SAR and QSAR in Environmental Research

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

Selection of appropriate training and validation set chemicals for modelling dermal permeability by U-optimal design

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

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

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

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

Published online: 16 Nov 2012.

To cite this article: G. Xu , J.M. Hughes-Oliver , J.D. Brooks , J.L. Yeatts & R.E. Baynes (2013): Selection of appropriate training and validation set chemicals for modelling dermal permeability by U-optimal design, SAR and QSAR in Environmental Research, 24:2, 135-156

To link to this article: http://dx.doi.org/10.1080/1062936X.2012.742458

SAR and QSAR in Environmental Research, 2013
Vol. 24, No. 2, 135–156, http://dx.doi.org/10.1080/1062936X.2012.742458

Selection of appropriate training and validation set chemicals for modelling dermal permeability by U-optimal design

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

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

(*Received 10 September 2012; in final form 18 October 2012*)

Quantitative structure-activity relationship (QSAR) models are being used increasingly in skin permeation studies. The main idea of QSAR modelling is to quantify the relationship between biological activities and chemical properties, and thus to predict the activity of chemical solutes. As a key step, the selection of a representative and structurally diverse training set is critical to the prediction power of a QSAR model. Early QSAR models selected training sets in a subjective way and solutes in the training set were relatively homogenous. More recently, statistical methods such as D-optimal design or space-filling design have been applied but such methods are not always ideal. This paper describes a comprehensive procedure to select training sets from a large candidate set of 4534 solutes. A newly proposed 'Baynes' rule', which is a modification of Lipinski's 'rule of five', was used to screen out solutes that were not qualified for the study. U-optimality was used as the selection criterion. A principal component analysis showed that the selected training set was representative of the chemical space. Gas chromatograph amenability was verified. A model built using the training set was shown to have greater predictive power than a model built using a previous dataset [1].

Keywords: QSAR model; training set selection; Baynes' Rule; U-optimal design; principal component analysis (PCA); applicability domain (AD)

1. Introduction

During metalworking, cutting fluids are used widely to prevent the cutting tools from overheating and losing their accuracy. To inhibit bacterial or fungal growth in these cutting fluids, it is a common practice to add certain chemical solutes known as biocides. Skin irritation is sometimes the result of contact with these cutting fluids containing biocides. Thus, it is of interest to study the permeation capability of the added chemical solutes through skin and to find safer chemical solutes that can be used as biocides in cutting fluids.

Quantitative structure-activity relationship (QSAR) modelling is a widely used approach for *in vitro* study of permeation of chemicals or solutes through skin [2,3].

*Corresponding author. Email: rebaynes@ncsu.edu

The QSAR approach assumes that biological activity (such as skin permeation) can be related quantitatively to chemical properties such as physicochemical properties [3] or molecular structure [4]. These chemical properties can be identified by molecular descriptors [4] through the use of computer software such as ADME BOXES [5]. Once determined, the QSAR model can be applied to any solute whose molecular structure is known in order to obtain a predicted level of biological activity such as skin permeability. Thus the bioactivity of a large number of previously untested solutes can be easily predicted based on their computer-generated molecular descriptors. This can result in great saving of resources and effort, provided the QSAR model has strong predictive power.

In order for a QSAR model to make an effective prediction, one key step is to select an appropriate training set [6], where 'appropriate' means that those solutes selected for the training set should be structurally diverse and representative of the entire solute space [6,7]. Early QSAR models such as the Potts and Guy model [3] and the Abraham model [4] were designed using sets of solutes without consideration of representativeness. For example, the Potts and Guy model was developed based on Flynn's data [8] having 93 solutes. This training set was a random collection of solutes from a number of different laboratories, with no selection criteria design employed. Since QSAR models are only valid on the trained and validated domains [9], when applied to large and diverse datasets, models built on unrepresentative training sets may not produce satisfying predictions. Indeed, QSAR models are fundamentally recognized as predictive models that can benefit from statistical techniques aimed at improving predictive power.

The literature has a number of applications of statistical methods proposed to select diverse training sets, including factorial designs, D-optimal design [7] and uniform coverage design based on space-filling analysis [10]. While successful in their respective applications, these methods are not ideal for all circumstances. For example, D-optimal designs can be very sensitive to model mis-specification. Although we illustrate our procedure by using a very popular linear model, there is strong evidence that this linear model is not adequate to explain experimental variability and hence we would be uncomfortable using it to select a D-optimal design. On the other hand, selection criteria such as those used by Vijay et al. [10] rely on extreme computational intensity or the use of non-commercial software tools. Thus, a more general and easily implemented method of selecting a training set is needed.

In this paper, a new and comprehensive procedure is proposed for selecting training and validation sets. In this procedure, a two-phase selection process was applied to a large candidate set. First, study unfriendly solutes were screened out with a Baynes' 'rule of four' adjusted from Lipinski's 'rule of five' [11] and then the U-optimal criterion was used, which is one of the distance-based design [12] approaches to select training and validation sets. The principal component analysis (PCA) technique was used to show the diversity and representativeness of the selected sets.

To demonstrate the effectiveness of this procedure, a linear free-energy relationship (LFER) [13] model (i.e. a QSAR model) was fitted with the experimental data on the training set and this model was used to predict the permeation capability of solutes in the validation set. The goodness-of-fit statistics, internal and external validation results and the applicability domain (AD) [14] of the training set were examined. The results were further compared to those of a LFER model fitted from a non-U-optimal set.

2. Materials and methods

2.1 *General overview of the methods*

Among the thousands of molecular descriptors that have been defined for each chemical, it has been found that the five solvatochromic descriptors (E, S, A, B, V) are most relevant to the solvation process during permeation [4,13]. 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 of the chemicals, V can be calculated directly, E can be obtained experimentally or calculated, A, B, and S are experimentally derived. The computer software ADME BOXES [5] was used to obtain these five descriptors. Our target was to fit the general LFER model [13]:

$$\log Kp = \beta_0 + \beta_1 E + \beta_2 S + \beta_3 A + \beta_4 B + \beta_5 V \tag{1}$$

where $\log Kp$ is the permeation coefficient, β_0 is the intercept and β_1 to β_5 are the regression coefficients that describe the relationships of descriptors E, S, A, B and V to $\log Kp$.

As mentioned in the introduction, a key step in building a QSAR model (thus a LFER model) is to find a diverse training set. Solvatochromic descriptors E, S, A, B and V and the permeation coefficients $\log Kp$ were obtained for all solutes in this set, and the resulting data was used to 'fit' the regression model in Equation (1). In other words, data from the training set were used to estimate the unknown intercept and regression coefficients, resulting in the fitted regression equation:

$$\widehat{\log Kp} = \hat{\beta}_0 + \hat{\beta}_1 E + \hat{\beta}_2 S + \hat{\beta}_3 A + \hat{\beta}_4 B + \hat{\beta}_5 V \tag{2}$$

Equation (2) will be applied to any solute whose solvatochromic descriptors are known in order to obtain an estimated permeation coefficient for that solute. But the quality of Equation (2) must be validated by an independent set of solutes (separate from the training set used to fit the equation) to determine whether the estimated permeation coefficients are close to the actual values. A validation set of solutes was selected and values obtained for all solvatochromic descriptors and permeation coefficients. The solvatochromic descriptors were used in Equation (2) to obtain estimated permeation coefficients which were then compared to the actual observed permeation coefficients. If this comparison yielded strong agreement, then the model had been validated and could now be applied to the larger set from which the training and validation sets were selected. As previously mentioned, it is important that both the training and validation sets are selected to be representative of the larger set.

The process began with a 'candidate set' of N solutes. This large set of solutes represented somewhat of a 'universal' set of interest, but not all of these solutes were appropriate for permeation studies. Previous experience in our laboratory, along with published literature, indicates that certain solutes in the candidate set were not study friendly (meaning that they were not gas chromatograph (GC) amenable, or were extremely toxic, were difficult to obtain from chemical supply houses, etc.) and thus should be removed by a screening process. After this preliminary screening process, a screened set with N_s solutes was derived. A statistical design criterion to select a training set of size n_t was then applied. Afterwards, a validation set of size n_v was selected under the same design rule from the leftover set of $N_s - n_t$ solutes.

2.2 *Original set*

A large candidate set of N = 4534 solutes formed the original set; each of these solutes have E, S, A, B and V descriptor values from the database supplied with ADME BOXES [5]. In order to rule out the study-unfriendly solutes in the screening process, the following additional predicted or experimental physicochemical properties were obtained from the literature or calculated: molecular weight (MW), number of hydrogen bond donators/acceptors, and octanol–water partition coefficient (log *P*) values.

The batch calculation option provided by ADME BOXES was used to obtain the above physicochemical properties. The batch calculation option accepts a text file containing records of SMILES (Simplified Molecular Input Line Entry Specification) strings. The SMILES strings are chemical notations of solutes designed especially for computer storage and use [15]. The outputs of the batch calculation are the desired physicochemical properties. ADME BOXES can handle calculation for more than 1000 solutes at a time, which saved effort because only five batches needed to be processed. The SMILES strings list can be generated either by ChemDraw [16], (ChemDraw is a GUI which accepts solute names and then translates them into SMILES strings) or JChem for Excel [17] (JChem for Excel takes the International Union of Pure and Applied Chemistry (IUPAC) named solutes as input and outputs SMILES strings in a batch mode).

With all nine descriptors (the five solvatochromic descriptors E, S, A, B, V plus the four physicochemical properties MW, number of hydrogen bond donators, number of hydrogen bond acceptors and log *P* values), the screening process was applied. Note that ADME BOXES provides two calculated log *P* values: log *P* (AB) and log *P* (ACD). Both values were used for the remainder of this work.

2.3 *Screening process*

Experiments have shown that the permeation capability of certain solutes is related to certain physicochemical properties such as MW, log *P*, etc. [18]. Solutes with certain property values are more or less likely to have good permeation capability. Thus in practice, empirical rules for solute screening are often used.

Lipinski et al. [11] suggested a 'rule of five' to provide an empirical way to assess potential solubility and intestinal permeation. If a solute violates any of the following criteria, then this solute is very likely to have poor intestinal absorption or permeation:

(1) There are more than five hydrogen bond donors.
(2) There are more than 10 hydrogen bond acceptors.
(3) The MW is greater than 500 grams/mole.
(4) The logP is over five.

Lipinski's rule was proven to be successful in solute identification in intestinal absorption [11]. However, in skin permeation, this rule may not apply. Magnusson et al. [18] suggested a rather strict rule for predicting the solutes' potential capability of transdermal delivery. For example, they proposed that when the MW is greater than 213, a solute will show poor permeation through dermal delivery.

Direct application of Magnusson's screening rule to our candidate set is far too stringent as it eliminated all of the 4534 solutes. The screening rule was therefore adjusted based on our experience to consider only two factors: the MW and log *P* values. This was termed 'Baynes' rule of four'. Solutes with MW equal to or greater than 400 or log *P* less

Table 1. Comparison of the three screening rules.

	Predictors – connect with 'and' for inclusion			
Inclusion rule	*MW*	*log P*	*HB-a**	*HB-d***
1. Lipinski (intestinal)	≤ 500	≤ 5	≤ 10	≤ 5
2. Magnusson (transdermal)	≤ 213	≤ 1.2	< 3	< 0
3. Baynes (transdermal)	< 400	$1 \leq \log P \leq 4$	not used	not used

*Hydrogen bond acceptors.
**Hydrogen bond donors.

Table 2. List of chemical classes and reasons for removal.

Chemical class	*Reason for removal*
'acid'	These solutes are ionisable.
'urea', 'isone', 'epam', 'bartital'	These four classes of solutes are usually not available or not GC-amenable and require derivatization.
'alol', 'olol'	These two classes of solutes require derivatization.

than one or greater than four were ruled out. Because ADME BOXES supplied two log *P* values, a solute was considered as a failure if either of the log *P* values violated the inclusion rule. Table 1 summarizes the differences between the three screening rules.

After applying Baynes' rule, the process was continued by removing solutes from certain chemical classes. The list of the chemical classes and the reasons for removal are summarized in Table 2.

Figure 1 shows the workflow of the general process. This general process was rather an ideal one, as it assumed that all the solutes left in the screened set are workable. In practice, however, the training and validation sets selected based on this screened set had some solutes that were not workable as described above. In such scenarios a small modification to the general process was required as indicated in Figure 2.

2.4 *Selection criteria*

To select a 'representative' subset from a large molecular database, there are mainly three types of statistical optimal design approaches that have been reported in the literature: information-based designs [12], distance-based designs [12] and cell-based designs [19].

In least squares estimation of the linear parameters of a model, the variance is proportional to the inverse of the information matrix. An information-based design considers a subset as good if some function of the variance is minimized or equivalently, some function of the information matrix is maximized [12]. D-optimal designs are commonly used information-based designs that seek to maximize the determinant of the information matrix; they are widely used in the literature [7]. However, D-optimal designs are good only if the underlying models are linear and correctly specified; it will be a bad choice if the true model is nonlinear. In other words, because a D-optimal design is based

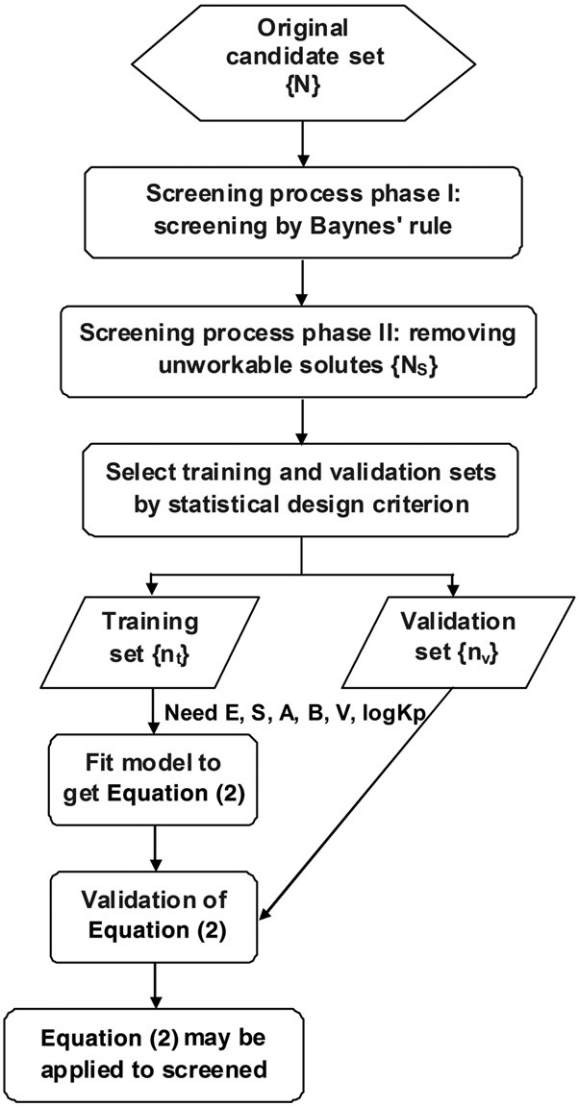

Figure 1. General process. The character in {} represents the size of the resulting dataset.

on the information matrix, it is dependent on the model to be fit. We want a 'model-free' training set selection process. For these reasons, D-optimality was not ideal for this work.

U-optimal designs and S-optimal designs are two types of distance-based designs with different focus. A distanced-based design checks the distance from a point x in p-dimensional Euclidean space R^p to a set A in that same p-dimensional Euclidean space, i.e. $A \subset R^p$. The distance $d(x, A)$ is defined as $d(x, A) = \min_{y \in A} \|x - y\|$, where $\|x - y\|$ is the p-dimensional distance in Euclidean space. According to Higgs [20], the U-optimal design focuses on the 'coverage' of a subset that has maximum similarity to the candidate set. The U-optimal criterion is evaluated by the mean distance from each candidate point (including the design points) to the design as $\frac{1}{N}\sum_{x \in C} d(x, D)$, where N is

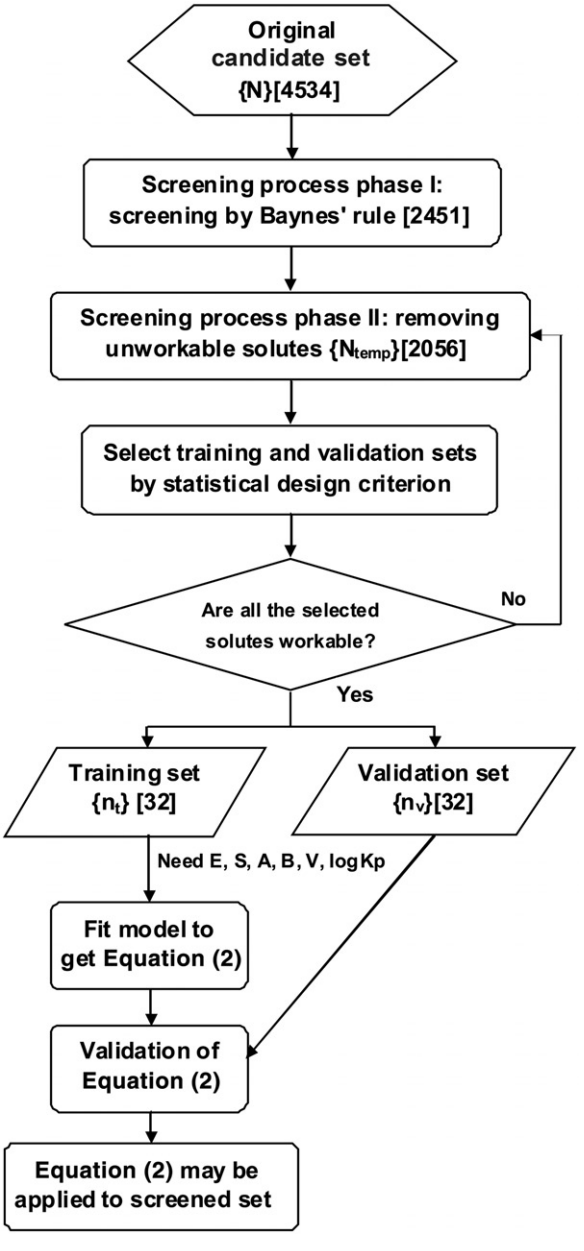

Figure 2. Modified process. The character in {} is general notation for the size of the resulting dataset. The character in [] represents size of the resulting dataset in the application discussed in this paper.

the number of points in the candidate set, C is the candidate set and D is the design set [12]. S-optimal designs emphasize the 'spread' of the design points with an aim to get a subset in which the design points are maximally dissimilar to each other. Different from the U-optimal design, S-optimality only examines the design set. It focuses on a subset in

which the harmonic mean distance from each design point to all the other points in the design is maximized [12]. The S-optimal criterion is $\frac{N_D}{\sum_{y \in D} 1/d(y, D-y)}$, where N_D is the number of design points. For both criteria, the smaller the value the better is the design. U-optimal and S-optimal are two popular designs and often serve as baselines for comparison with other newly introduced designs [21].

A third design class is called the cell-based design. The basic idea in the cell-based design is: for each of the k descriptors of the solutes, divide them into m bins to form m^k cells. The desired design will have at least one solute from each cell and thus the space is covered or filled. This conventional design has the problem of too many cells in high-dimensional space and a large portion of uncovered cells [19].

Lam et al. [19] suggested a Uniform Coverage Design to handle the 'curse of dimensionality' with the aim being to select a design that has the highest average uniform coverage design rate in all subspaces. Vijay et al. [10] adopted this design to select $n = 25$ solutes from 4098. Though this design claimed faster exchange time and better coverage than a U-optimal design and an S-optimal design, the algorithm is rather complicated and is only seen in a non-commercial software package. Thus, this design method was not considered for the current study.

For the information-based design (D-optimal) and the distance-based designs (U-optimal and S-optimal), implementations are available for SAS 9.2 [22] with the OPTEX procedure. The U-optimal design was chosen because it adequately fit the purposes of this study, i.e. to select a set that is diverse, covers the space well and has maximal similarity to the candidate set. In contrast, D-optimal designs seek to minimize the variance of the regression parameters and hence are very sensitive to model mis-specification while S-optimal designs seek to maximize the spread of the design points.

3. Results and analysis

3.1 *Principal component analysis (PCA)*

Principal component analysis (PCA) is a statistical procedure aimed at dimension reduction. It identifies a few linear combinations of the original variables that can explain the most variation in the original data [23]. These linear combinations are orthogonal to one another and are called 'principal components'. Usually, the majority of the variation in the original data can be explained by much fewer principal components than original variables. Thus, we can view these principal components as a new set of variables that represent the original variables and use them to interpret the variation in the data.

Data compressions of this form are highly desirable for graphically demonstrating key features of a dataset that has many variables, or for comparing this dataset to another set that may also have many variables. Table 3 summarizes PCA output for the large original set of 4534 solutes. With five variables in the set (E, S, A, B and V), five principal components are identified, where the first captures 47.0% of the variability seen in the original set, the first and second together capture 69.5% of the variability, etc. One would rarely consider using the first four principal components because 97.0% of variability explained is usually more than necessary. The first three principal components explain a comfortable 85.6% of the variability, but for ease of graphical display, the focus was only the first two principal components. Graphs of these principal components are discussed in Section 3.3.

Table 3. Summary of PCA on the original set of 4534 solutes.

Principal component	*Eigenvalue*	*Proportion of variability explained*	*Cumulative percentage of variability explained*
1	2.3507	0.4701	47.01
2	1.1243	0.2249	69.50
3	0.8037	0.1607	85.57
4	0.5718	0.1144	97.01
5	0.1495	0.0299	100.0

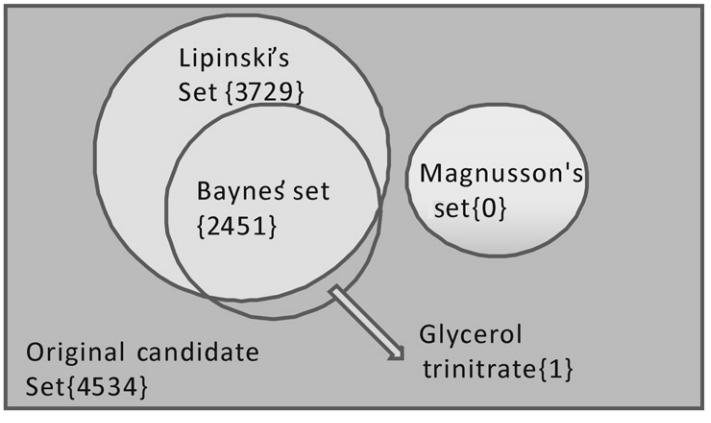

Figure 3. Screened results of the three inclusion rules. The character in {} represents the size of the resulting set.

3.2 *Screened set*

The candidate set (= 4534 solutes) was the starting point and obtained from the database supplied with ADME BOXES version 4.95 [5]. For the first phase of the screening process, Baynes' rule was used to eliminate solutes that suggest poor permeation through dermal delivery by limiting MW to less than 400 and log *P* between one and four, inclusively. This process resulted in 2451 solutes left in the screened set. It should be noted that, in addition, the Baynes' screened set was almost a subset of Lipinski's screened set, with the exception of the solute glycerol trinitrate. Magnusson's rule [18] was too strict for this study and resulted in an empty set after screening. The result of screening phase I is shown in Figure 3. The number in {} is the size of each set.

In practice, the second phase of screening was combined with the preliminary U-optimal selections. To be specific, starting with the 2451 solutes remaining after the first phase of screening, additional solutes were removed from certain chemical classes (see Section 2.2), thus resulting in an updated screened set of size N_{temp1}. This screened set of size N_{temp1} contained more solutes than the final screened set size N_s. U-optimal designs were obtained based on the updated screened set of size N_{temp1} to obtain a training set and a validation set. A careful examination of the solutes in these two selected sets showed that not all solutes were workable, e.g. some were not available from chemical supply houses, some were too expensive, or others were too toxic or difficult to use experimentally.

Table 4. Summary of PCA on the final screened set of 2056 solutes.

Principal component	*Eigenvalue*	*Proportion of variability explained*	*Cumulative percentage of variability explained*
1	2.6566	0.5313	0.5313
2	1.0574	0.2115	0.7428
3	0.8148	0.163	0.9058
4	0.2797	0.0559	0.9617
5	0.1915	0.0383	1.0000

Those unworkable solutes were removed from N_{temp1} to give N_{temp2} screened solutes, and then the U-optimal selection process was started again. The newly selected training set and validation set were checked and the unworkable solutes were removed. This process was repeated multiple times until the final workable U-optimal designs were obtained. The last updated screened set of size $N_s = 2056$ was the final screened set, from which the training set and validation set were selected. The modified process is illustrated in Figure 2. Compared with the general process in Figure 1, the difference is the U-optimal designs may be run several times to obtain the final training set and validation set.

The solutes in the screened set should be more similar to one another than those in the original candidate set, but at the same time, the screened set retains the majority of the variation seen in the original candidate set. PCA analysis shows that the first two principal components account for 69.5% of the total variation in the original set, while the percentage is 74.3% in the screened set. The eigenvalues in descending order, their proportions and the cumulative percentages for the two sets are shown in Tables 3 and 4.

3.3 *Training set and validation set*

Based on the screened set of size $N_s = 2056$, the selected training set and validation set are shown in Tables 5 and 6. To visualize coverage of the training set relative to the entire screened set, Figure 4 displays all 2056 screened solutes using their first two principal components and highlights the solutes selected for the $n_t = 32$ training set. With only a $32/2056 = 0.016$ proportional selection, the training set appears to be both diverse and representative of the screened set. Depicted in Figure 4, the training sets selected by both the U-optimal (marked as black dots) and S-optimal (marked as open squares, solutes not tabulated in this paper due to space restriction) criteria were plotted in the solute space for comparison.

From Figure 4, one can see the apparent difference between the two optimization criteria, which supports the discussion of the two designs in Section 2.3. Also depicted in Figure 4 is the training set selected by the D-optimal criterion (marked as grey dots). Only 10 unique solutes are selected for this training set of size 32 (i.e. many solutes are selected to be replicated) and six of these 10 unique solutes are also selected by the S-optimal criterion. As indicated by a reviewer, the D-optimal criterion tends to select solutes on the fringes of the solute space while ignoring solutes in the middle of the space, and this was not our desire.

Table 5. Training set ($n_t = 32$) with log P, MW, E, S, A, B and V descriptor values.

No	Solute	log P (AB)	log P (ACD)	MW	E	S	A	B	V
1	1-Bromo-2-methylpropane	2.5	2.56	137.02	0.34	0.37	0	0.12	0.8472
2	1-Fluoronaphthalene	3.1	3.5	146.16	1.14	0.97	0	0.13	1.103
3	1-Methylisoquinoline	2.57	2.42	143.18	1.28	1.03	0	0.51	1.1852
4	1-Naphthaleneethanol	2.71	2.61	172.22	1.54	1.31	0.3	0.7	1.4259
5	2,3,5-Trimethylphenol	2.6	2.86	136.19	0.86	0.84	0.52	0.41	1.1978
6	2,4-Dimethylaniline	1.88	1.86	121.18	0.95	0.95	0.2	0.49	1.098
7	2-Acetylnaphthalene	3	2.9	170.21	1.5	1.4	0	0.55	1.3829
8	2-Ethylthiophene	2.76	2.89	112.19	0.69	0.58	0	0.13	0.9229
9	2-Methyl-3-methoxypyrazine	1.37	1.21	124.14	0.65	0.81	0	0.6	0.9747
10	2-Methyl-3-nitroaniline	1.55	1.83	152.15	1.2	1.72	0.37	0.44	1.1313
11	2-Nitrotoluene	2.25	2.41	137.14	0.87	1.11	0	0.28	1.0315
12	3,3-Dimethylbutanal	2.14	1.6	100.16	0.14	0.59	0	0.46	0.9697
13	3,4-Dimethylbenzyl-alcohol	2.05	1.95	136.19	0.83	0.9	0.39	0.61	1.1978
14	3-Butylpyridine	2.51	2.78	135.21	0.63	0.73	0	0.45	1.2389
15	3-Methylpent-1-ene	2.99	3.25	84.16	0.08	0.08	0	0.06	0.911
16	3-Nitrobenzonitrile	1.26	1.17	148.12	1.02	1.6	0	0.47	1.0453
17	4-Chloro-2-methylaniline	2.15	2.22	141.6	1.08	1.13	0.3	0.31	1.0795
18	4-Chloro-3-methylphenol	2.83	2.89	142.58	0.92	1.02	0.67	0.22	1.0384
19	4-Ethoxyacetanilide	1.5	1.63	179.22	0.94	1.48	0.48	0.86	1.4542
20	6-Methylindole	2.58	2.6	131.17	1.2	1.09	0.44	0.49	1.0873
21	Dibutylamine	2.45	2.76	129.24	0.11	0.3	0.08	0.69	1.3356
22	Dimethyl-phthalate	1.93	1.64	194.18	0.78	1.4	0	0.84	1.4288
23	Dimethyl-pimelate	1.54	1.19	188.22	0.15	1	0	0.77	1.5255
24	Diphenylsulfone	2.42	2.4	218.27	1.57	2.15	0	0.7	1.6051
25	Ethyl-4-methylpentanoate	2.54	2.72	144.21	0.07	0.58	0	0.4	1.3102
26	Heptan-2-ol	2.48	2.29	116.2	0.19	0.36	0.33	0.56	1.1536
27	Methyl-trichloroacetate	2.01	2.11	177.41	0.41	0.81	0	0.42	0.9729
28	Methyl-2-chlorobenzoate	2.57	2.48	170.59	0.82	0.99	0	0.47	1.195
29	Methyl-4-hydroxybenzoate	1.97	1.86	152.15	0.9	1.37	0.69	0.45	1.1313
30	Pirimicarb	1.64	1.7	238.29	1.18	1.34	0	1.34	1.8945
31	4,4′-Isopropylidenediphenol	3.12	3.61	228.29	1.61	1.56	0.99	0.91	1.8643
32	Phenylbutazone	2.86	3.16	308.37	1.85	2.62	0	1.28	2.4329

3.4 *Model fitting and validation utilizing experimental data*

Ideally, experimental data for the selected training set of 32 solutes would be used to fit the model to evaluate our method of training set selection. However, during dose solution preparation, it was discovered that five solutes were toxic, one required derivatization and three solutes had very low boiling points; thus these nine solutes were not dosed. The remaining 23 solutes dosed in the experiments are listed in Table 7.

The experiments used flow-through (FT) diffusion cells with porcine skin obtained from weaning female Yorkshire pigs. Experimental details of this type of experiment can be found in Riviere and Brooks [24] and Baynes et al. [25]. All the 23 solutes were dissolved in water to deliver a dose of each solute of 100 μg/ml. Methanol was added at 4.6% to facilitate production of the dosing solution. The solution was dosed in seven porcine skin diffusion cells. The solutes that permeated the pig skin were collected in the receptor fluid of the flow-through cells over time. Gas chromatography–mass spectrometry (GC-MS) was used for quantitative analysis. The absorbed amount of each solute was denoted as

Table 6. Validation set ($n_v = 32$) with log P, MW, E, S, A, B and V descriptor values.

ID	Solute	*log P (AB)*	*log P (ACD)*	*MW*	*E*	*S*	*A*	*B*	*V*
1	1,4-Benzodioxan	1.62	1.88	136.15	0.87	1.07	0	0.35	1.007
2	1-Acetylnaphthalene	3	2.9	170.21	1.52	1.41	0	0.54	1.3829
3	1-Chloro-2-bromoethane	1.91	1.61	143.41	0.57	0.7	0.1	0.09	0.6878
4	1-Phenylbutan-2-one	2.23	1.97	148.2	0.75	1.14	0	0.66	1.2957
5	2,2-Dichloropropane	2.71	1.88	112.98	0.37	0.32	0	0.11	0.7761
6	2,2-Diethoxypropane	1.57	1.49	132.2	0	0.32	0	0.57	1.2123
7	2,5-Dibromobenzaldehyde	3.38	3.35	263.91	1.4	1.25	0	0.25	1.223
8	2-Isopropylphenol	2.97	2.82	136.19	0.84	0.88	0.52	0.38	1.1978
9	3,5-Dichloroanisole	3.74	3.8	177.03	0.94	0.87	0	0.18	1.1608
10	3-Chlorophenol	2.54	2.4	128.56	0.91	1.06	0.69	0.15	0.8975
11	3-Ethylaniline	1.78	1.93	121.18	0.94	0.95	0.23	0.45	1.098
12	3-Methylbenzyl-alcohol	1.69	1.49	122.16	0.82	0.9	0.39	0.59	1.0569
13	4-Chloroacetanilide	1.86	2.05	169.61	0.98	1.47	0.64	0.51	1.2361
14	4-Methyl-3-nitroaniline	1.55	1.83	152.15	1.2	1.72	0.37	0.43	1.1313
15	5-Methylindole	2.58	2.6	131.17	1.2	1.08	0.44	0.49	1.0873
16	6-Methylquinoline	2.57	2.54	143.18	1.31	0.95	0	0.55	1.1852
17	Benzylmethylether	2.24	1.96	122.16	0.65	0.77	0	0.48	1.0569
18	Cyanazine	2.46	2.19	240.69	1.73	2.24	0.45	0.97	1.7743
19	Diethyl-phthalate	2.8	2.7	222.24	0.73	1.4	0	0.86	1.7106
20	Ethyl-4-methylbenzoate	2.98	3.19	164.2	0.73	0.88	0	0.47	1.3544
21	Heptan-3-ol	2.18	2.29	116.2	0.18	0.36	0.33	0.56	1.1536
22	Hex-1-ene	3.01	3.43	84.16	0.08	0.08	0	0.07	0.911
23	Hexyl-propanoate	3.43	3.37	158.24	0.07	0.56	0	0.45	1.4511
24	Methyl-4-methylpentanoate	2.03	2.12	130.18	0.07	0.57	0	0.47	1.1693
25	*N,N*-Diphenylacetamide	2.49	2.05	211.26	1.77	1.85	0	0.56	1.7215
26	*N,N*-Dipropylacetamide	2.36	1.38	143.23	0.27	1.03	0	0.8	1.3513
27	*N*-Methylformanilide	1.38	1.09	135.16	0.98	1.45	0	0.52	1.1137
28	*o*-xylene	3.08	3.14	106.16	0.66	0.56	0	0.16	0.9982
29	Propyl-4-aminobenzoate	2.21	2.48	179.22	1.02	1.2	0.31	0.73	1.4542
30	2-Methyl-4-nitrophenol	2.1	2.03	153.14	1.1	1.6	0.78	0.25	1.0902
31	Triphenylphosphineoxide	3.99	2.87	278.28	2.01	2.08	0	1.5	2.1953
32	Propyl acrylate	1.76	1.86	114.14	0.19	0.62	0	0.42	0.9854

'Absorption' (μg). The steady state flux (Js) was calculated as the slope of the cumulative absorption/area vs. time curve. The permeability coefficient (*Kp* in cm/h) was calculated as the ratio between the steady state flux and the initial drug concentration applied to the pig skin surface (in μg/ml). Unfortunately, two of the 23 solutes were not absorbed by the fibre from the flow-through receptor fluid, and four of the dosed solutes had data below the limit of detection from our water dose. This resulted in 17 solutes and a total of 119 ($=7 \times 17$) valid permeation coefficients.

The data was fitted to Abraham's LFER model (Equation (1)) and the following version of Equation (2) was derived from the selected training set:

$$\widehat{\log Kp_i} = -1.14(0.26) + 0.92(0.16)E_i - 0.78(0.14)S_i + 0.03(0.14)A_i - 0.77(0.26)B_i - 0.39(0.33)V_i \quad (3)$$

where the numbers in parentheses are standard errors. The fit is summarized in Table 8 and discussed below.

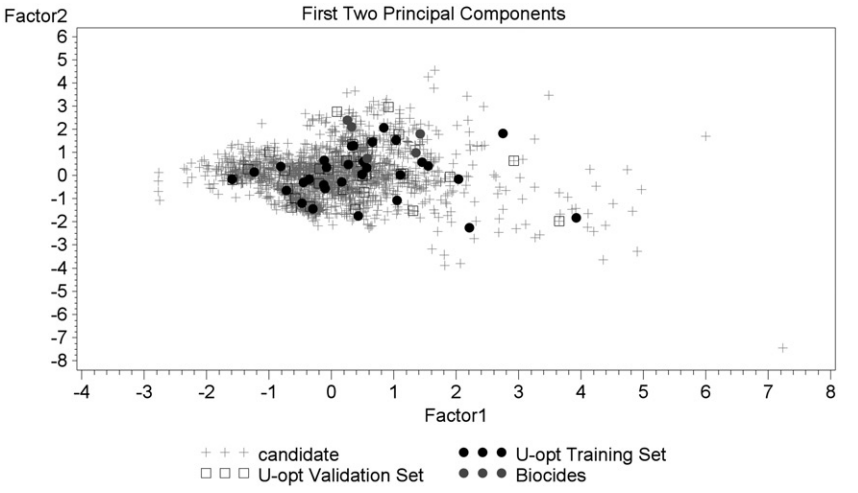

Figure 4. Scatterplot of first principal component (x axis) vs. second principal component (y axis) of U-optimal, S-optimal and D-optimal training sets each of size 32, relative to the screened set of size 2056. The black dots are U-optimal training set solutes, the open squares are S-optimal training set solutes and the grey dots are D-optimal training set solutes.
Note: The D-optimal set only has 10 unique solutes, of which six overlapped with the S-optimal set.

The goal of any design-of-experiments activity is to identify a training set that is 'better than' other training sets that were either arbitrarily created or created using alternative methods. The U-optimally-selected training set from Table 7 is to be compared with another set that will serve as a competitor 'training set'. The set from Vijay et al. [1] (henceforward referred to as 'Vijay's set') was chosen for this purpose because it has been successfully applied in multiple studies, the data were collected under similar experimental conditions in the same laboratory, and the set contains information on five biocides that are particularly relevant for our study on dermal permeation. The 35 solutes of Vijay's set were dosed in water and each solute had four replicate measurements, resulting in a total of 140 observations in this set. Using Vijay's set as training data, Abraham's LFER model (Equation (1)) was fitted and the following version of Equation (2) was derived:

$$\widehat{\log Kp_i} = -2.50(0.17) + 0.04(0.15)E_i + 0.83(0.19)S_i - 0.13(0.09)A_i - 1.00(0.21)B_i + 0.27(0.16)V_i \quad (4)$$

The fit is summarized in Table 9 and discussed below.

3.4.1 *Model comparison by Q^2_{LOO} and Q^2_{EXT}*

As discussed by Gramatica [26], it is important to assess the predictive power of a given QSAR model. One method of assessing the predictive power is by a leave-one-out (LOO) internal validation. In the leave-one-out method, each observation is left out once and a QSAR model is fitted with the remainder of the data until all the observations have been

Table 7. Final training set of dosed solutes with observed log *Kp* values.

ID	*Solute*	*Obs. #1*	*Obs. #2*	*Obs. #3*	*Obs. #4*	*Obs. #5*	*Obs. #6*	*Obs. #7*	*Mean (SD)*
1	Heptan-2-ol	−1.73	−1.73	−1.86	−1.74	−1.69	−1.74	−1.74	−1.75(0.05)
2	2-Ethylthiophene	−1.94	−2.33	−2.69	−1.39	−1.52	−1.65	−1.64	−1.88(0.47)
*3	Ethyl-4-methylpentanoate	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
4	2-Methyl-3-methoxypyrazine	−1.87	−1.83	−1.96	−1.94	−1.81	−1.86	−1.87	−1.88(0.06)
5	3-Butylpyridine	−1.52	−1.60	−1.67	−1.51	−1.46	−1.56	−1.51	−1.55(0.07)
6	1-Fluoronaphthalene	−1.38	−1.52	−1.58	−1.17	−1.17	−1.35	−1.38	−1.36(0.16)
7	2,3,5-Trimethylphenol	−1.77	−1.85	−1.98	−1.73	−1.66	−1.78	−1.76	−1.79(0.10)
8	3,4-Dimethylbenzyl-alcohol	−2.06	−2.05	−2.23	−1.97	−1.87	−1.98	−1.98	−2.02(0.11)
9	4-Chloro-3-methylphenol	−1.73	−1.83	−1.90	−1.67	−1.58	−1.71	−1.68	−1.73(0.11)
*10	Methyl-2-chlorobenzoate	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
*11	Dimethyl-pimelate	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
12	1-Methylisoquinoline	−1.72	−1.78	−1.82	−1.71	−1.63	−1.73	−1.71	−1.73(0.06)
13	6-Methylindole	−1.40	−1.48	−1.56	−1.40	−1.34	−1.45	−1.41	−1.43(0.07)
14	3-Nitrobenzonitrile	−1.90	−1.89	−2.07	−1.91	−1.79	−1.90	−1.93	−1.91(0.08)
15	Methyl-4-hydroxybenzoate	−2.29	−2.21	−3.22	−2.35	−2.17	−2.23	−1.84	−2.33(0.42)
16	Dimethyl-phthalate	−3.46	−3.46	−4.20	−3.93	−3.07	−3.43	−3.71	−3.61(0.37)
17	2-Acetylnaphthalene	−1.88	−1.86	−1.94	−2.04	−1.77	−2.15	−1.75	−1.91(0.14)
18	1-Naphthaleneethanol	−1.96	−2.02	−2.15	−1.90	−1.83	−1.99	−2.06	−1.99(0.10)
19	Pirimicarb	−2.79	−2.69	−3.01	−2.84	−2.65	−2.85	−2.81	−2.81(0.12)
20	Diphenylsulfone	−2.12	−2.16	−2.28	−2.07	−2.00	−2.18	−2.18	−2.14(0.09)
*21	4,4′-Isopropylidenediphenol	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
**22	Dibutylamine	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
**23	Methyl-trichloroacetate	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

*Below the limit of detection, no data available.
**Not absorbed by fibres, no data available.

Table 8. Fit and prediction statistics of Equation (3), built from the U-optimally selected training set and applied to the external Vijay's set.

Statistics	*Model results*
n	119
RMSE	0.38
r^2	0.55
adj-r^2	0.53
Q^2_{LOO}	0.50
Q^2_{EXT}	0.02
Q^2_{EXT} for the five biocides	0.81

Table 9. Fit and prediction statistics of Equation (4), built from Vijay's set and applied to the external U-optimally selected training set.

Statistics	*Model results*
n	140
RMSE	0.29
r^2	0.27
adj-r^2	0.25
Q^2_{LOO}	0.19
Q^2_{EXT}	−0.31

predicted. The predicted response of the fitted model is $\hat{y}_{i,-i}$. The final measured Q^2 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 ith observed response value, $\hat{y}_{i,-i}$ is the leave-one-out predicted value of the ith observation obtained from the model fitted without observation 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.

Compared to internal validation, external validation is another and usually more reliable manner to evaluate the predictive power of a model. In external validation, the fitted model is used to predict the response for observations from an additional set (a validation set) different from the training set. Q^2_{EXT} is used to measure the predictive power of the model [26]. If there are m observations in the validation set, the Q^2_{EXT} is defined as follows:

$$Q^2_{EXT} = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{6}$$

Table 10. Five biocides from Vijay et al. [1] with E, S, A, B and V descriptors.

ID	*Solute*	*E*	*S*	*A*	*B*	*V*
1	1,2-Benzisothiazolinone	1.47	1.68	0.26	0.77	1.0277
2	*o*-Phenylphenol	1.55	1.4	0.56	0.49	1.3829
3	*p*-Chloro-*m*-xylenol	0.93	0.96	0.64	0.21	1.1793
4	*p*-Chloro-*m*-cresol	0.92	1.02	0.67	0.22	1.0384
5	4-tert-Amylphenol	0.79	0.8	0.5	0.44	1.4796

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. Like Q^2_{LOO}, large values of Q^2_{EXT} are preferred and Q^2_{EXT} is not limited to the range of [0,1].

Because experimental log *Kp* values have not yet been obtained for the validation set of Table 6, Vijay's set was used as a validation set for the model shown in Equation (3) built from our U-optimally-selected training set. Reversing roles, the model shown in Equation (4) built from Vijay's set used our U-optimally-selected training set as a validation set. Comparison of Tables 8 and 9 indicates that the U-optimally selected training set has both a better model fit (higher r^2, adj-r^2) and better predictive power (higher Q^2_{LOO} and Q^2_{EXT}). The Q^2_{EXT} in Table 9 is negative, which means the numerator is greater than the denominator in Equation (6). This indicates that prediction of the validation set by the model fitted with Vijay's set is not as good as prediction using just the average of the training set log *Kp* values. Vijay's set was not selected using the U-optimal design criterion and this negative value for Q^2_{EXT} further indicates that a non-U-optimal set may be insufficient. Although neither model gave adequate external prediction, the relative benefit of our carefully selected training set is evidenced by the dramatic differences in performance of the two models. Additional comments regarding performance may be found in the Discussion.

Among the 35 solutes in Vijay's set, there are five biocides (listed in Table 10) that are widely used in the metalworking industry. To measure how well the QSAR model built from our approach (Equation (3)) can predict these biocides, external validation was conducted separately with the five biocides (in total 20 observations). A much higher Q^2_{EXT} of 0.81 was obtained, which indicates that Equation (3) is better at predicting the biocides than predicting Vijay's set as a whole. The excellent predictivity can be explained by the locations of the five biocides in the solute space as shown in the first two principal components plot (Figure 5). For reference, both the U-optimal training set and the U-optimal validation set were plotted with the five biocides in the same figure. One can see that the five biocides are very close to the U-optimal training set and validation set, which suggests the five biocides can be predicted well by the model built on the U-optimal training set.

3.4.2 *Model comparison by applicability domain*

No model can be robust enough to reliably predict all possible solute permeations. The applicability domain (AD) of a given model must be defined. The AD is defined as the response and chemical structure space in which the model makes predictions with a given reliability [27]. The Williams plot was used to visualize the AD [14,28]. This is a plot of the standardized residuals versus the leverage values (h). Standardized residuals result from the Q^2_{EXT} calculations, obtained as $\frac{r_i - mean(r)}{stdev(r)}$, where r_i is the ith raw residual calculated as the

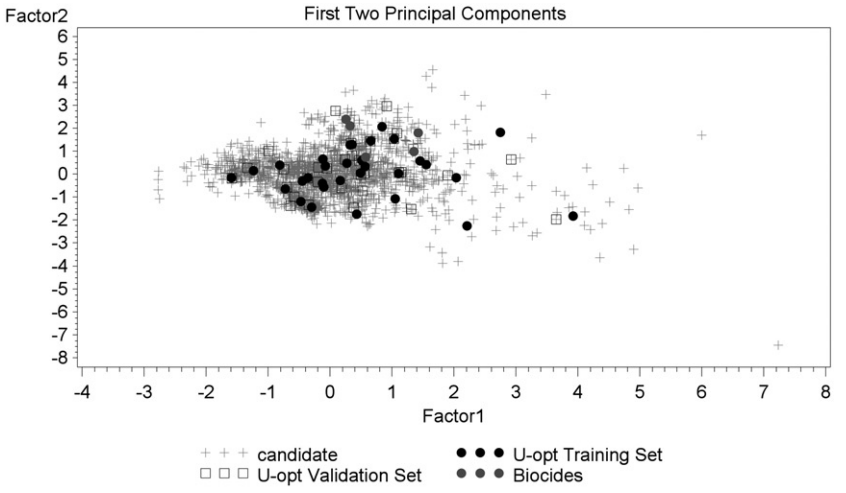

Figure 5. Scatterplot of first principal component (x axis) vs. second principal component (y axis) of U-optimal training set (size 32), U-optimal validation set (size 32) and biocides (size 5), relative to the screened set of size 2056. The grey dots are U-optimal training set solutes, while the open squares are U-optimal validation set solutes and the black dots are biocides.

difference between the observed log Kp and predicted log Kp, $i = 1, \ldots, n+m$. Leverage for observation i is calculated as: $h_i = x_i^T(X^TX)^{-1}x_i$, where X^TX is the information matrix from the training set, x_i is a vector (for each observation either in the training set or the validation set) containing the constant 1 as the first element and p predictors as the remaining p elements. A cutoff leverage or critical hat value $h^* = 3(p+1)/n$ is defined, where, in this case, p is the number of model variables and n is the number of observations used to fit the model. The AD is the area of the plot where $h < h^*$ [14,28].

Figure 6 is the Williams plot produced using the U-optimally selected training set and using Vijay's set as the validation set. The vertical reference line in the figure is the critical hat value h^*. It is clear that only one solute (observations displayed in exactly the same column belong to the same solute) from the validation set falls outside the AD, which indicates that the selected training set has an excellent coverage of the validation set. This solute that falls outside the AD is 1,2-benzisothiazolinone, one of the biocides. Figure 7 is the observed vs. predicted log Kp for both the U-optimally selected training set and Vijay's set. From Figure 7 one can see that the five biocides were predicted relatively well.

Using Vijay's set as the training set and the U-optimally selected training set as a validation set, Figures 8 and 9 show the resulting Williams plot and plot of observed vs. predicted log Kp values. Figure 8 suggests that, when using Vijay's set as the training set, two solutes in the validation set fall outside the AD. A further calculation shows that among the 2056 solutes in the screened set, only 295 (14.3%) of the solutes fall outside the AD defined by the model fitted from the U-optimal set, while there are 592 (28.8%) solutes falling outside the AD defined by the model fitted from Vijay's set. This means that of the 2056 solutes, more solutes can be predicted reliably by the LFER model built upon the U-optimal set than the model built on Vijay's set. Compared with Figure 7, the observations seen in Figure 9 for the validation set are more widely scattered around the regression line, indicating a weaker potential to predict the validation set. Figures 8 and 9 again support the advantage of the training set selection approach.

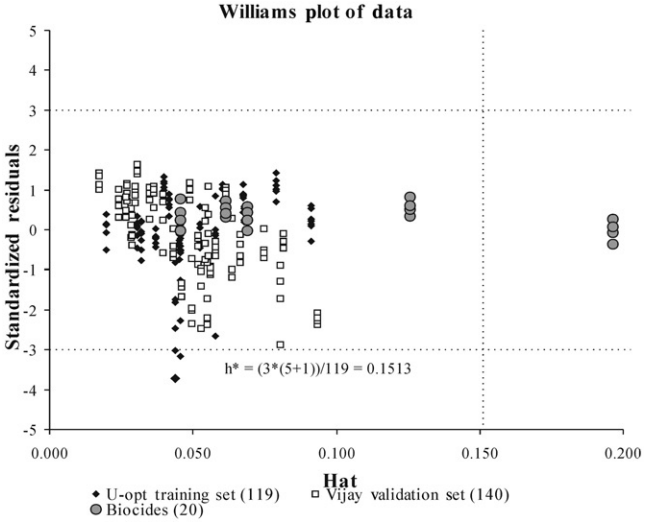

Figure 6. Williams plot of data with the U-optimal set in Table 7 as the training set and the Vijay et al. [1] data as the validation set.

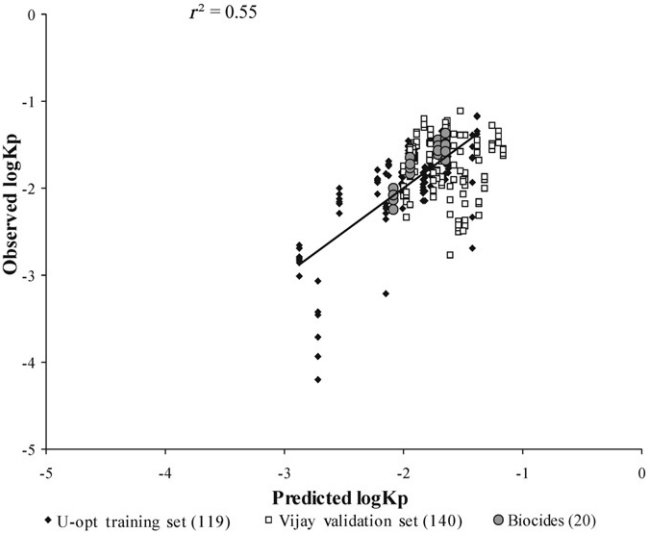

Figure 7. Observed log *Kp* vs. predicted log *Kp* with the U-optimal set in Table 7 as the training set and the Vijay et al. [1] data as the validation set.

4. Summary and discussion

In summary, a principled yet practical and transparent procedure was applied to select diverse and representative training and validation sets. Two steps were carried out based on the original large candidate set. The first step was the screening process, which removed solutes with possible poor permeability by 'Baynes' rule' and unavailable or extremely toxic solutes by identification of their chemical suffix. The second step was to select

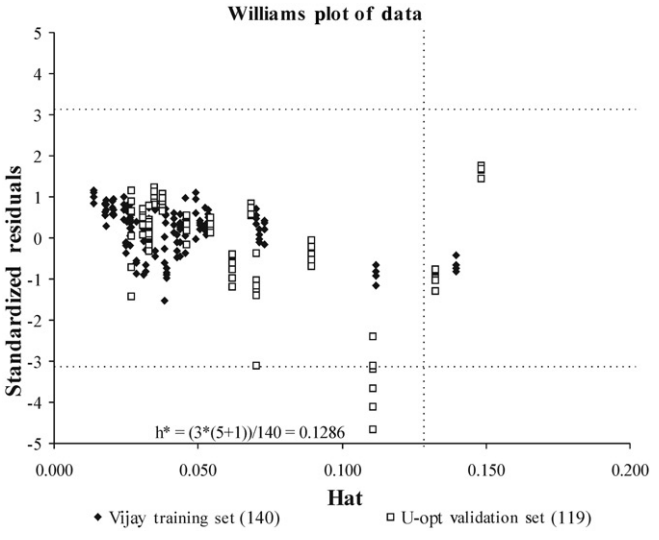

Figure 8. Williams plot of data with the Vijay et al. [1] data as the training set and the U-optimal data as the validation set.

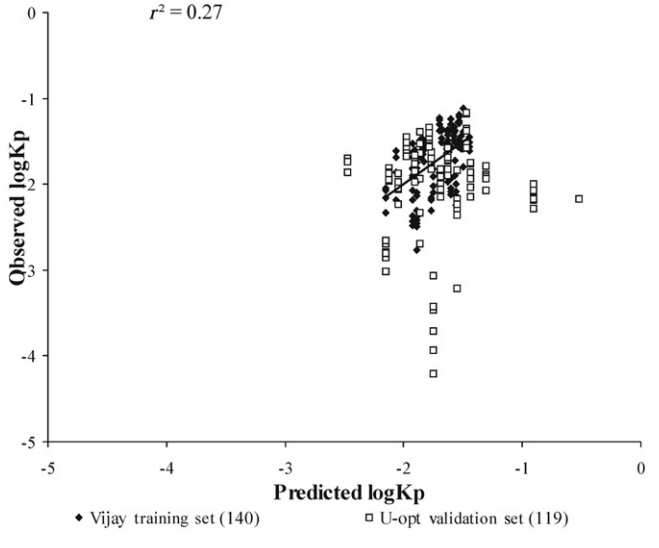

Figure 9. Observed log *Kp* vs. predicted log *Kp* with the Vijay et al. [1] data as the training set and the U-optimal data as the validation set.

training and validation sets using the U-optimal criterion. Principal component analysis showed that the proposed procedure selected solutes representative of the solute space. Another merit of the procedure was the ease of use of the standard U-optimal design using SAS software. The resulting predictive models have demonstrated a relative benefit of the selected set compared to an existing set that has been successfully used in the past.

Another aim of this paper was to remain true to the Abraham LFER model as much as possible due to its popularity in modelling dermal permeability. This aim limited our ability to obtain more predictive models. There are many additional things, however, that may be done to further improve predictive performance. One option is to use one of the many alternative LFER models that are available, especially those that use different molecular descriptors from the Abraham LFER model. Increasing the number of solutes would provide another major move in the direction of improved prediction, but due to limited resources, we were not able to make such increases for the current experiment. Another option is to consider any number of nonlinear models, especially those that can automatically accommodate interactions and the effects of variable thresholding. One such model is the random forest technique that is comprised of many recursively partitioned trees [29]. Using default settings of the randomForest package in R [30], for the model built from our selected set and applied to Vijay's set, Q^2_{EXT} increased from 0.02 to 0.27. This increase is dramatic, although a Q^2_{EXT} of 0.27 is still not impressive. There was a similar increase for the model built from Vijay's set and applied to our selected set, namely −0.31 to 0.07. Other model improvements are possible, and these investigations form the basis of future studies. It is also relevant to point out that the Q^2_{EXT} for predicting the five biocides is much higher than that for predicting the entire Vijay's set, which suggests the selected training set is better at predicting biocides that are a very important class of solutes for our study regarding dermal permeation.

Some may claim that instead of splitting to create a small training set and a small validation set, better fitted models may be achieved if we used the U-optimal criterion to select a larger set and implemented some appropriate cross-validation technique [31]. We agree that combining the training set and validation set into a larger single sample and using cross-validation could lead to a better and more powerful model, especially in situations such as ours where training and validation sets are small. However, this is not the standard practice in QSAR studies concerning dermal permeation (see OECD guidelines on QSAR model validation [32]). As a result, we restrict ourselves to the philosophy of separate training and validation sets for this project.

As mentioned in the Introduction, several other criteria have been proposed for selecting training and validation sets for QSAR studies. We recommend that all these alternative proposals be fully considered for each application, since all have admirable qualities. The U-optimal criterion that we recommend in this paper is not driven by the same sort of mathematical guarantees that accompany the D-optimal criterion. At the same time, however, those mathematical guarantees of D-optimality are only supported under rather strict assumptions of properly specifying the yet-to-be-determined model. As mentioned above, Abraham's LFER model does not appear to provide as good predictive power when compared to a nonlinear random forest model, so building a D-optimal design from Abraham's LFER model would not be desirable. On the other hand, the U-optimal criterion is a model-free approach that is intuitively attractive and has now been shown to provide relative improvements in predictive ability.

This study suggests several avenues of possible future improvement. The screening process is critical. Ideally, screening should occur once with the subsequently identified training and validation sets being those solutes used for experimentation. In this study, it was not possible to anticipate the many practical limitations that resulted in redefining the screened set and ultimately using a training set of only 17 solutes when the original intent was to use 32 solutes. Given the particular application area of dermal permeability where experimentation can be costly, sample sizes are necessarily small to begin with and any

reductions discovered during experimentation can have undue influence on the resulting predictive models. Despite these difficulties, the steps proposed in this paper are repeatable and effective for principled identification of training and validation sets starting from large candidate sets.

In the future, further experiments using porcine skin will be conducted in our laboratory with the U-optimal validation set. It is expected that the resulting permeation coefficients (log *Kp*) will add to the robustness of the LFER models.

Acknowledgements

The authors would like to express their gratitude to the reviewers for their valuable comments and suggestions. This research was funded by the NIOSH grant #OH-003669.

References

[1] 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.

[2] M.D. Barratt, *Quantitative structure-activity relationships for skin permeability*, Toxicol. Vitro 9 (1995), pp. 27–37.

[3] R.O. Potts and R.H. Guy, *Predicting skin permeability*, Pharm. Res. 9 (1992), pp. 663–669.

[4] 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.

[5] *ADME BOXES 4.95*, Pharma Algorithms, Toronto, Canada, 2009, software available at http://pharma-algorithms.com/adme_boxes.htm

[6] S. Wold and W.J. Dunn, *Multivariate quantitative structure-activity relationships (QSAR): Conditions for their applicability*, J. Chem. Inf. Comput. Sci. 23 (1983), pp. 6–13.

[7] L. Eriksson and E. Johansson, *Multivariate design and modeling in QSAR*, Chemom. Intell. Lab. Syst. 34 (1996), pp. 1–19.

[8] G.L. Flynn, *Physicochemical determinants of skin absorption*, in *Principles of Route-to-Route Extrapolation for Risk Assessment*, T.R. Gerrity and C.J. Henry, eds., Elsevier, Amsterdam, Amsterdam, 1990, pp. 93–127.

[9] E. Furusjö, A. Svenson, M. Rahmberg, and M. Andersson, *The importance of outlier detection and training set selection for reliable environmental QSAR predictions*, Chemosphere 63 (2006), pp. 99–108.

[10] V. Vijay, R.E. Baynes, S.S. Young, and J.E. Riviere, *Selection of appropriate training set of chemicals for modeling dermal permeability using uniform coverage design*, QSAR Comb. Sci. 28 (2009), pp. 1478–1486.

[11] C.A. Lipinski, F. Lombardo, B.W. Dominy, and P.J. Feeney, *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*, Adv. Drug Delivery Rev. 23 (1997), pp. 3–25.

[12] SAS Institute, *SAS/QC software: Usage and reference, Version 6*, Institute SAS, Cary, NC, 1995.

[13] 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.

[14] P. Gramatica, *Principles of QSAR models validation: Internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701.

[15] D. Weininger, *Smiles, a chemical language and information-system. 1. Introduction to methodology and encoding rules*, J. Chem. Inf. Comput. Sci. 28 (1988), pp. 31–36.

[16] *ChemDraw Ultra 6.0*, CambridgeSoft Corporation, Cambridge, MA, 2001, software available at http://products.camsoft.com

[17] *JChem for Excel 5.4.1.157*, ChemAxon, Budapest, Hungary, 2001, software available at http://www.chemaxon.com
[18] B.M. Magnusson, W.J. Pugh, and M.S. Roberts, *Simple rules defining the potential of compounds for transdermal delivery or toxicity*, Pharm. Res. 21 (2004), pp. 1047–1054.
[19] R.L.H. Lam, W.J. Welch, and S.S. Young, *Uniform coverage designs for molecule selection*, Technometrics 44 (2002), pp. 99–109.
[20] R.E. Higgs, K.G. Bemis, I.A. Watson, and J.H. Wikel, *Experimental designs for selecting molecules from large chemical databases*, J. Chem. Inf. Comput. Sci. 37 (1997), pp. 861–870.
[21] C.H. Reynolds, A. Tropsha, L.B. Pfahler, R. Druker, S. Chakravorty, G. Ethiraj, and W. Zheng, *Diversity and coverage of structural sublibraries selected using the SAGE and SCA algorithms*, J. Chem. Inf. Comput. Sci. 41 (2001), pp. 1470–1477.
[22] SAS 9.2, SAS Institute Inc., Cary, NC, 2008, software available at http://www.sas.com
[23] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed., Pearson Prentice Hall, Upper Saddle River, NJ, 2006, pp. 430–436.
[24] J.E. Riviere and J.D. Brooks, *Predicting skin permeability from complex chemical mixtures*, Toxicol. Appl. Pharmacol. 208 (2005), pp. 99–110.
[25] 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.
[26] P. Gramatica, *Evaluation of Different Statistical Approaches for the Validation of Quantitative Structure-Activity Relationships*, European Centre for the Validation of Alternative Methods, Ispra, 2004.
[27] T.I. Netzeva, G. Saliner, and A.P. Worth, *Comparison of the applicability domain of a QSAR for estrogenicity with a large chemical inventory*, Environ. Toxicol. Chem. 25 (2005), pp. 1223–1230.
[28] M. Pavan, T.I. Ntezeva, and A.P. Worth, *Validation of a QSAR model for acute toxicity*, SAR QSAR Environ. Res. 17 (2006), pp. 147–171.
[29] L. Breiman, *Arcing classifiers (with discussion and a rejoinder by the author)*, Ann. Stat. 26 (1998), pp. 801–849.
[30] A. Liaw and M. Wiener, *Classification and Regression by randomForest*, R News 2(3) (2002), pp. 18–22.
[31] D.M. Hawkins, J.J. Kraker, S.C. Basak, and D. Mills, *QSPR checking and validation: A case study with hydroxy radical reaction rate constant*, SAR QSAR Environ. Res. 19 (2008), pp. 525–539.
[32] OECD, *Guidance Document on the Validation of (Quantitative) Structure–Activity Relationships [(Q)SAR] Models*, ENV/JM/MONO(2007)2, OECD Environmental Health and Safety Publications Series on Testing and Assessment No. 69, Organisation for Economic Co-operation and Development, Paris, 2007.