The Application of Stochastic Machine Learning Methods in the Prediction of Skin Penetration

Y. Sun*[,a], M.B. Brown[b], M. Prapopoulou[b], N. Davey[a], R.G. Adams[a], G.P. Moss[c]

[a]*School of Engineering & Information Science,*
[b]*School of Pharmacy,University of Hertfordshire, Hatfield, UK*
[c]*School of Pharmacy, Keele University, Keele, UK*

Abstract

Improving predictions of skin permeability is a significant problem for which mathematical solutions have been sought for around twenty years. However, the current approaches are limited by the nature of the models chosen and the nature of the dataset. This is an important problem, particularly with the increased use of transdermal and topical drug delivery systems. In this work, we apply K-nearest-neighbour regression, single layer networks, mixture of experts and Gaussian processes to predict the skin permeability coefficient of penetrants. A considerable improvement, both statistically and in terms of the accuracy of predictions, over the current quantitative structure-permeability relationship (QSPRs) was found. Gaussian processes provided the most accurate predictions, when compared to experimentally generated results. It was also shown that using five molecular descriptors - molecular weight, solubility parameter, lipophilicity, the number of hydrogen bonding acceptor and donor groups - can produce better predictions than when using only lipophilicity and the molecular weight, which is an approach commonly found with QSPRs. The Gaussian process regression with five compound features was shown to give the best performance in this work. Therefore, Gaussian processes would appear to

*Corresponding author
Email addresses: comrys@herts.ac.uk (Y. Sun), marc.brown@mdepharm.co.uk (M.B. Brown), maria.2.prapopoulou@kcl.ac.uk (M. Prapopoulou), n.davey@herts.ac.uk (N. Davey), r.g.adams@herts.ac.uk (R.G. Adams), g.p.j.moss@mema.keele.ac.uk (G.P. Moss)

provide a viable alternative to the development of predictive models for skin absorption, and underpin more realistically mechanistic understandings of the physical process of the percutaneous absorption of exogenous chemicals.

Key words: Skin Penetration, QSARs, QSPRs, Gaussian Process Regression

---

1. Introduction

Predicting percutaneous absorption accurately has proven to be a major challenge and one which has substantial implications for pharmaceutical and cosmetic industries, as well as toxicological issues in fields such as pesticides usage. Several approaches have been used to try to quantify and predict skin absorption. One such method involves the use of quantitative structure-activity (or permeability) relationships (QSARs, or QSPRs), and another is the use of mathematical modelling [6]. These approaches have been extensively reviewed [17]. Recently, more new approaches, for example, artificial neural network and fuzzy modelling, have been applied to this problem domain [4], with varying degrees of success.

Therapeutically relevant percutaneous absorption has presented a significant challenge for pharmaceutical scientists for the last 50 years. As knowledge of the detailed structure of the skin barrier - the *stratum corneum*, the skin's outermost layer - increased, new technologies gradually became available for the treatment of medical conditions by transdermal therapy. The stratum corneum is the main barrier to percutaneous absorption, due to its unique structure and properties. It is a very thin layer, commonly $15 - 30mm$ on the volar forearm, for example, although it may be thicker or thinner at different sites on the body. This layer effectively governs the rate of passage of exogenous chemicals across the skin and into the viable tissues from the external environment. It is a densely packed layer consisting of dead, flattened keratin cells enmeshed in a lipid domain [5]. It is generally held that the most common route of absorption across the skin is via the lipid pathway [17].

While qualitative estimates of percutaneous absorption were common until

the 1980's, it was not until 1990, and the publication of the Flynn dataset [7] that a quantitative approach to skin absorption was proposed. Flynn determined, in a semi-quantitative manner, that skin absorption was influenced predominately by two compound descriptors - the lipophilicity of a molecule (P) and its molecular weight (MW). The former term, P, is the ratio of the solubility of a molecule between two phases; octanol, to represent the lipid phase, and water (or a buffered aqueous solution) to represent the aqueous phase. Normally, this gives quite a range as some molecules will prefer one phase to another, often across as wide a range as 10^{-7} to 10^{7}. Hence, a log scale is used to simplify the notation in common use. Potts and Guy [25] used the Flynn dataset to derive a linear equation that quantified percutaneous absorption:

$$\log K_p = 0.71 \log P - 0.061 MW - 6.3, \tag{1}$$

where K_p is the permeability coefficient, $\log P$ the octanol-water partition coefficient and MW the molecular weight of the penetrant. It is important to note that $\log K_p$ is a completely different term to $\log P$. The amount of drug that passes across the skin is measured as concentration (in suitable units) against time. This gives us a rate term which we call flux (J). However, to compare the relative rates of drug release for molecules which may have different properties (particularly different solubility and $\log P$) we have to correct for differences in concentration. K_p is defined as follows:

$$K_p = J/\Delta C_m \tag{2}$$

where ΔC_m denotes the concentration difference across the membrane. Thus, K_p is a concentration corrected version of flux that allows comparison of permeation for different molecules. A number of similar equations have been derived since the publication of Potts and Guy's model. For example, Moss and Cronin [16] developed Potts and Guy's model by evaluating a slightly larger and more robust dataset. The model is represented by the following equation:

$$\log K_p(cm/s) = 0.74 \log P - 0.0091 MW - 2.39, \tag{3}$$

where $\log K_p$, $\log P$ and MW are as defined earlier. In [17], authors have reviewed extensively similar QSAR equations. In general, these models offer linear relationships to quantify percutaneous absorption. It is worth reflecting on the implications of these consistent findings in the context of recent work by Moss et al., [19], which suggests that the dataset employed for skin absorption is fundamentally non-linear in nature.

Moss et al., [18] investigated this further, and compared a series of published models. They showed that there were significant differences between $\log K_p$ values that were measured experimentally and those that were determined using the Potts and Guy (and other, similar) equations. Interestingly, they showed that the greatest difference between experimental and predicted values was found at high $\log P$ values. This was reinforced by a detailed examination of the distribution of the permeability data, which showed no linear trends and a clear Gaussian distribution, suggesting that the use of linear models to represent skin permeability might not provide the most accurate of predictive models, and that their approach to predicting permeability was limited to molecules with $\log P < 3.0$. It should also be noted that the use of such models in this manner is inappropriate, as it does not fully reflect the spread of the dataset.

One problem addressed in the current study is how predictions of K_p may be improved by applying advanced machine learning techniques, such as Gaussian Processes [27]. One key feature of this problem domain is that the target, the skin permeability coefficient (K_p), has a strongly non-linear relationship with the compound descriptors (features). This has been determined previously by Moss et al., [19], who used principal component analysis to explore the mathematical nature of the dataset commonly used to generate mathematical models of skin absorption. As this work clearly shows the inherently non-linear nature of the data underpinning these models, it clearly raises issues over the extensive prior use of linear models and their validity and accuracy in estimating percutaneous absorption. It may also be suggested that this study shows the limitations of the range of previous models, compared with previous models.

Currently, most QSPR-type models used to predict skin absorption suggest

in general that only two molecular parameters, molecular weight and lipophilicity (indicated by the octanol-water partition coefficient, $\log P$) are of relevance to the percutaneous absorption of exogenous chemicals. However, a more specific analysis, such as that conducted by Potts and Guy [26], Pugh et al., [23] and others [24], has shown that other parameters may be of significance for certain types of molecule, and may indeed give a more detailed description of a penetrant's ability to pass into and across the skin.

Hydrogen-bonding, despite being absent from the Potts and Guy (1992) model [25] and from its variants, has been considered as a key parameter in percutaneous absorption for just over thirty-five years [28]. Further, consideration of partition phenomena, particularly the development of the solvatochromic theory [13] and developments in the understanding of epidermal permeability ([1], [26], [29], [30]) clearly indicated the importance of hydrogen-bonding acceptor and donor properties in understanding the underlying mechanisms governing the percutaneous absorption of exogenous chemicals. For example, Roberts et al. ([30]) showed that the introduction of even one hydrogen-bonding group to a molecule resulted in a significant decrease to its ability to permeate successfully across the skin [30]. Addition of further groups to the molecule results in further decreases, which were non-linear and not as large as the addition of the first hydrogen-bonding group. They concluded that hydrogen-bonding was the major factor in diffusion across the *stratum corneum*, and that lipophilicity, usually represented by $\log P$, was more important for partitioning.

Therefore, the aims of the current study are to demonstrate the feasibility of prediction improvement by using computational regression modelling methods, particularly Gaussian processes. Further, it is also the aim of this current study to investigate the introduction of new compound descriptors to aid the problem and to provide a mechanistic insight to the nature of percutaneous absorption.

2. Methods

2.1. Theoretical background: modelling methods

2.1.1. QSPR analysis

Prior to the application of the modeling methods described below to the dataset, the QSPR methods were applied to the data in order to provide a comparison between machine learning methods and previous approaches to this matter. The methods used are those reported previously (eqs. (1) and (3)). Further details on the nature of these models may also be found elsewhere ([3] and [17]).

2.1.2. Single layer networks

Regression analysis was initially carried out on the dataset using a single layer network (SLN). This simple linear regression considers the output y as the weighted sum of the components of an input vector $\mathbf{x}$, which can be written as follows [2]:

$$y = y(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^{d} w_i x_i + w_0 \ , \tag{4}$$

where d is dimensionality of the input space and $\mathbf{w} = (w_1, \ldots, w_d, w_0)$ is the weight vector. The weights are set so that the sum squared error function is minimised on a training set.

2.1.3. K-nearest-neighbour (KNN) regression

Given a test input vector $\mathbf{x}$, the algorithm finds the K closest points to $\mathbf{x}$ among all the training inputs. The prediction of the model is therefore the average of those K target values.

2.1.4. Mixture of experts - MIXEXP

The mixture of experts [11] divides the input space into a nested set of regions. In each region a simple surface is fitted to the data. It consists of a gating network and experts. The function of the gating network is to partition the input space so that each expert only needs to model a small region. The gating network receives the input $\mathbf{x}$, and outputs a scalar value p_i with the property that $p_i \geq 0$ and $\sum_i p_i = 1$. The final prediction of the model is a

sum of the expert predictions weighted by p_i. In this work, all local experts are linear regression models.

2.1.5. Gaussian process regression - GPR

Gaussian process (GP) modelling is a non-parametric method. It does not produce an explicit functional representation of the data, as QSPR modeling does in the form of an equation where the permeability is usually related to statistically significant physicochemical descriptors of a dataset. In GPR modelling it is assumed that the underlying function, $f(\mathbf{x})$, that produces the data will remain unknown, but that the data is produced from a (infinite) set of functions, with a Gaussian distribution in the function space.

A Gaussian process is completely characterised by its mean and covariance function. For simplicity, we usually consider the mean function to be the zero everywhere function. The covariance function, $k(\mathbf{x}_i, \mathbf{x}_j)$, is crucial to GP modelling. It expresses the expected correlation between the values of $f(\mathbf{x})$ at the two points $\mathbf{x}_i, \mathbf{x}_j$. In other words, it defines nearness or similarity between data points.

In this work, we apply the squared exponential covariance function, which incorporates noise into the model, as follows:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\Big(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T M(\mathbf{x}_i - \mathbf{x}_j)\Big) + \sigma_n^2 \delta_{ij} \ , \tag{5}$$

where $M = l^{-2}I$, l is *characteristic length-scale*, σ_f is *signal variance*, σ_n is *noise variance*, and δ_{ij} is the Kronecker delta which is one if $i = j$ and zero otherwise.

To make a prediction y_* at a new input $\mathbf{x}_*$, we need to compute the conditional distribution $p(y_*|y_1, \ldots, y_{N_{trn}})$ on the observed vector $[y_1, \ldots, y_{N_{trn}}]$, where N_{trn} denotes the number of training examples. Since our model is a Gaussian process, this distribution is also a Gaussian and is completely defined by its mean and variance. The mean at $\mathbf{x}_*$ is given by

$$E[y_*] = \mathbf{k}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} \ . \tag{6}$$

In eq(6), $\mathbf{k}_*$ denotes the vector of covariances between the test point and the N_{trn} training data; $\mathbf{K}$ denotes the covariance matrix of the training data; σ_n^2 denotes the variance of an independent identically distributed Gaussian noise, which means observations are noisy; and $\mathbf{y}$ denotes the vector of training targets.

The predictive variance at $\mathbf{x}_*$ is given by

$$var[y_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_*. \tag{7}$$

We use the mean as our prediction and the variance as error bars on the prediction.

2.1.6. GPR with automatic relevance determination - GPRARD

To implement *automatic relevance determination* [20] in GPR, one can redefine the characteristic length-scale matrix M in eq.(5) as a diagonal matrix containing the elements of vector $\mathbf{L} = [l_1^{-2}, \ldots, l_d^{-2}]$, and $l_1, \ldots, l_d$ on the diagonal are the characteristic length scales for each input dimension, determining how relevant an input is to the task. If the length-scale has a very large value, it suggests that the corresponding input could be removed from the inference. These characteristic length-scales can be optimised from the data by Bayesian inference.

2.2. Description of the Dataset Employed

The dataset employed in this study has been collated with reference to a range of literature sources. It predominately consists of the Flynn dataset, used by Potts and Guy, and others. It contains several additions, including those described in [18] and whose origins are described in [17], covering a wide range of molecular properties. The whole dataset consists of 149 compounds. Usually, $\log P$ and MW appear to be the only significant features in QSAR forms. However, in some cases (such as [17]) other features achieve significance; these features are often calculated using expensive and specialist software. Since they often provide only marginal improvements in the prediction of $\log K_p$ compared to other QSAR models, there is little application of them in the field [17].

In this work, five molecular features in total are involved. They are *molecular weight* (MW), *solubility parameter* (SP), $\log P$ (often described, for example by Potts and Guy, as $\log P_{known}$), counts of the number of hydrogen-bonding acceptor (HA) and donor groups (HD), respectively, that can be found on a molecule. These descriptors are described in detail elsewhere ([17] and [19]).

2.2.1. Visualisation of the data

The scatter plot matrix in Figure 1 shows data for all 149 compounds with five features plotted against each other. The diagonal is different in that it shows the shape of the distribution of each feature. The subplot appearing in the first row and last column shows MW against $\log K_p$. It suggests that very similar $\log K_p$ values can correspond to many different MW values. This is also true of $\log P$ (shown as $\log P_{known}$) and $\log K_p$. It can also be seen that there is no simple linear relationship between any pair of descriptors. For example, the correlation coefficient for SP and $\log P$ is -0.32; for SP and HD is 0.21; for SP and HA is 0.30. These correlation coefficients would suggest that there is no linear correlation between these descriptors.

2.2.2. Canonical correlation analysis

Canonical correlation analysis (CCA) [10] can be used to find a projection that maximises the correlation between two sets of variables. In this study, MW, SP, $\log P$, HA and HD were grouped into one set, denoted by $\mathbf{x}$, and $\log K_p$ into another set, denoted by $\mathbf{y}$, in order to investigate the correlating linear relationship between $\log K_p$ and the five compound descriptors. CCA seeks vectors $\mathbf{m}$ and $\mathbf{n}$ so that the correlation between the random variables $\mathbf{m}'\mathbf{x}$ and $\mathbf{n}'\mathbf{y}$ is maximised. The random variables $\mathbf{m}'\mathbf{x}$ and $\mathbf{n}'\mathbf{y}$ are called canonical variables.

The canonical variable 1 ($CV1$) in Figure 2 is a combination of five descriptors used in this work:

$$CV1 \quad = \quad 0.002MW - 0.116SP + 0.033 \log P + 0.107HA + 0.6655HD \ ,$$

while the canonical variable 2 ($CV2$) in Figure 2 is given by $CV2 = -0.686 \log K_p$. Figure 2 demonstrates clearly that there is no linear relationship between the

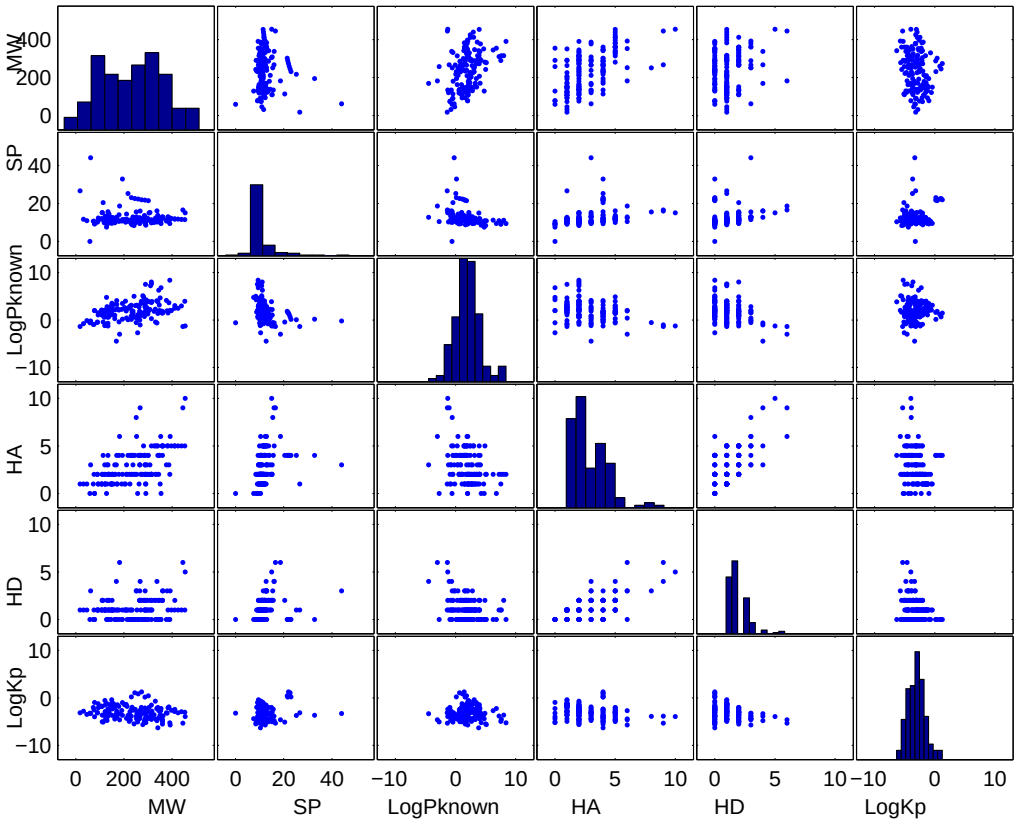

Figure 1: A scatter plot matrix of the skin dataset. The diagonal shows the shape of the distribution of each feature. The graphs in the lower triangle are the transpose of the graphs in the upper triangle.

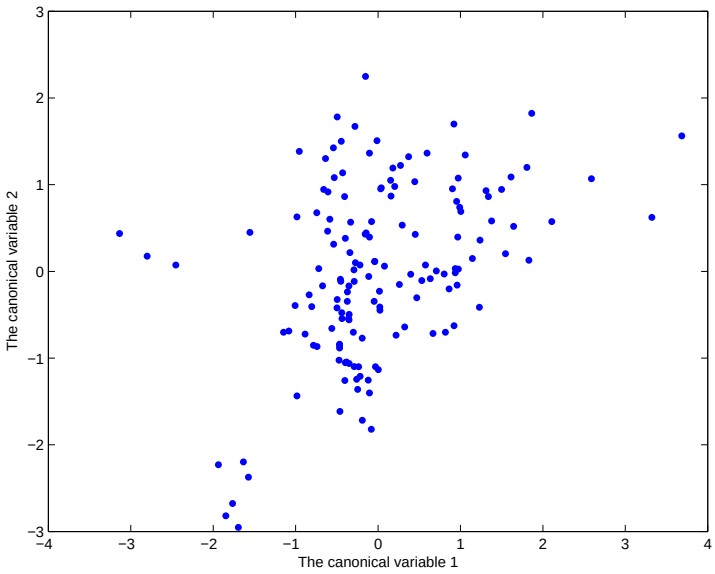

Figure 2: The canonical correlation between five compound descriptors and $\log K_p$.

two sets of variables. It is interesting to note that in $CV1$ the least important features (those with lowest coefficients) are MW and $\log P$. Actually, the *canonical correlation coefficient* is approximately 0.42, while the canonical correlation coefficient between $\log K_p$ and a group of two variables, MW and $\log P$ is about 0.24.

The use of both the above visualisation and the canonical correlation analysis indicates that a non-linear approach to predicting skin permeability is essential, given the inherent nature of the skin dataset being employed.

2.3. Experimental setup

The whole dataset was randomly divided into a training set and an independent test set. There are 130 compounds in the training set, while the test set consists of the remaining 19 compounds. Those modelling methods described in Section 2.1 were applied to the training set to develop predictions on the independent test set using the trained models. This process was repeated ten times, each time for different randomly assigned training and test sets.

To investigate whether predictions can be improved by involving all five

features rather than the original two features used in the QSAR forms, (MW and $\log P$), we employed regression modelling methods with both two and five compound features as an input vector.

In K-nearest-neighbour modelling, we varied the number of neighbours, K, between one and ten; in the mixture of experts, we set the number of experts between two and five. In Gaussian process modelling, we chose the initial values of the logarithm of *characteristic length-scale*, the logarithm of *signal variance*, and *noise variance* using cross validation from ten user defined pre-sets.

We used a five-fold cross-validation procedure to select optimal parameters for each of K-nearest-neighbour, the mixture of experts, and Gaussian process. In these cases, each training set is further divided into training and validation sets five times.

To further investigate which compound descriptors contribute significantly to the prediction, we apply GPRARD (see section 2.1.6) to the data. Again, we undertake experiments on ten randomly selected training and test sets. However, this time the hyperparameters are optimised by maximising the *marginal likelihood* using the derivative rather than selecting from pre-set hyperparameters using a cross validation procedure. More details can be found in [27]. Each time we initialise the logarithm of *characteristic length-scale* for each input dimension, the logarithm of *signal variance*, and *noise variance* as $[0;\ 0;\ 0;\ 0;\ 0;\ 0;\ \log(sqrt(0.1))]$.

We applied Rasmussen and Williams's GP toolbox [27] to do Gaussian process modelling; and employed the Bayes Net Toolbox to carry out the mixture of experts modelling (publicly available at
http://www.cs.ubc.ca/∼murphyk/Software/BNT/bnt.html#ack).

2.4. Influence of descriptors on the model

To explore the effect of particular descriptors on permeability and subsequent predictive models, an analysis of dependence was carried out using the trained GP model and the method reported previously in [21]. Firstly, one of ten trained GP models, using all five descriptors, was randomly selected as the final model to be analysed. Next, six new test sets were constructed. In each of the first five

datasets, one of five descriptors was varied and the other four descriptors were set to the median values in the training set. For the last dataset both $\log P$ and MW were varied and the remaining descriptors were set to their median values. In this study, the six test sets varied MW (range 1 to 600; in increments of 1), $\log P$ (-5 to 9; 0.1), SP (0 to 50; 0.1), HA (0 to 12; 1), HD (0 to 8; 1); other descriptors were set to their median values as described above. Table 1 summarises the statistics of the corresponding training set.

Table 1: : Summary of the training set used.

Descriptor	**Mean**	**Std**	**Min**	**Max**	**Median**
MW	231.5778	103.8051	18.0200	454.4500	236.1700
$\log P$	2.0007	2.1257	-4.4700	8.3900	1.9750
SP	12.3001	4.5697	0	44.0600	11.4300
HA	2.7769	1.8184	0	10.0000	2.0000
HD	1.1615	1.1123	0	6.0000	1.0000

3. Performance measures

Suppose we are given N_{trn} and N_{tst} training and test input-target pairs $(\mathbf{x}_n^{trn}, y_n^{trn})$ and $(\mathbf{x}_n^{tst}, y_n^{tst})$, respectively. Given a test input $\mathbf{x}_n^{tst}$, the model prediction is denoted by $\hat{y}_n$.

3.1. Mean squared error

The *mean squared error* measures the average squared difference between model predictions $\hat{y}_n$ and the corresponding targets y_n^{tst}. Here we report the normalised mean squared error (NMSE) which is shown in the following equation:

$$\text{NMSE} = \frac{1}{N_{tst}} \sum_{n=1}^{N_{tst}} \frac{(y_n^{tst} - \hat{y}_n)^2}{var(y^{trn})}. \tag{8}$$

3.2. Percent improvement over a naive model

In the naive model for any input the prediction is always the same value, namely the mean of $\log K_p$ in the training set, defined by

$$\hat{y}_{naive} = \frac{1}{N_{trn}} \sum_{n=1}^{N_{trn}} y_n^{trn} \ . \tag{9}$$

Thus, the mean squared error of a naive model is given by

$$MSE_{naive} = \frac{1}{N_{tst}} \sum_{n=1}^{N_{tst}} (\hat{y}_{naive} - y_n^{tst})^2 \ . \tag{10}$$

The degree of improvement of the model over the *Naive* predictor can be quantified by the *improvement over Naive* (ION) measure [32]

$$ION = \frac{MSE_{naive} - MSE}{MSE_{naive}} \times 100\% \ . \tag{11}$$

3.3. Negative log loss (NLL)

When we investigate GP's results, we also consider the average *negative log estimated predictive density NLL*, given by

$$NLL = \frac{1}{N_{tst}} \sum_{n=1}^{N_{tst}} (-\log p(y_n^{tst}|\mathbf{x}_n^{tst})) \ , \tag{12}$$

where $-\log p(y_n^{tst}|\mathbf{x}_n^{tst}) = \frac{1}{2}\log(2\pi\sigma_*^2) + \frac{(y_n^{tst} - \hat{y}_n)^2}{2\sigma_*^2}$, in which case σ_*^2 is the predictive variance obtained from eq. (7) plus the noise variance σ_n^2. A small value of NLL shows good performance.

With regard to the performance of our models, and their comparison with previous work [19], the aim of the current study is to obtain a model whose statistical veracity is confirmed where, on the test set, low values of both NMSE and NLL are obtained, as well as high values of both ION and the correlation coefficient (CORR).

4. Experimental results

Prior to the application of modeling methods using the trainable regression models, established methods (eqs. (1) and (3)) used to generate QSPR models were applied to the whole dataset. The results of this analysis are summarized in Table 2, where eq. (1) is denoted as Potts; eq. (3) is denoted as Moss.

Table 2 shows the results using the two QSAR forms discussed in this paper. The results are the averages on the ten independent test sets. For comparison,

Table 2: The results on test sets using different QSAR models.

Models	**NMSE** low better	**ION (%)** high better	**CORR** high better
Naive	$\mathbf{1}.08 \pm \mathbf{0}.13$	**0**	-
Moss	1.46 ± 0.28	-34.71 ± 18.45	0.21 ± 0.21
Potts	5.75 ± 1.14	-430.33 ± 74.38	0.18 ± 0.22

Table 3: The results on test sets using different machine learning methods with only two features.

Models	**NMSE** low better	**ION (%)** high better	**CORR** high better	**NLL** low better
Naive	1.08 ± 0.13	0	-	-
KNN	$\mathbf{0}.87 \pm \mathbf{0}.14$	$\mathbf{1}9.27 \pm \mathbf{6}.74$	$\mathbf{0}.44 \pm \mathbf{0}.15$	-
SLN	1.07 ± 0.17	1.54 ± 4.11	0.21 ± 0.16	-
MIXEXP	1.03 ± 0.14	4.76 ± 6.76	0.28 ± 0.12	-
GPR	0.98 ± 0.11	9.85 ± 5.92	0.32 ± 0.13	3.06 ± 0.48

Table 2 also shows results from the Naive model. In general, all QSAR predictions are less robust than naive predictions, especially with Potts' QSAR form.

Table 3 shows results obtained using computational modelling methods from the machine learning field with MW and $\log P$ as descriptors. One can see that all four methods have improved on the naive predictions, with K-nearest-neighbour giving the best results. The average of the optimal number of neighbours, K, was equal to 8.4. Not surprisingly, the single layer network, which is a simple linear regression model, performed worst. However, it should be noted that the SLN still produced a statistically more robust model than either QSPR model assessed. The mean weights from ten separate runs of SLN with two features, are $0.18(\pm 0.05)$ and $-0.38(\pm 0.03)$ for $\log P$ and MW. The bias in each run is almost zero. This shows that the SLN gives more weights to MW than $\log P$ compared with eqs. (1) and (3).

Results obtained with five compound descriptors are shown in Table 4. Comparing with Table 3, one can see that all four regression modelling methods have improved their performance when using five features rather than two. This would suggest the importance of these terms - solubility parameter and

Table 4: The results on test sets using different machine learning methods with all five features.

Models	**NMSE** low better	**ION (%)** high better	**CORR** high better	**NLL** low better
KNN	0.63 ± 0.11	42.04 ± 7.59	0.67 ± 0.08	-
SLN	0.87 ± 0.12	20.12 ± 5.37	0.50 ± 0.12	-
MIXEXP	0.44 ± 0.24	59.28 ± 22.03	0.81 ± 0.10	-
GPR	$\mathbf{0}.30 \pm \mathbf{0}.07$	$\mathbf{7}2.62 \pm \mathbf{5}.03$	$\mathbf{0}.86 \pm \mathbf{0}.05$	$\mathbf{1}.48 \pm \mathbf{0}.13$
GPRARD	0.30 ± 0.06	71.99 ± 5.53	0.86 ± 0.04	1.50 ± 0.21

hydrogen-bonding descriptors - in the prediction of percutaneous absorption. Comparison of SLN results shown in Tables 3 and 4 indicates improvements in all performance metrics when five descriptors are used instead of two. This suggests that the use of five descriptors can potentially improve predictions, even on a linear model of this type.

Of the models summarised in Table 4 the Gaussian process regression and its modified form, GPRARD, give the best performance. There is no significant difference between the results obtained from these two methods. Figure 3 displays a box plot of normalised mean squared errors from ten independent test sets on the Naive model, the Moss QSAR form, and those four computational modelling methods with five features. It shows that the Gaussian process regression with five features (GPRf5) gives the lowest upper quartile, median and lower quartile values on NMSE. Although the mixture of experts with five features (MIXEXPf5) has comparable low median and lower quartile values, its upper quartile value and the largest NMSE value are much bigger than those obtained from GPRf5. It suggests that GPRf5 has a relatively stable and robust performance. On the other hand, one can see the QSAR form (Moss) has the highest lower quartile, median and upper quartile values. Both K-nearest-neighbour with five features (KNNf5) and single layer network with five features (SLNf5) were relatively stable, but in general not as good as GPRf5 and MIXEXPf5.

Each length-scale in GPRARD for the corresponding individual compound descriptors is shown in Table 5. It shows that all five descriptors have a similar length-scale, with HD having the shortest length-scale. Since length scale is inversely related to the relevance of the descriptor, this suggests that

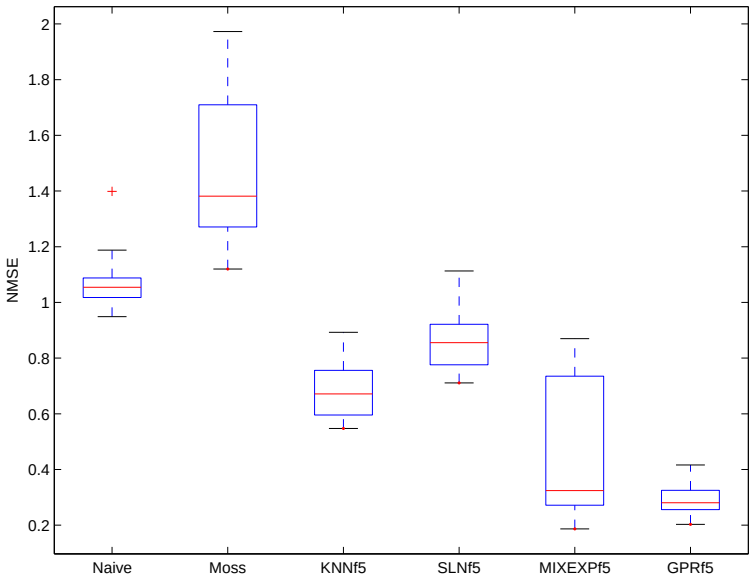

Figure 3: Box plot of normalised mean squared errors from ten independent test sets on six different models with five features.

Table 5: Lengthscales with five features.

	MW	**SP**	$\log$ **P**	**HA**	**HD**
lengthscale	0.79 ± 0.10	1.04 ± 0.29	0.99 ± 0.12	2.02 ± 4.23	0.55 ± 0.20

all inputs are fairly equally relevant to the task. However, HA gives a relatively bigger mean length-scale with a large standard deviation. One outlier is with HA, where the results on the test set with the trained model are $NMSE = 0.43, ION(\%) = 58.14, CORR = 0.81,$ and $NLL = 2.01$. Comparing this results with the last row in Table 4, it can be seen that all these performance measurements are worse than the mean values. This suggests that in this particular case the trained GPRARD model did not capture the underlying distribution very well.

The dependency of molecular descriptors on skin permeability is shown in Figures 4-9, where each of the descriptors is plotted separately. Figure 9 shows the effect of both $\log P$ and MW on $\log K_p$. Variables not shown in a particular plot were set to their median values. The central line represents the prediction, and the outlying lines the 95% confidence intervals. It can be seen from

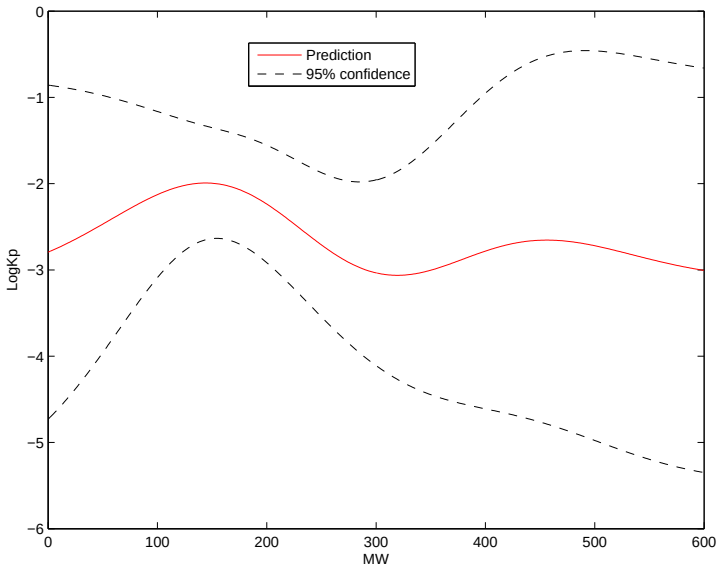

Figure 4: Dependency of permeability $\log K_p$ on MW for the final model.

this model that permeability increases with MW $(0-150)$, decreases thereafter $(150-320)$ and then increases slightly again (Figure 4). This last increase may be an artifact of the Gaussian process due to the small number of data points present in this part of the plot (and the associated increase in variance at such points in the plot), or it may indicate a particular effect, such as ionisation, on the data. The relationship between $\log P$ and $\log K_p$ (Figure 5) is not linear and a bell-shaped distribution is observed in the data (see Figure 1). A similar trend is observed between $\log K_p$ and SP (Figure 6). The permeability coefficient decreases from 7 to 15 and increases thereafter, falling away at around 30. This matches the SP associated with the stratum corneum, and suggests that permeability is at its lowest where it reflects the solubility in the stratum corneum best, suggesting a bimodal inverse relationship between SP and $\log K_p$.

Figures 7 and 8 show the influence of HA and HD on $\log K_p$. Figure 9 is an insight into why the original linear regression models perform poorly across the full range of $\log P$ values. The contour plot clearly shows the relationship between MW, $\log P$ and $\log K_p$ is highly non-linear.

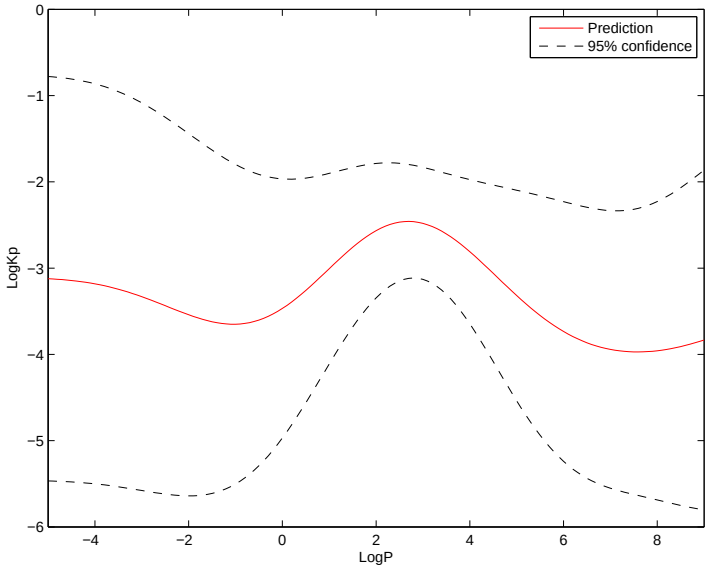

Figure 5: Dependency of permeability $\log K_p$ on $\log P$ for the final model.

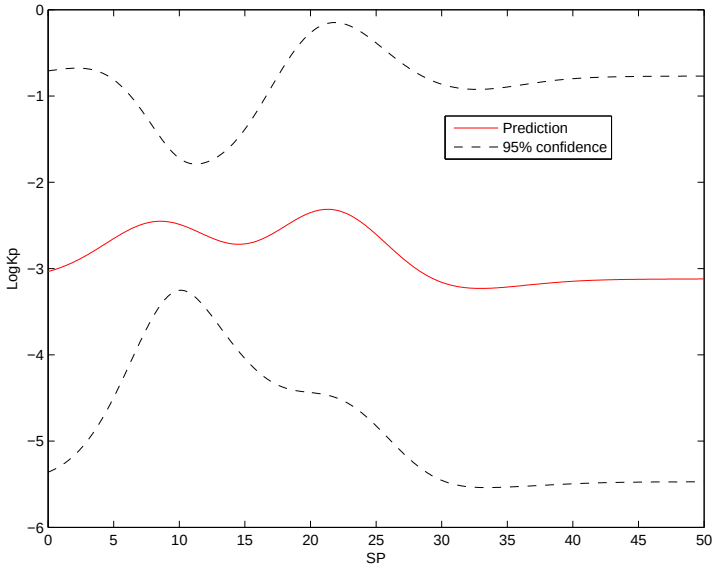

Figure 6: Dependency of permeability $\log K_p$ on SP for the final model.

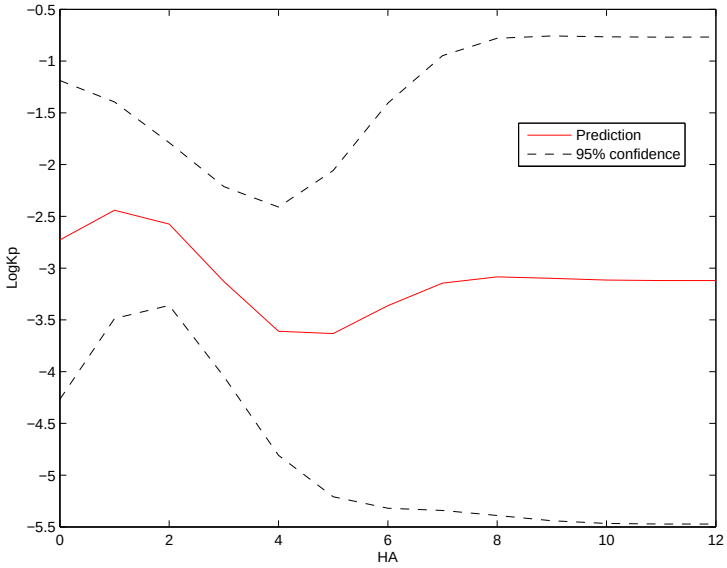

Figure 7: Dependency of permeability $\log K_p$ on HA for the final model.

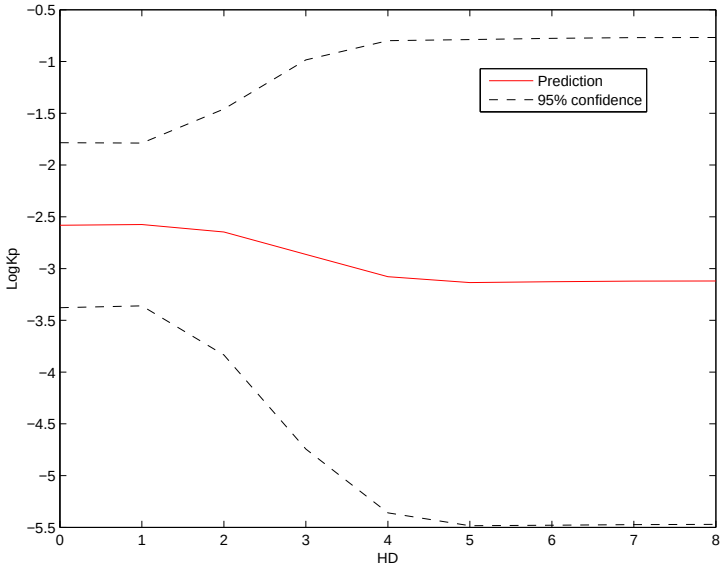

Figure 8: Dependency of permeability $\log K_p$ on HD for the final model.

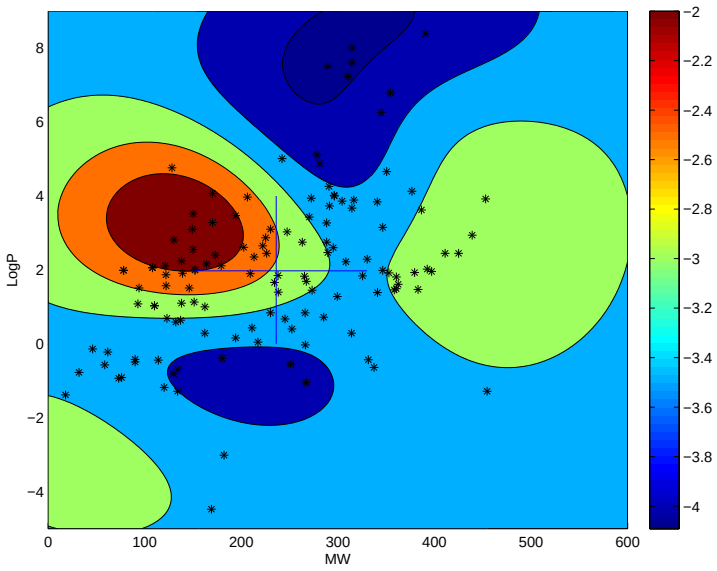

Figure 9: Dependency of permeability $\log K_p$ on MW and $\log P$ for the final model. Asterisks represent training data points; lines mark the median values of the descriptors shown in the plot.

5. Discussions

Most methods which have yielded models of percutaneous absorption have involved the use of quantitative structure-permeability relationships (QSPRs). While some of these models (i.e. those derived in [33]) used non-linear methods, the vast majority of models employed linear representations of the data. This field has been reviewed extensively, and the reader is directed to detailed reviews of this subject ([9] and [17]). An advantage of the models derived (those of the "Potts and Guy" form, where permeability is commonly seen to be a function of $\log P$ and molecular weight) is the ease of use, as the relevant molecular parameters can be easily determined - in the case of $\log P$, this is either carried out in a laboratory or by computational methods. However, one of the criticisms made against any model that does not conform to this type is that it is difficult to use and, if complex mathematics or molecular descriptors are involved, the models will have limited applicability to an audience who may not have access to the costly software often required to develop such models.

The use of Gaussian processes takes this to an extreme, as it does not directly result in a quantitative, descriptive output (such as a "Potts and Guy"-type equation) that may be interpreted appropriately by those interested in percutaneous absorption. However, the use of related methods, including length-scale analysis [14], provides additional details of the importance of particular molecular descriptors. Moss et al., [19] explored the viability of the GP approach for modelling skin absorption data, and demonstrated its statistical superiority over a series of other models. In particular, the QSPR-type models (specifically, those by [25] and [16]) were shown to be significantly worse, in terms of their descriptive statistics, than single layer networks or Gaussian processes. This perhaps reflects the nature of the dataset employed in QSPR studies, which was derived from [7]. This dataset - a substantially expanded form of which is used herein - is predominately comprised of data points at the lower end of the scale, in terms of physicochemical descriptors. Moss et al [19] likened this to the up-slope of a Gaussian distribution curve, which may explain why such statistically acceptable models were developed from this dataset. However, in expanding this dataset, particularly with molecules that are predominately lipophilic (i.e. those which may reside on the down-slope of a Gaussian distribution), Moss et al [19] were able to develop a model of percutaneous absorption that not only modelled better, in a statistical sense, but which fitted empirical and experimental observations of skin absorption, which is not considered to be a linear process in the context of the molecular descriptors. This is due to the nature of the stratum corneum skin barrier and its interaction of exogenous chemicals.

Classical QSPR-based models of percutaneous absorption output a defined mathematical relationship between permeability (as K_p or J). This provides mechanistic information regarding the significant physicochemical descriptors of a molecule that influence its percutaneous absorption. It is also transparent, allowing a wide range of users to apply the model for their needs. Clearly, the GP model does not allow the same breadth of use due to its "black box" approach. However, this method does offer a different approach to the issue

of modelling percutaneous absorption. One must consider that the use of such models extends beyond merely being a tool for researchers to estimate the permeability of their novel compounds. Clearly, as the vast body of work in this field, some of which is cited herein, demonstrates, these models offer a deep and quite specific mechanistic understanding of percutaneous absorption. While usage of, in particular, the Potts and Guy (1992) equation [25] is common, this body of research has provided detailed and invaluable information on the mechanism of percutaneous absorption. We feel therefore that the GP approach, while currently limited by its "black box" approach compared to QSPR-based models, offers significant advantages over the previously employed methods, as highlighted in the previous section. It should also be noted that the use of GP methods is a novel approach to the problem of modelling skin absorption. Work of this type, and using such methods, has only begun to be published ([14], [19]) in the field of percutaneous absorption.

The present study expands the concept of non-linear modelling of skin absorption. Figure 1 shows a visualisation of the dataset and its inherently Gaussian distribution. Figure 2 shows the results from canonical correlation analysis, which demonstrates clearly the lack of a linear relationship between the variables. Clearly, the use of these methods show that the inherent nature of the dataset is non-linear, suggesting that non-linear methods of analysis would be the most appropriate in accurately predicting skin absorption.

In the prediction of percutaneous absorption, both the method used to derive a model, and the physicochemical descriptors associated with such models, have varied significantly despite the perception of the applicability of the generic algorithm associated with Potts and Guy's (1992) work. Indeed, Potts and Guy subsequently re-analysed the dataset associated with their initial work [26] and found that, for a subset of the dataset, hydrogen-bonding was an important descriptor for permeability for a specific class of molecules. Other researchers have explored the importance of a range of molecular descriptors to percutaneous absorption.

In studies such as this the nature of the dataset can play a key role in the

nature of the resulting model. For example, a comparison of the Potts and Guy studies ([25], [26]) indicates very clear difference in the output based on the nature of the dataset used. Similarly, the study by Moss and Cronin [18] saw the removal of the steroid data from the Flynn ([7]) dataset (the Scheuplein data ([31])) and the substitution of additional data that had been collated by others ([12]). This saw, for example, the inclusion of eight values for estradiol, where only one had been included in the Flynn dataset ([7]). When the model was recomputed an equation very similar to the Potts and Guy ([25]) equation resulted. However, the Moss and Cronin ([16]) equation importantly found that steroids were no longer listed as outliers due to the re-modelling.

This is an important point in considering the nature of the dataset used in this study. For example, it infers that simply increasing the number of chemicals in the dataset may have little or no effect on the quality of the resulting model and that the distribution of the data (shown, for example, in Figures (1) and (2)) is of greater significance in terms of representative modelling of percutaneous absorption. It should also be noted that the dataset employed in this study is one of the largest used in any study modelling skin absorption.

While absent from the widely accepted Potts and Guy model [25], hydrogen-bonding has been considered as a key influence in percutaneous absorption for just over thirty years ([28]). Development of the solvatochromic theory in explaining partition phenomena ([13]) and epidermal permeability ([1], [29]) suggested that, among other physicochemical properties, both hydrogen-bonding acceptor and donor properties of a molecule play key roles in determining penetrant permeation.

Roberts et al., [30] showed that the introduction of even one hydrogen-bonding group to a molecule resulted in a substantial decreases in permeability. Addition of further groups resulted in further decreases, which were non-linear. In general, they found that acids seemed to diffuse more slowly than alcohols or phenols, and suggested that hydrogen-bonding was the key factor in diffusion across the stratum corneum, whereas lipophilicity (i.e. $\log P$) was more important for partitioning. This phenomenon may be related to the acidity constant,

pKa, of the penetrant and its ionisation state, and suggests that ionisation may have a substantial role in understanding how hydrogen-bonding influences skin absorption. The results in the current study, particularly those shown in Figure 7, would also suggest that the introduction of a hydrogen-bonding group onto a potential penetrant exerts a significant influence on permeability. Indeed, the trend shown in this figure follows closely the argument used by Roberts and co-workers in their study.

The role of hydrogen-bonding in skin absorption has also been explored by other authors (i.e. [29], [24], [22]). While it is difficult to directly compare such studies to other approaches (specifically, those used to develop "Potts and Guy"-type models of skin permeation) due to differences of dataset composition and mathematical approaches, it may be argued that the use of methods that do not properly consider the nature of whichever dataset is used undermines the veracity of any resultant model.

While Moss et al. [19] compared the statistical accuracy of Gaussian processes, single linear networks and QSPRs, they did not explore in detail the effect of particular physicochemical descriptors on the resultant models. This is explored in the current study, where models developed with five molecular descriptors ($\log P$, MW, HA, HD, SP) performed significantly better than those developed with two descriptors ($\log P$ and MW). In addition, Table 5 summarises the length-scales from GPRARD analysis for each of individual physicochemical descriptors. It shows that, with the exception of HA, all parameters contribute relatively equally to the development of the predictive model. The length-scale for HA is higher than for the other descriptors, but this value is swamped by a very large standard deviation. This might suggest an error associated with the ionisation state of a chemical and may indicate the importance of normalising the K_p values from the dataset to percentage ionisation state. While this is not a straightforward task, and one which may skew other parameters (due to considerations of, for example, solubility and the effect on $\log P$) it may provide an understanding of the mechanistic importance of hydrogen-bonding and ionisation in percutaneous transport.

It also demonstrates the difficulty of separating a group of such inter-dependant descriptors and yielding specific mechanistic information that relates to a specific molecular functionality. For example, while du Plessis et al. [22] indicated that hydrogen-bonding was important for permeability, they were unable to fully decouple any such effects from other parameters, including molecular symmetry and the substitution of the molecules in their dataset, and suggested that this may be due to the similarity of lipophilic molecular features in their data. Further, while Magnusson et al. [15] indicated that molecular weight was the main parameter for predicting flux (a term related to permeability, K_p, and the concentration of a permeant) across skin, they also suggested that melting point and HA are also of significance.

The main focus of this study is in developing and validating the use of GP methods, and also in showing how they compare to existing models. This study, in common with a large number of other studies, focuses on K_p. While more recent studies, notably Magnusson et al. [15] use flux (as J_{max}) we have focused on K_p in this study in order to allow ready comparisons with classical studies in this field, such as Potts and Guy [25]. This is an important aspect of validating the novel GP method and comparing it directly with existing benchmarks.

Recently, other novel methods have been employed in this field. For example, Fransch [8] used a 4-parameter algebraic model to examine percutaneous absorption. This differs significantly from the work reported herein, which is fundamentally different to Fransch's approach in that it is a statistical-based approach to modelling. In addition, it should be noted that Frasch uses a model based on the structural organisation of mouse stratum corneum, including the covering of the upper and lower surfaces of the stratum corneum with a lipid film. The work in the current manuscript makes no such assumptions.

6. Conclusions

The results presented herein suggest substantial limitations to current QSPR-type models, both in terms of the significance of the descriptors used and the manner in which the data is interpreted and analysed. They indicate that, in

terms of statistical performance, the following rank order is observed: GP > Mixture of Experts > SLN > QSPR. The distribution of the dataset has been shown in this study to be non-linear, and that increasing the number of descriptors improves the model significantly. It should be noted that the dataset used herein is different from those used to produce QSAR-type models, as it contains more lipophilic members. Further, analysis of the descriptors used suggests that they are all of similar weighting and all contribute to the models produced. This is consistent with previous observations reported in the literature, where hydrogen-bonding in particular is an important factor in skin permeability. As shown in Figure 9, the results should be treated with caution due to the limitations of the dataset, and any interpretation should be made with this, and an underlying knowledge of the nature of the Gaussian process, in mind.

References

[1] M. Abraham, H. Chadha, and R. Mitchell, (1995) The factors that influence skin penetration of solutes. *Journal of Pharmacy and Pharmacology*, 47, pp.8-16.

[2] C. M. Bishop, (1995) *Neural Networks for Pattern Recognition*. Oxford University Press, New York.

[3] M. T. D. Cronin, J. C. Dearden, G. P. Moss, and G. Murray-Dickson, (1999) Investigation of the mechanism of flux across human skin in vitro by quantitative structure-permeability relationships. *European Journal of Pharmacy and Pharmacology*, 7, pp. 325-330.

[4] I. Tuncer Degim, (2006) New tools and approaches for predicting skin permeability, *Drug Discovery Today*. Vol 11, pp.517-523.

[5] P. M. Elias, (1983) Epidermal lipids, barrier function and desquamation. *J. Invest. Dermatol*. Vol 80, pp.44-50.

[6] D. Fitzpatrick, J. Corish and B. Hayes, (2004) Modelling skin permeability in risk assessment - the future, *Chemosphere*. Vol 55, pp.1309-1314.

[7] G. L. Flynn, (1990) Physicochemical determinants of skin absorption. In *Principles of Route-to-Route Extrapolation for Risk Assessment*, T. R. Gerrity and C. J. Henry (eds.), Elseviet, New Kork, 1990, pp.93-127.

[8] H. F. Frasch, (2002) A Random Walk model of skin permeation. *Risk Analysis*, 22, pp.265-276.

[9] S. Geinoz, R. H. Guy, B. Testa, and P. A. Carrupt, (2004) Quantitative structure-permeation relationships (QSPeRs) to predict skin permeation: a critical evaluation. *Pharmaceutical Research*, Vol 21 pp.83-92.

[10] H. Hotelling, (1936) Relations between two sets of variates. *Biometrika*, Vol 28 pp.312-377.

[11] R. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, (1991) Adaptive mixtures of local experts. Neural Computation, 3, pp.79-87.

[12] M. E. Johnson, D. Blankschtein, R. Langer, (1995) Permeation of steroids through human skin. *Journal of Pharmaceutical Sciences*, 84, pp.1144-1146.

[13] M. J. Kamlet, J. L. Abboud, M. H. Abraham, and R. W. Taft, (1983) Linear Solvation Energy Relationships. 23. A comprehensive collection of the solvatochromic parameters, π^*, α, and β, and some methods for simplifying the generalized solvatochromic equation. *Journal of Organic Chemistry*, 48, pp.2877-2887.

[14] L. T. Lam, Y. Sun, N. Davey, R. Adams, M. Prapopoulou, M. B. Brown, and G. P. Moss, (2010) The application of feature selection to the development of Gaussian process models for percutaneous absorption. *Journal of Pharmacy and Pharmacology*, 62, pp.738-749.

[15] B. M. Magnusson, Y. G. Anissimov, S. E. Cross, and M. S. Roberts, (2004) Molecular size as the main determinant of solute maximum flux across the skin. *Journal of Investigative Dermatology*, 122, pp.993-999.

[16] G. P. Moss and M. T. D. Cronin, (2002) Quantitative structure-permeatibility relationships (QSPRs) for percutaneous absorption: re-analysis of steroid data. *International Journal of Pharmaceutics*, Vol. 238, pp.105-109.

[17] G. P. Moss, J. C. Dearden, H. Patel, and M. T. D. Cronin, (2002) Quantitative structure-permeatibility relationships (QSPRs) for percutaneous absorption. *Toxicology in Vitro*, Vol. 16, pp299-317.

[18] G. P. Moss, D. R. Gullick, P. A. Cox, C. Alexander, M. J. Ingram, J. D. Smart and W. J. Pugh, (2006) Design, synthesis and characterization of captopril produgs for enhanced percutaneous absorption. *Journal of Pharmacy and Pharmacology*, Vol 58, pp.167-177.

[19] G. P. Moss, Y. Sun, M. Prapopoulou, N. Davey, R. Adams, W. J. Pugh and M. B. Brown, (2009) The application of Gaussian processes in the prediction of percutaneous absorption. *Journal of Pharmacy & Pharmacology*, Vol 61, pp.1147-1153.

[20] R. M. Neal (1996) *Bayesian Learning for Neural Networks*. Springer, New York. Lecture Notes in Statistics 118.

[21] D. Neumann, O. Kohlbacher, C. Merkwirth, and T. Lengauer, (2006) A fully computational model for predicting percutaneous drug absorption, *J. Chem. Inf. Model.* 46, pp.424-429.

[22] J. du Plessis, W. J. Pugh, A. Judefeindb, and J. Hadgraft, (2002) Physicochemical determinants of dermal drug delivery: effects of the number and substitution pattern of polar groups. *European Journal of Pharmaceutical Sciences*, 16, pp.107-112.

[23] W. J. Pugh, and J. Hadgraft, (1994) Ab initio prediction of human skin permeability coefficients. *International Journal of Pharmaceutics*, 103, pp.163-178.

[24] W. J. Pugh, M. Roberts, and J. Hadgraft, (1996) Epidermal permeability-penetrant structure relationships: 3. The effect of hydrogen bonding interactions and molecular size on diffusion across the stratum corneum. *International Journal of Pharmaceutics*, 138, pp.149-165.

[25] R. O. Potts and R. H. Guy, (1992) Predicting skin permeability, *Pharm. Res.* vol(12), pp.663-669.

[26] R. O. Potts and R. H. Guy, (1995) A predictive algorithm for skin permeability: the effects of molecular size and hydrogen bond activity. *Pharmaceutical Research*, vol(12) pp.1628 - 1633.

[27] C. E. Rasmussen and C. K. I. Williams, (2006) *Gaussian Processes for Machine Learning*. The MIT Press.

[28] M. Roberts, (1976) *Percutaneous absorption of phenolic compounds*; PhD Thesis, University of Sydney, Sydney.

[29] M. Roberts, W. J. Pugh, J. Hadgraft, and A. Watkinson, (1995) Epidermal permeability-penetrant structure relationships: 1. An analysis of methods of predicting penetration of monofunctional solutes from aqueous solutions. *International Journal of Pharmaceutics*, 126, pp.219-233.

[30] M. Roberts, W. J. Pugh, and J. Hadgraft, (1996) Epidermal permeability: Penetrant structure relationships. 2. The effect of H-bonding groups in penetrants on their diffusion through the stratum corneum. International Journal of Pharmaceutics, 132, pp.23-32.

[31] R. J. Scheuplein, and I. H. Blank, (1971) Permeability of the skin. *Physiological Reviews*, 51, pp.702-747.

[32] P. Tino, I. Nabney, B. S. Williams, J. Losel, and Y. Sun, (2004) Nonlinear Prediction of Quantitative Structure-Activity Relationships. *Journal of Chemical Information and Computer Sciences*, 44(5), pp.1647-1653. (c) ACM

[33] A. Wilschut, W. F. ten Berge, P. J. Robinson, and T. E. McKone, (1995) Estimating skin permeation - the validation of 5 mathematical skin permeation models. *Chemosphere* 30, pp.1275-1296.