: the minimum value of the data; the value below which 25 percent of the data fall, also known as the first quartile; the middle value of the data, or median; the value below which 75 percent of the data fall, also known as the third quartile; and the maximum of the data. The diamond represents the sample mean. From Figure 1, one can see that there is a clear trend of the response variable across the MWF formulations. The mean values of the four types of MWF formulations follow the pattern: water > SYN > SSYN > SO. This indicates that a single regression equation may not be sufficient to fit the data well.

To show how the regression coefficients may differ quantitatively under various experimental conditions, the original data were grouped into subsets according to the MWF formulations and an Abraham's LFER model was fit for each subset. The estimated regression coefficients and corresponding standard errors (in parentheses) are listed in Table 3. It can be seen that even accounting for the standard errors, the regression coefficients vary a great deal. For example, the 95% confidence intervals (calculated as the estimate$\pm$1.96*SE) for the regression coefficient of descriptor *E* are (0.61, 1.23) from water; (0.35, 1.05) from synthetic oil; ($-$0.28, 0.26) from semi-synthetic oil; and ($-$0.66, $-$0.26) from soluble oil.



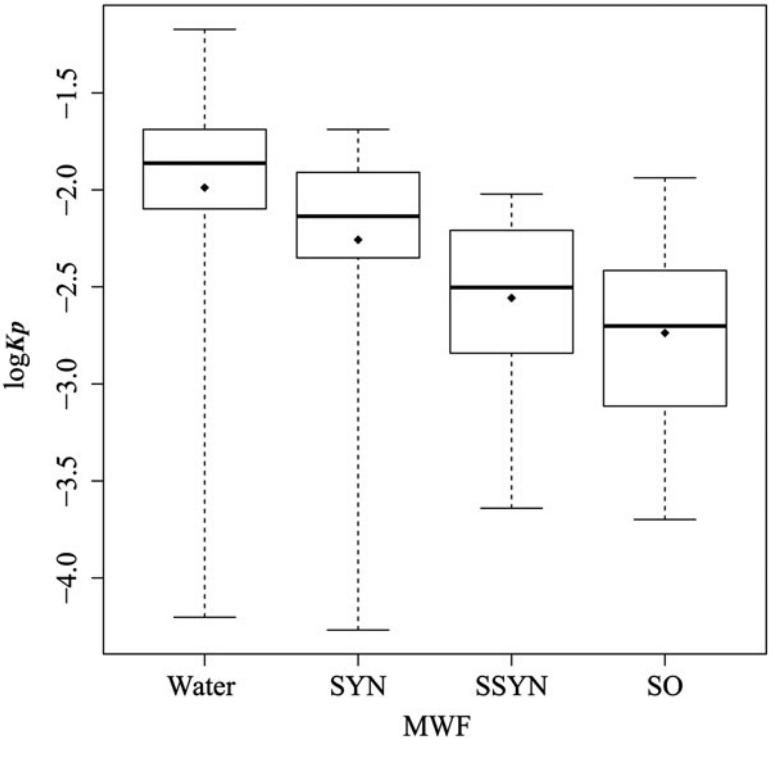


Figure 1. Boxplot of log *Kp* across MWF formulations of the training set. Numbers of observations in each MWF are: 119 (water), 92 (SYN), 84 (SSYN) and 92 (SO). The mean permeability follows the order: water > SYN > SSYN > SO. The five lines from top to bottom of each boxplot represent: maximum, 3rd quartile, median, 1st quartile and minimum. The diamond represents the sample mean.

Differences in the regression coefficients are also graphically displayed in Figure 2. The first panel from the left in Figure 2 shows the estimated partial slopes for $E$ with its confidence intervals from the four MWF formulations: each vertical line with two bars at the ends represents the 95% confidence interval of $\beta_1$ for one MWF formulation while the circle

Table 3. Regression coefficients from different types of MWF formulations.

| *MWF* | *Intercept* | *E* | *S* | *A* | *B* | *V* |
|---|---|---|---|---|---|---|
| Water | −1.14(0.26) | 0.92(0.16) | −0.78(0.14) | 0.03(0.14) | −0.77(0.26) | −0.39(0.33) |
| SYN | −0.55(0.27) | 0.71(0.18) | −0.62(0.16) | −0.20(0.17) | −0.28(0.28) | −1.28(0.35) |
| SSYN | −0.92(0.22) | −0.01(0.14) | −0.24(0.13) | −0.30(0.13) | 1.40(0.22) | −1.67(0.27) |
| SO | −0.84(0.18) | −0.46(0.10) | 0.20(0.10) | 0.10(0.10) | 2.00(0.18) | −2.21(0.20) |



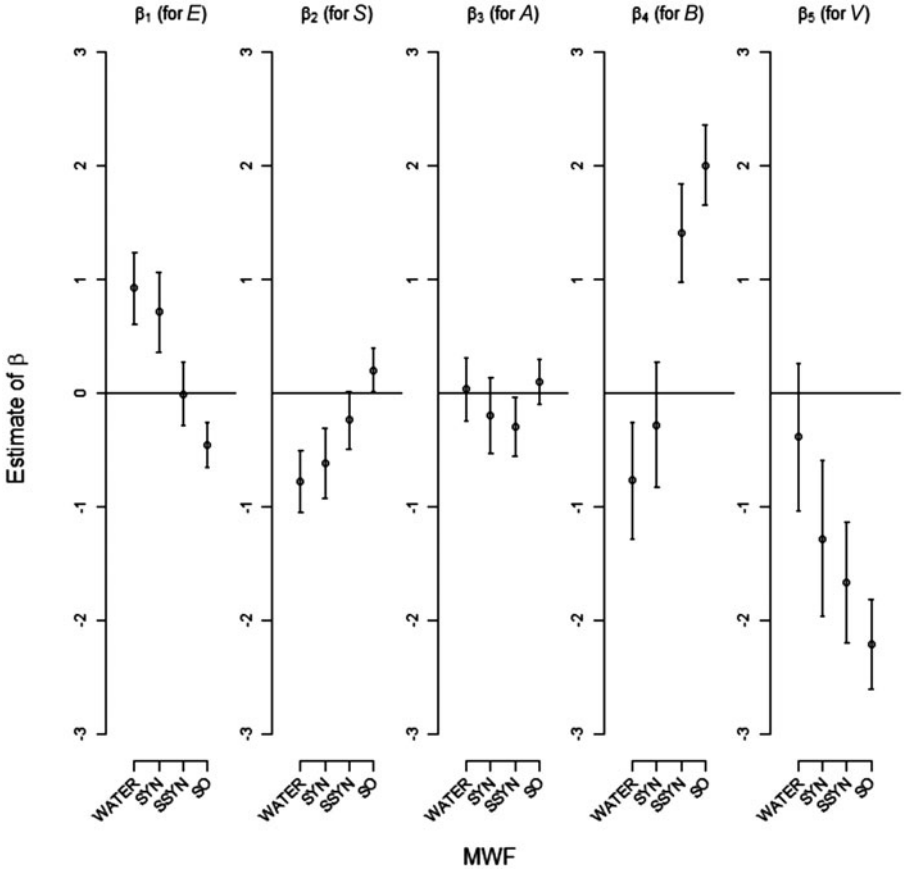


Figure 2. Estimated partial slopes (circles) for molecular descriptors $E$, $S$, $A$, $B$ and $V$ across the four MWFs of the expanded model of Equation (4). Also shown are 95% confidence intervals for these slopes, as vertical lines with two bars at the ends. Each panel corresponds to a different molecular descriptor, and within a panel the four sets of estimates and confidence intervals correspond to a different MWF formulation. For example, molecular descriptor $B$ (hydrogen bond basicity) has estimated slopes that are negative for MWF formulations water and synthetic oil, and clearly positive for MWF formulations semi-synthetic and soluble oil.

is the estimated value. One can see that the estimated partial slopes for descriptor $E$ can be very different across different MWF formulations, even after accounting for the standard errors. Similarly, the remaining panels in Figure 2 show estimated partial slopes and confidence intervals for descriptors $S$, $A$, $B$, and $V$, respectively.

### 2.4 *Fully expanded model and condensed model*

From Section 2.3 one can see that a single Abraham's LFER model may not do well at fitting a dataset resulting from varying experimental conditions. For our dataset, one simple and natural solution might be to fit a separate Abraham's model for each of the four MWF formulations. For the four formulations, four simple models can be fit. However, when experimental conditions get more complex, for example more types of MWF formulations, the number of models to fit could become quite large. The use and evaluation of that many models could become extensive.

Instead of fitting one simple model for each experimental condition, we propose an expanded version of Abraham's LFER model to adjust for the heterogeneity introduced by various MWF formulations. In the expanded model of Equation (2), the partial slopes of $E$, $S$, $A$, $B$, and $V$ are allowed to differ according to MWF formulations:

$$\begin{aligned}\log Kp_{i,j} = {} & F_{1j}(\beta_{01} + \beta_{11}E_j + \beta_{21}S_j + \beta_{31}A_j + \beta_{41}B_j + \beta_{51}V_j) \\ & + F_{2j}(\beta_{02} + \beta_{12}E_j + \beta_{22}S_j + \beta_{32}A_j + \beta_{42}B_j + \beta_{52}V_j) \\ & + F_{3j}(\beta_{03} + \beta_{13}E_j + \beta_{23}S_j + \beta_{33}A_j + \beta_{43}B_j + \beta_{53}V_j) \\ & + F_{4j}(\beta_{04} + \beta_{14}E_j + \beta_{24}S_j + \beta_{34}A_j + \beta_{44}B_j + \beta_{54}V_j)\end{aligned} \tag{2}$$

where $\beta_{li}$ denotes the partial slope for MWF formulation $i$ ($i = 1$ for water, $i = 2$ for SYN, $i = 3$ for SSYN, $i = 4$ for SO) of descriptor $l$ (with $l = 0$ for intercept, $l = 1$ for $E$, $l = 2$ for $S$, $l = 3$ for $A$, $l = 4$ for $B$, and $l = 5$ for $V$). For example, $\beta_{11}$ is the partial slope for descriptor $E$ from MWF formulation 1 (Water). Letting $j$ be the index of observation, the variable $F_{ij} = 1$ if observation $(i,j)$ comes from MWF formulation $i$ and $F_{ij} = 0$ otherwise. The expanded model in Equation (2) is quite large, having a maximum of four intercepts (one for each MWF formulation) and $5 \times 4 = 20$ partial slopes (slopes corresponding to each of $E$, $S$, $A$, $B$, and $V$ for each MWF formulation). The part in each parenthesis in Equation (2) is a separate Abraham's LFER model for a given experimental condition. For example, the first portion $\beta_{01} + \beta_{11}E_j + \beta_{21}S_j + \beta_{31}A_j + \beta_{41}B_j + \beta_{51}V_j$ is the regression model for the MWF formulation water.

There are many benefits to fitting a single large model of the form in Equation (2) rather than four separate models (one for each MWF formulation). Obvious benefits include improved fit statistics and higher predictive power, with details to be discussed in Section 3. Another benefit is that users of the model can conveniently conduct hypothesis testing for significance of the MWF formulation effect. For example, suppose one wishes to determine whether the partial slopes should change as a function of MWF formulation. This is equivalent to testing $H_0$: $\beta_{l1} = \beta_{l2} = \beta_{l3} = \beta_{l4}$ for all $l = 1,2,3,4,5$ thus collapsing from a maximum possible $5 \times 4 = 20$ partial slopes to only $5 \times 1 = 5$ partial slopes (the intercepts are not required to be the same). A pairwise test between any two of the MWF formulations can also be performed, from which one can find out which two MWF formulations are more "alike" in regard to the impact on the permeation process.

It is possible that some parameters in Equation (2) are not significant, thus sometimes a more condensed model is attainable. To be specific, to get a more condensed model, the

non-significant parameters in Equation (2) can be removed and a lack-of-fit $F$-test will be performed. If the $F$-statistic of the lack-of-fit test is not significant, the final condensed model will be defined. Details will be discussed in Section 3.

## 3. Results and discussion

### 3.1 *Fitted simple model and expanded model*

SAS 9.3 [14] and R [15] were used for data analysis. The simple Abraham's LFER model in Equation (1) was fit and Equation (3) was obtained, where the numbers in parentheses are standard errors:

$$\widehat{\log Kp_j} = -0.78(0.19) + 0.35(0.12)E_j - 0.35(0.11)S_j - 0.02(0.11)A_j + 0.48(0.19)B_j - 1.47(0.23)V_j \quad (3)$$

The expanded model in Equation (2) was fit to produce Equation (4):

$$\begin{aligned}\widehat{\log Kp_{i,j}} = {} & F_{1j}(-1.14(0.22) + 0.92(0.14)E_j - 0.78(0.12)S_j + 0.03(0.12)A_j - 0.77(0.22)B_j - 0.39(0.28)V_j). \\ & + F_{2j}(-0.55(0.24) + 0.71(0.16)E_j - 0.62(0.14)S_j - 0.20(0.15)A_j - 0.28(0.25)B_j - 1.28(0.31)V_j) \\ & + F_{3j}(-0.92(0.27) - 0.01(0.17)E_j - 0.24(0.15)S_j - 0.30(0.16)A_j + 1.40(0.26)B_j - 1.67(0.33)V_j) \\ & + F_{4j}(-0.84(0.26) - 0.46(0.14)E_j - 0.20(0.14)S_j + 0.10(0.13)A_j + 2.00(0.25)B_j - 2.21(0.29)V_j).\end{aligned} \quad (4)$$

Each row in Equation (4) represents the fitted partial slopes for one MWF formulation. For example, row one is for water, row two is for SYN, etc. The use of the fitted model (4) is not as complex as it might seem. When given a solute with its five descriptors and the MWF formulation in which it was dosed, one only needs to find the corresponding regression coefficients for the given formulation and plug the descriptors into this particular Abraham's LFER model. The estimated regression coefficients in Equation (4) are the same as those of the corresponding MWF formulation in Table 3, but with different standard errors. The difference in standard errors is understandable since Table 3 was the fitted results of four separate subsets, while Equation (4) was fitted based on the entire dataset.

The regression statistics of the models in Equations (3) and (4) can be found in Table 4. The improvement of the goodness-of-fit is obvious, with $r^2$ increasing from 0.21 to 0.70.

Comparison of plots of the observed versus the predicted log $Kp$ for the simple model in Figure 3 and the expanded model in Figure 4 demonstrates the superior goodness-of-fit by the expanded model. In Figure 3, the data points are grouped by solutes, meaning that each column of data points in Figure 3 represents a solute. Because the descriptors belonging to a certain solute remain the same across MWF formulations, the resulting predicted values are

Table 4. Regression statistics of both simple model Equation (3) and expanded model Equation (4).

| *Statistics* | *Simple model* | *Expanded model* |
|---|---|---|
| $n$ | 387 | 387 |
| RMSE | 0.51 | 0.32 |
| $r^2$ | 0.21 | 0.70 |
| adj– $r^2$ | 0.20 | 0.69 |

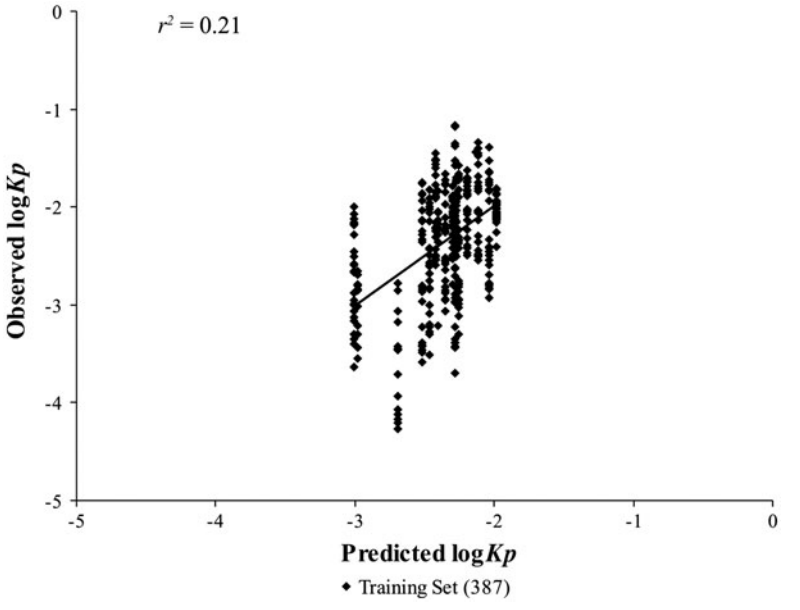


Figure 3. Observed vs. predicted log *Kp* of the training set for the simple model of Equation (3).

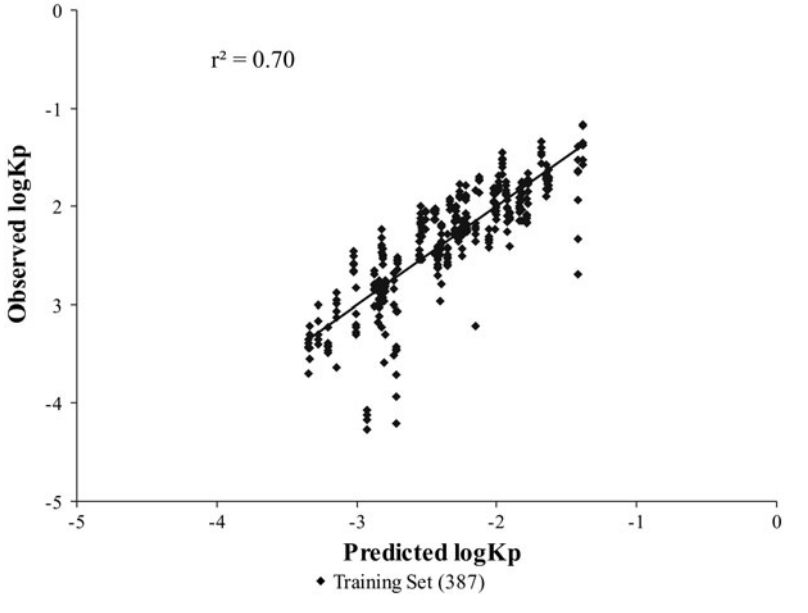


Figure 4. Observed vs. predicted log *Kp* of the training set for the expanded model of Equation (4).

all the same for that solute irrespective of the widely varying observed values for different MWF formulations. In contrast, the expanded model includes the MWF formulation information in the regression model, thus allowing predicted values to change according to MWF formulation. The result demonstrated in Figure 4 is that values are predicted with greater accuracy.

### 3.2 *Internal validation: leave-one-out and leave-one-solute-out*

As suggested by the OECD guidelines [16], one of the principles for a QSAR model to be valid is to have appropriate measures of predictivity. When no additional data are available, a common practice for evaluating the predictive power is by internal cross-validation. There are several approaches of internal validation, including leave-one-out, leave-many-out, bootstrapping, etc. [17]. We use leave-one-out and calculate the corresponding statistic $Q^2_{LOO}$. In the leave-one-out method, each observation is left out once and a QSAR model is fit with the remainder of the data until all the observations have been predicted. The predicted response of the fitted model is $\hat{y}_{i,-i}$. The final measured $Q^2_{LOO}$ is calculated as

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i,-i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

where $y_i$ is the $i$th observed response value, $\hat{y}_{i,-i}$ is the leave-one-out predicted value of the $i$th observation with the model fitted without $i$, and $\bar{y}$ is the average of all the observed responses. Large values of $Q^2_{LOO}$ are preferred but $Q^2_{LOO}$ is not limited to the range of [0,1] as the more familiar fit assessment measure $r^2$.

Though the leave-one-out method is used widely as an internal validation approach [11,17], one problem with using this usual $Q^2_{LOO}$ statistic directly is that a solute may have an experimental replicate in the training set and another in the validation set (the validation set only has one single observation in this leave-one-out case). Thus those two sets are not independent, which may impact evaluation of the predictive power, resulting in misleading indications of good predictive power. To avoid this issue, cross-validation by leaving one solute out each time was conducted and the QSAR model was refitted with the remaining solutes; this approach gives a better representation of how well a QSAR model will perform when applied to a solute that has not previously been tested. Accordingly, the $Q^2_{LOO}$ is adjusted as

$$Q^2_{LOO-adj} = 1 - \frac{\sum_{i=1}^{23}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij,-i})^2}{\sum_{i=1}^{23}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2} \tag{6}$$

where $y_{ij}$ is the $j$th observation of the $i$th solute, $\bar{y}$ is the average of all the observed responses, and $\hat{y}_{ij,-i}$ is the predicted value of $y_{ij}$ with the model fitted leaving out all the observations belonging to the $i$th solute, $n_i$ is the replicate number of $i$th solute.

Table 5 shows the $Q^2_{LOO}$ and $Q^2_{LOO\text{-}adj}$ statistics for the simple model in Equation (3) and the expanded model in Equation (4), respectively. As expected, the expanded model has larger $Q^2_{LOO}$ and $Q^2_{LOO\text{-}adj}$ than those of the simple model, which indicates better predictive power in both of the two types of validations.

Table 5. $Q^2_{LOO}$, $Q^2_{LOO\text{-}adj}$ and $Q^2_{EXT}$ statistics of the simple model Equation (3) and the expanded model Equation (4).

| *Statistics* | *Simple Model* | *Expanded Model* |
|---|---|---|
| $Q^2_{LOO}$ | 0.18 | 0.67 |
| $Q^2_{LOO\text{-}adj}$ | −0.03 | 0.35 |
| $Q^2_{EXT}$ | 0.20 | 0.33 |

### 3.3 *Improved external validation statistics*

Recently, there have been concerns about using internal validation to measure the predictive power of models [18,19]. For example, Golbraikh and Tropsha [19] discussed in their study that the $Q^2$ from internal validation may overestimate the predictive power of a QSAR model, thus it can only be used as a necessary condition of a good predictive model. For this reason, more researchers turn to external validation. In external validation, a training set is used to fit a QSAR model and then the results of the model are used to predict the response of interest for observations from an external validation set. $Q^2_{EXT}$ is used to measure the predictive power of the model [16]. Suppose there are $m$ observations in the validation set, the $Q^2_{EXT}$ is defined as

$$Q^2_{EXT} = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{7}$$

where $\hat{y}_i$ is the value predicted by the model fitted from the training set and $\bar{y}$ is the mean of the training set.

The data from Vijay et al. [5] were used as the validation set. Each of the 35 validation set solutes was dosed in four similar MWF formulations: water, 10% synthetic oil (SYN), 10% semi-synthetic oil (SSYN) and 10% soluble oil (SO). For each solute in each MWF formulation, there were two to four replicates. In total, there were 486 observations in the validation set, among which 140 observations were from water, 105 from synthetic oil, 138 from semi-synthetic oil and 103 from soluble oil. In this paper, the focus is on incorporating the MWF formulation type into the LFER model, thus the concentration difference in the training set and validation set was neglected, and the two sets were considered to be using the same types of MWF formulations. In a future paper, the MWF concentration will be used as another factor in the expanded model and the impact of MWF concentrations to the permeation capability will be investigated.

The summary statistics of the response variable and predictors in the validation set are listed in Table 6. The complete data are available as supplementary information on the article's web page.

As we did for the training set, boxplots of the validation set grouped by MWF formulation are depicted in Figure 5. From Figure 5, a similar trend is seen in the log *Kp* means across the MWF formulations. That is, solutes in water show the highest permeability, followed by SYN and SSYN, and soluble oil has the lowest permeability. These boxplots of the validation set again support our argument that log *Kp* values vary in different MWF formulations and thus a single Abraham's LFER model may not be sufficient.

Table 6. Summary statistics of variables in validation set.

| *Variable* | *Minimum* | *Lower Quartile* | *Mean* | *Median* | *Upper Quartile* | *Maximum* | *Std Dev* | *N* |
|---|---|---|---|---|---|---|---|---|
| log *Kp* | −4.088 | −2.436 | −2.142 | −2.070 | −1.738 | −1.115 | 0.530 | 486 |
| E | 0.600 | 0.740 | 0.912 | 0.820 | 1.030 | 1.550 | 0.242 | 486 |
| S | 0.500 | 0.840 | 0.941 | 0.900 | 1.090 | 1.680 | 0.238 | 486 |
| A | 0 | 0 | 0.228 | 0 | 0.56 | 0.7 | 0.279 | 486 |
| B | 0.050 | 0.160 | 0.319 | 0.280 | 0.440 | 0.830 | 0.185 | 486 |
| V | 0.775 | 0.939 | 1.065 | 1.038 | 1.139 | 1.786 | 0.196 | 486 |

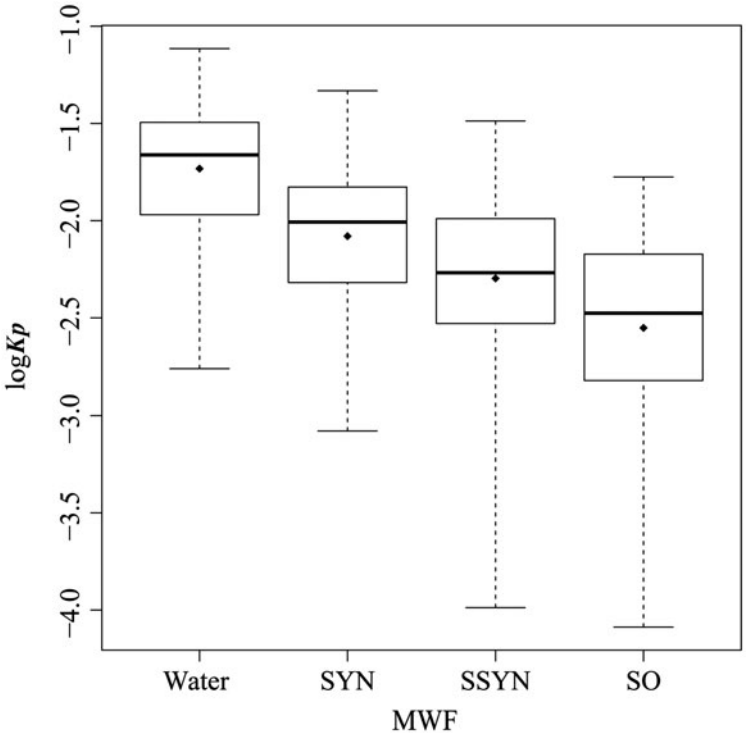


Figure 5. Boxplot of log *Kp* across MWF formulations of the validation set. Numbers of observations in each MWF are: 140 (water), 105 (SYN), 138 (SSYN) and 103 (SO). The mean permeability follows the order: water > SYN > SSYN > SO. The five lines from top to bottom of each boxplot represent: maximum, 3rd quartile, median, 1st quartile and minimum. The diamond represents the sample mean.



With the fitted models in Equations (3) and (4), the log *Kp* values of the validation set were predicted and then the $Q^2_{EXT}$ statistics were calculated for each model and included in Table 5. The $Q^2_{EXT}$ for the simple model is 0.20, and increases to 0.33 for the expanded model, which indicates that the expanded model has improved predictive power.

To visually demonstrate the improved predictive power by the expanded model, the observed vs. predicted log *Kp* values are plotted for both the training set and the validation set together in Figures 6 and 7.

### 3.4 *Hypothesis testing and condensed model*

One advantage of fitting an expanded model instead of four separate Abraham's LFER models is that hypothesis testing can be easily conducted. From Figure 2, one can see that the regression coefficients vary across the four types of MWF formulations. With the expanded model, we can formally test for statistical significance of the MWF formulation effect. The null hypothesis for testing the MWF formulation effect will be based on Equation (2) as

$H_{01}$: $\beta_{l1} = \beta_{l2} = \beta_{l3} = \beta_{l4}$ for $l$ = 1,..,5

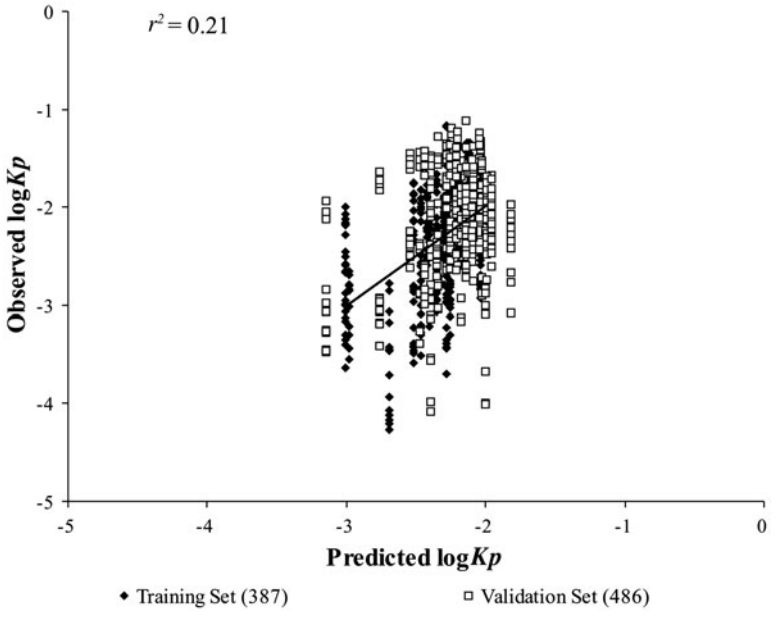


Figure 6. Observed vs. predicted log *Kp* of the training set and the validation set for the simple model of Equation (3).

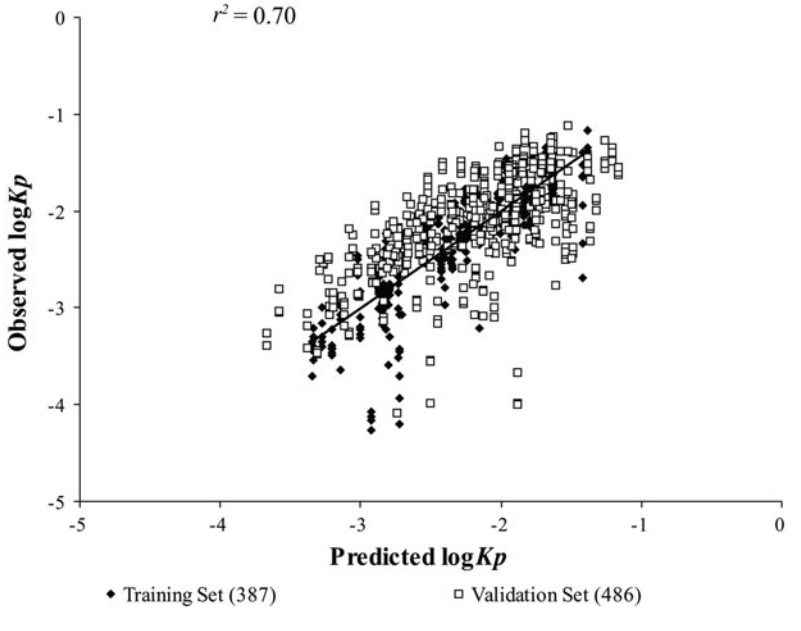


Figure 7. Observed vs. predicted log *Kp* of the training set and the validation set for the expanded model of Equation (4).

i.e. that the partial slopes for each of $E$, $S$, $A$, $B$, and $V$ do not differ with changing MWF formulation. The test result of the above null hypothesis $H_{01}$ is shown in Table 7. The $p$-value is less than 0.0001, which indicates the MWF formulation effect is very significant.

Sometimes, it may be of interest to group the MWF formulation effect, for example to test whether the formulations water and synthetic oil play a similar role in affecting the permeability. This is testing the null hypothesis $H_{02}$: $\beta_{l1} = \beta_{l2}$ for $l$ = 1,..,5. Table 8 shows all the paired testing results. To adjust for the multiple testing and control the family-wise error rate at $\alpha = 0.05$, we compare the $p$-values with the Bonferroni-adjusted significance level $\alpha^* = 0.05/6 = 0.0083$ and find that there is no significant difference between water and synthetic oil, or between semi-synthetic oil and soluble oil. In other words, the four types of MWF formulations can be combined into two groups; one group includes water and synthetic oil while the other includes semi-synthetic oil and soluble oil. Within each group, the MWF formulation effect is much more similar than between the two groups. This grouping is not surprising. The base oil used for formulating soluble and semi-synthetic MWF formulations is generally heavy naphthenic mineral oil, while synthetic MWF formulations contain no petroleum oil and are made with organic and inorganic salts dissolved in water [20].

Since some MWF formulation effects are more similar than others, a natural question inspired by Table 8 might be can we simplify the expanded model in Equation (2) (that is, use fewer regression coefficients by collapsing the model among the MWF formulations) without impairing its predictive power significantly?

The answer is yes. We can start with testing a model with only two sets of regression coefficients, one for the water and synthetic oil group, the other for the semi-synthetic oil and soluble oil group. The testing hypothesis will be $H_{03}$: $\beta_{l1} = \beta_{l2}$, $\beta_{l3} = \beta_{l4}$, for $l$ = 1,…,5. By doing this, we expect to get a model of only half the coefficients of the expanded model. The hypothesis test for $H_{03}$ (which means simultaneously collapsing SSYN/SO and SYN/water) results in a $p$-value of 0.0082, which is significant. This $p$-value means collapsing both groups will be too aggressive. Thus we should consider collapsing SSYN/SO and water/SYN separately. We start by collapsing water and SYN first, as the $p$-value of 0.0640 for Water = SYN in Table 8 is less significant. We then continue the process by removing the insignificant coefficients to get the condensed model in Equation (8):

Table 7. Test of MWF formulation effect.

| *Source* | *DF* | *Mean Square* | *F Value* | *Pr > F* |
|---|---|---|---|---|
| Numerator | 15 | 1.5232 | 14.7 | <.0001 |
| Denominator | 363 | 0.1036 | | |

Table 8. Paired testing of partial slopes for the four MWF formulations.

| *Null hypothesis* | *P-value* |
|---|---|
| Water = SYN | 0.0640 |
| Water = SSYN | <.0001 |
| Water = SO | <.0001 |
| SYN = SSYN | <.0001 |
| SYN = SO | <.0001 |
| SSYN = SO | 0.0187 |

Table 9. Lack-of-fit test result for the condensed model in Equation (8).

| *Source* | *DF* | *Error Sum of Squares* | *F Value* | *Pr > F* |
|---|---|---|---|---|
| Condensed model | 372 | 39.0584 | 1.5425 | 0.1312 |
| Expanded model | 363 | 37.6197 | | |

Table 10. Regression statistics for the condensed model in Equation (8).

| *Statistics* | *Condensed model* |
|---|---|
| n | 387 |
| RMSE | 0.32 |
| $r^2$ | 0.69 |
| *adj*-$r^2$ | 0.68 |
| $Q^2_{LOO}$ | 0.67 |
| $Q^2_{LOO\text{-}adj}$ | 0.41 |
| $Q^2_{EXT}$ | 0.31 |

$$\begin{aligned}\widehat{\log Kp}_{i,j} = {} & F_{1j}* - 0.77(0.16) + F_{2j}* - 1.08(0.16) + F_{12j}*(0.83(0.10)E_j \\ & - 0.71(0.09)S_j - 0.53(0.17)B_j - 0.79(0.21)V_j) + F_{3j}*(-0.92(0.27) \\ & - 0.25(0.11)S_j - 0.30(0.16)A_j + 1.41(0.25)B_j - 1.67(0.31)V_j) \\ & + F_{4j}*(-0.83(0.25) - 0.32(0.10)E_j + 2.07(0.25)B_j \\ & - 2.16(0.28)V_j) \end{aligned} \quad (8)$$

where $F_{12j} = 1$ if the $j$th observation is from MWF formulation 1 (water) or 2 (SYN), $F_{12j} = 0$ otherwise.

To test the sufficiency of the condensed model in Equation (8), a lack-of-fit test is performed. The sum of squared errors and degrees of freedom of the reduced model (i.e. condensed model) and the full model (i.e. expanded model in Equation 4) are listed in Table 9. The $F$-statistic is 1.5425 with a $p$-value of 0.1312. Since the $p$-value is larger than 0.05, we regard the condensed model in Equation (8) as sufficient.

It is also possible to collapse the coefficients of SSYN and SO first, but the lack-of-fit test produces a significant $p$-value of 0.0187, thus we take the model in Equation (8) as the final condensed model.

The regression statistics of the condensed model are listed in Table 10. Compared with the statistics of the expanded model in Tables 4, 5 and it is suggested that the condensed model is a good approximation of the expanded model with respect to the goodness-of-fit statistics ($r^2$, adj-$r^2$), and it has improved the predictive internal power with higher $Q^2_{LOO\text{-}adj}$ values. Moreover, the condensed model is simple and easy to apply, and it explains the relationship between MWF formulations water and synthetic oil.

### 3.5 *Model interpretation*

We now explain how to interpret the condensed model in Equation (8). For convenience, the model is repeated in Equation (9) with some convenient formatting:

$$\begin{aligned}\widehat{\log Kp}_{i,j} &= F_{1j}* - 0.77(0.16) + F_{2j}* - 1.08(0.16) \\ &+ F_{12j}*(0.83(0.10)E_j - 0.71(0.09)S_j - 0.53(0.17)B_j \\ &- 0.79(0.21)V_j) + F_{3j}*(-0.92(0.27) - 0.25(0.11)S_j \\ &- 0.30(0.16)A_j + 1.41(0.25)B_j - 1.67(0.31)V_j) \\ &+ F_{4j}*(-0.83(0.25) - 0.32(0.10)E_j + 2.07(0.25)B_j - 2.16(0.28)V_j) \qquad (9)\end{aligned}$$

There are five rows in the above Equation (9). The first and second rows represent the intercepts for water and synthetic oil, respectively. The third row is the common regression function for water and synthetic oil based on the hypothesis testing results. Row four and row five are regression functions for semi-synthetic oil and soluble oil, respectively. The blank spaces in rows three to five indicate that the corresponding descriptors are not important for predicting permeability in that particular MWF formulation (except for the intercept position in row three). The coefficient is the partial slope for the corresponding descriptor in that particular MWF formulation.

According to Equation (9), the effect of the MWF formulation (i.e. solvent) is not as simple as an additive or linear effect. Given the increasing amounts of oil content of the varying formulations, from water (no oils), to SYN (no mineral oil but other oils), to SSYN (moderate mineral oil), to SO (heavy mineral oil), one may have easily predicted the observed ordering of permeability across the various MWF formulations. The model in Equation (9), however, implies much more than this predictable ordering. As we change the MWF formulation, permeability coefficients do not change in a simple linear additive fashion. In other words, permeability coefficients for water are not simply a constant increase above permeability coefficients for soluble oil. Instead, the structure of dependence between permeability and solute characteristics changes with the changing MWF formulations, and this is reflected in different coefficients for each MWF formulation group. For example, the coefficient for $B$ (hydrogen bonding basicity) is 2.07 in soluble oil, 1.41 in semi-synthetic oil and -0.53 in both water and synthetic oil. Recalling the interpretation of coefficients in linear regression, a coefficient of 2.07 for $B$ in soluble oil means that a one unit increase in overall hydrogen bond basicity of a solute will increase log $Kp$ by 2.07 units (thus increasing permeability by a factor of $10^{2.07} = 117.5$) provided all other solute descriptors are held fixed. This represents a very large impact of a solute's hydrogen bond basicity. In contrast, when either water or synthetic oil are used as the solvent, a one unit increase in overall hydrogen bond basicity will decrease log $Kp$ by 0.53 unit (thus decreasing permeability by a factor of $10^{-0.53} = 0.295$). Both the directional change (increase versus decrease) and the factor of change (above 100 times versus less than 1/3) are relevant for understanding the mechanisms of permeation in the various MWF formulations.

Figure 8 further demonstrates how descriptor $B$, hydrogen bond basicity, impacts the permeability coefficient differently according to MWFs; it is an added-variable plot (Weisberg [21], pp. 49–50) of descriptor $B$. Added-variable plots show dependence of a response on an individual predictor after accounting for other predictors. If only one predictor were used in Equation (9), a simple scatterplot of log $Kp$ against that predictor would have been sufficient to show the dependence relationship. But because our model contained five predictors, we must first account for the other predictors before graphically illustrating the effect of predictor $B$. The $x$-axis in the added-variable plot is the residual of regressing $B$ against all predictors except $B$ (we refer to this as $B^*$), and the $y$-axis is the residual of regressing log $Kp$ against all predictors except $B$ (we refer to this as log $Kp^*$). In Figure 8, different plotting symbols

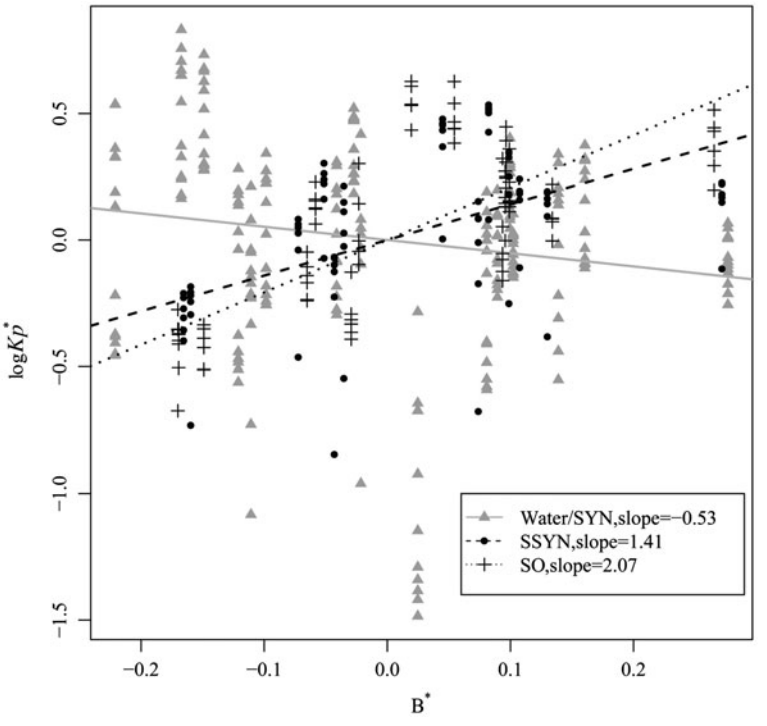


Figure 8. Added-variable plot of descriptor *B* for three MWF formulation groups: water/synthetic oil (gray triangle, slope of -0.53), semi-synthetic oil (black dot, slope of 1.41) and soluble oil (black plus, slope of 2.07). The slope of each regression line is equal to the regression coefficient of *B* for corresponding MWF in Equation (9). The *x*-axis is the residual of regressing *B* against all predictors except *B* (we refer to this as $B^*$), and the *y*-axis is the residual of regressing log *Kp* against all predictors except *B* (we refer to this as log $Kp^*$).

are used to indicate the three MWF formulation groups: Water/SYN, SSYN, and SO. The slopes of the three regression lines in Figure 8 exactly equal the partial coefficients of *B* in each of the three groups of MWF formulation, indicating that the effect of overall hydrogen bond basicity can be very different in the various MWF formulations, after accounting for the other descriptors.

In addition, Equation (9) tells us that the solute excess molar refraction (*E*) is not important for SSYN, at the same time it shows opposite impacts to solute permeability in water/SYN and SO. Similarly, the solute dipolarity/polarizability (*S*) is not important for SO and the overall hydrogen bond acidity (*A*) is not important for either water or SYN or SO and is only marginally important for SSYN. The McGowan characteristic volume (*V*) has significant impact in all three MWF formulation groups, with larger molecules being less permeable and the extent of reduced permeability is greatest in SO.

Figure 9(a) is the observed vs. predicted log *Kp* of the training set and the validation set for the condensed model of Equation (9). Separating the observations in Figure 9(a) by MWF formulation will show that different MWF formulations have different prediction error, with permeability in soluble oil being the least predictive and permeability in synthetic oil being the most predictive. (The plots for each MWF formulation are not listed in this paper due to

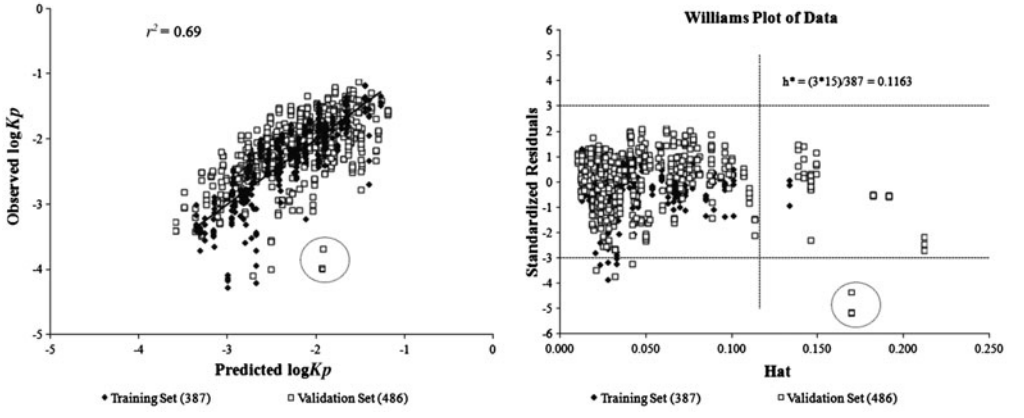


Figure 9. (a) Observed log *Kp* vs. predicted log *Kp* of the training and validation sets for the condensed model of Equation (9). The three most poorly predicted observations have been identified by the circle in the figure, and they all come from the validation set; all three have model predictions about a factor of two away from actual observed values. (b) Williams plot of the training and validation sets. The three circled observations are the same as those circled in Figure 9(a), and are here seen to have the largest standardized residuals. These same three observations also fall outside the applicability domain (AD) of the model, as indicated by their position to the right of the vertical line critical hat value ($h^*$).

space limitations.) The relatively poor prediction in soluble oil is also suggested in Figure 9 (a), where the three most obvious outliers are all from soluble oil. These outliers can be explained by a Williams plot ([6,17]) as shown in Figure 9(b). The Williams plot is used to visualize the applicability domain (AD) of a model, with AD being defined as response and chemical structure space in which the model makes predictions with a given reliability [22]. It is a plot of the standardized residuals versus the leverage values (h). Observations to the right of the vertical line ($h > h^* = 0.1163$ in Figure 9(b), see Xu et al. [6] for more details about the calculation of leverage h and cutoff value $h^*$) are said to be outside of the AD and thus are predicted with less reliability. The three outliers from Figure 9(a) all fall outside the AD, which explains why they are poorly predicted.

## 4. Conclusions

In this paper, we make contributions to two aspects of fitting and assessing QSAR models: the first is that an expanded version of Abraham's LFER model is introduced, which not only incorporates the experimental factors and thus gives a better fit of the data, but also provides an easy way of hypothesis testing of the effects of factors. The second contribution is that the usual leave-one-out cross-validation method is modified to a "leave-one-solute-out" method, which considers the correlation between replicates from the same solute and measures the predictivity of the QSAR model more precisely.

In recent years, Riviere and Brooks [10,23,24] have produced a series of articles describing the use of a mixture factor (MF) to predict permeation from complex chemical mixtures. This MF is a sixth term added to Abraham's model that is based on the weight percentage of all the constituents in each of the chemical mixtures. That is, the five Abraham parameters describe each of the solutes, while the MF describes all the constituents in each of the mixtures. This approach proved useful in predicting permeation from mixtures other than binary mixtures of a single solute delivered from water.

The current use of the expanded model is an extension of this approach of predicting permeation from complex chemical mixtures when the mixtures are too complex to be described by weight percentages. The four MWF formulations utilized in this permeation study each contained water and a mixture of 23 solutes. In addition, three of the formulations contained either synthetic oil, semi-synthetic oil, or soluble oil. Description of the weight percent of each constituent in these oils would be extensive, if at all possible. The current expanded model, which accounts for four different formulations, eliminates the necessity of knowing the exact chemical identity and percentage of every constituent in the mixture.

Our results suggest that when the experimental conditions are complex, for example when there is more than one type of MWF formulation, an expanded model that accounts for the heterogeneity across the MWF formulations will greatly improve model fitting and predictive power. The proposed leave-one-solute-out internal validation approach along with $Q^2_{LOO\text{-}adj}$ statistic provides a more reasonable way of evaluating the predictive power when there are pseudo or real replicates. With the expanded model, hypothesis testing can also be applied, from which one can get insights into the similarity among the different MWF formulations.

In the literature there has been a proposal to incorporate a vehicle indicator variable into a QSAR model to account for the vehicle impact [25]. Though this will increase the model fit and predictive power, it was not adopted in this paper as we believe a QSAR model with an indicator variable will not be sufficiently flexible to capture the many potential inherent differences of solubility of compounds in varying MWF formulations as discussed in Section 3.5. Benefits of this model flexibility include being able to reasonably predict the observed permeabilities of Vijay et al. [5] even though their MWF formulations differed from those of the experiments reported in this paper (10% oils in Vijay et al. [5] compared with 5% oils here).

The advantage of hypothesis testing will be especially obvious in more complicated experiments, for example experiments with more than four MWF formulations, or with different solute concentrations, etc. The power of the hypothesis testing will be demonstrated in a future paper with a more complicated experimental design. For example, the incorporation of solvent concentration to expand Equation (9) may have minimized the degree of outlying behaviour seen in Figures 9(a) and 9(b) from predicting permeabilities from Vijay et al. [5] without acknowledging the differences in experimental conditions of the experiment reported here compared with the one reported by Vijay et al. [5].

Other QSAR approaches, not based on Abraham's LFER model, are certainly feasible (see for example Livingstone [7]) and may also be extended to account for experimental conditions. Alternatively, non-QSAR approaches are also possible for modelling dermal permeability. For example, the spreadsheet-based model of Dancik et al. [26] requires input of physico-chemical and dose parameters of the permeant as well as the vehicle. Unfortunately, the vehicle is limited to mixtures containing only five selections: water, olive oil, ethanol, acetone, and "other". The model showed negative correlations to our training set semi-synthetic and soluble oil data. This was due to a lack of access to the solubility of each permeant in 95% water + 5% MWF. The solubility of the permeant in the vehicle is the limiting factor. That is, when water solubility was input into the model as an estimate of these 95% water formulations, the resultant effective *K*p values were the same as the effective *K*p values for the 100% water formulations. Experimental determination of the solubility of these 23 compounds in the 95% water + 5% MWF would be necessary to run this on-line model effectively with our experimental doses.

Individuals do not universally agree upon the techniques for model validation. In this paper, we use separate training and validation sets, while Hawkins et al. [27] alternatively suggest internal cross-validation is more efficient than the approach of creating separate

training and validation sets. Our choice of using separate training and validation sets emanate from the fact that it is considered the standard practice in QSAR studies concerning dermal permeability (see OECD guidelines on QSAR model validation [16]). Irrespective of whether an external validation set is used for model assessment, any internal validation steps should be conducted using a leave-one-solute-out approach, as we have presented here.

In conclusion, this study demonstrated how a QSAR model with 5% MWF information can be used to predict changes in skin permeability according to the MWF formulation. For example, a 95% lower confidence bound of 0.5225 for the difference between log *Kp* from water-based or synthetic oil-based MWF formulation and log *Kp* from soluble oil fluids suggests that workers should be more concerned about dermal absorption of formulation components in water-based and synthetic MWFs than soluble oil fluids. In fact, the lower confidence bound implies that permeability in water or synthetic oil is likely to be greater than $10^{0.5225} = 3.33$ times the permeability in soluble oil, and this information may be highly relevant for controlling exposure and consequential health effects in workers.



## Supplementary material

Supplementary material can be found under the 'Supplemental' tab on the article's web page: http://dx.doi.org/10.1080/1062936X.2013.792875.

## Acknowledgements

The authors would like to thank James L. Yeatts for his analytical work, Milacron Inc. for kindly providing generic commercial MWFs, and the anonymous referees whose comments led to an improved paper. This work was supported by NIOSH Grant #OH-003669.

## References

[1] R.O. Potts and R.H. Guy, *Predicting skin permeability*, Pharm. Res. 9 (1992), pp. 663–669.
[2] M.H. Abraham, *Scales of solute hydrogen-bonding: Their construction and application to physicochemical and biochemical processes*, Chem. Soc. Rev. 22 (1993), pp. 73–83.
[3] ACD/ADME BOXES, version 4.95, Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2012. Available at www.acdlabs.com.
[4] M.H. Abraham and F. Martins, *Human skin permeation and partition: General linear free-energy relationship analyses*, J. Pharm. Sci. 93 (2004), pp. 1508–1523.
[5] V. Vijay, E.M. White, M.D. Kaminski, Jr., J.E. Riviere, and R.E. Baynes, *Dermal permeation of biocides and aromatic chemicals in three generic formulations of metalworking fluids*, J. Toxic. Environ. Health, Part A: Current Issues 72 (2009), pp. 832–841.
[6] G. Xu, J.M. Hughes-Oliver, J.D. Brooks, J.L. Yeatts, and R.E. Baynes, *Selection of appropriate training and validation set chemicals for modelling dermal permeability by U-optimal design*, SAR QSAR Environ. Res. 24 (2013), pp. 135–156.
[7] D.J. Livingstone, *A Practical Guide to Scientific Data Analysis*. Wiley, 2010.
[8] P. Gramatica, *Evaluation of Different Statistical Approaches for the Validation of Quantitative Structure-Activity Relationships*, ECVAM, Ispra, 2004.
[9] R.E. Baynes, J.D. Brooks, B.M. Barlow, and J.E. Riviere, *Physiochemical determinants of linear alkylbenzene sulfonate (LAS) disposition in skin exposed to aqueous cutting fluid mixtures*, Toxicol. Indust. Health 18 (2002), pp. 237–248.
[10] J.E. Riviere and J.D. Brooks, *Predicting skin permeability from complex chemical mixtures*, Toxicol. Appl. Pharmacol. 208 (2005), pp. 99–110.
[11] R.E. Baynes and V. Vijay, *Dermal absorption and cutaneous toxicity of metalworking fluids*, Toxicology of the Skin, 1st edition (2010), pp. 361–374.

[12] G. Foltz, *Definitions of metalworking fluids*, in *Waste Minimization and Wastewater Treatment of Metalworking Fluids*, R.M. Dick, ed., Independent Lubrication Manufacturers Association, Alexandria, VA, (1990), pp. 2–3.
[13] Q. Zhang, J.E. Grice, P. Li, O.G. Jepps, G.J. Wang, and M.S. Roberts, *Skin solubility determines maximum transepidermal flux for similar size molecules*, Pharm. Res. 26 (2009), pp. 1974–1985.
[14] *SAS 9.3*, SAS Institute Inc., Cary, NC, (2011), software available at http://www.sas.com
[15] R Development Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/
[16] OECD, *Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models*, Series on Testing and Assessment, No. 69. OECD, Paris, (2007). Available at http://www.oecd.org/dataoecd/55/35/38130292.pdf
[17] P. Gramatica, *Principles of QSAR models validation: internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701.
[18] K. Baumann and N. Stiefl, *Validation tools for variable subset regression*, J. Comput-Aided Mol. Design 18 (2004), pp. 549–562.
[19] A. Golbraikh and A. Tropsha, *Beware of $q^2$!*, J. Mol. Graph. Mod. 20 (2002), pp. 269–276.
[20] J.E. Anderson, B.R. Kim, S.A. Mueller, and T.V. Lofton, *Composition and analysis of mineral oils and other organic compounds in metalworking and hydraulic fluids*, Crit. Rev. Environ. Sci. Tech. 33 (2003), pp. 73–109.
[21] S. Weisberg, *Applied Linear Regression*. Vol. 528. Wiley, 2005.
[22] T.I. Netzeva, G. Saliner, and A.P. Worth, *Comparison of the applicability domain of a QSAR for estrogenicity with a large chemical inventory*, Env. Toxicol. Chem. 25 (2005), pp. 1223–1230.
[23] J.E. Riviere and J.D. Brooks, *Prediction of dermal absorption from complex chemical mixtures: Incorporation of vehicle effects and interactions into a QSPR framework*, SAR QSAR Environ. Res. 18 (2007), pp. 31–44.
[24] J.E. Riviere and J.D. Brooks, *Predicting skin permeability from complex chemical mixtures: Dependency of quantitative structure permeation relationships on biology of skin model used*, Toxicol. Sci. 119 (2011), pp. 224–232.
[25] J.J. Hostynek and P.S. Magee, *Modeling in vivo human skin absorption*, Quant. Struct.-Act. Relat. 16 (1997), pp. 473–479.
[26] Y. Dancik, M.A. Miller, J. Jaworska, and G.B. Kasting, *Design and performance of a spreadsheet-based model for estimating bioavailability of chemicals from dermal exposure*, Adv. Drug. Deliver. Rev. 65 (2012), pp. 221–236.
[27] D.M. Hawkins, B.C. Subhash, and D. Mills, *Assessing model fit by cross-validation*, J. Chem. Info. Comput. Sci. 43 (2003), pp. 579–586.