THE APPLICATION OF FEATURE SELECTION TO THE DEVELOPMENT OF GAUSSIAN PROCESS MODELS FOR PERCUTANEOUS ABSORPTION

Lun Tak Lam[1], Yi Sun[2], Neil Davey[2], Rod Adams[2], Maria Prapopoulou[1], Marc B. Brown[1], Gary P. Moss[3*]

[1]The School of Pharmacy and [2]The School of Engineering & Information Sciences, University of Hertfordshire, College Lane, Hertfordshire, AL10 9AB, UK.

[3]The School of Pharmacy, Keele University, Staffordshire, ST5 5BG, UK.

Corresponding Author:

Gary Moss

The School of Pharmacy

Keele University

Staffordshire

ST5 5BG

UK

Tel: 01782 733 833

Email: gpjmoss@yahoo.co.uk

Keywords:

Percutaneous absorption

Gaussian Process

Machine Learning Methods

Quantitative Structure-Permeability Relationships

ABSTRACT

Objectives

The aim was to employ Gaussian Processes to assess mathematically the nature of a skin permeability dataset and to employ these methods, particularly feature selection, to determine the key physicochemical descriptors which exert the most significant influence on percutaneous absorption, and to compare such models to established existing models.

Methods

Gaussian Processes (GPR), including Automatic Relevance Detection (GPRARD) methods, were employed to develop models of percutaneous absorption that identified key physicochemical descriptors of percutaneous absorption. Using MatLab software, the statistical performance of these models were compared to single linear networks (SLN) and quantitative structure–permeability relationships (QSPRs). Feature selection methods were used to examine in more detail the physicochemical parameters used in this study. A range of statistical measures to determine model quality.

Key findings

The inherently non-linear nature of the skin data set was confirmed. The GPR methods yielded predictive models that offered statistically significant improvements over SLN and QSPR models with regard to predictivity (where the rank order was: GPR > SLN > QSPR). Feature selection analysis determined that the best GPR models were those that contained log P, melting point and the number of hydrogen bond donor groups as significant descriptors. Further statistical analysis also found that great synergy existed between certain parameters. It further suggested that a number of the descriptors employed were effectively interchangeable, thus questioning the use of models where discrete variables are output, usually in the form of an equation.

Conclusions

The use of a non-linear GPR method produced models with significantly improved predictivity, compared to SLN or QSPR models. Feature selection methods were able to provide important mechanistic information. However, it was also shown that significant synergy existed between

certain parameters, and as such it was possible to interchange certain descriptors (i.e. molecular weight and melting point) without incurring a loss of model quality. Such synergy suggests that a model constructed from discrete terms in an equation may not be the most appropriate way of representing mechanistic understandings of skin absorption.

INTRODUCTION

The prediction of skin absorption is of interest to many fields, including topical and transdermal drug delivery, cosmetics and risk assessment for dermal exposure. The development of viable, quantitative models has been an area of substantial interest for almost twenty years, and offers considerable advantages in reducing or replacing time-consuming and costly experiments. It is known that the physicochemical properties of a molecule exert a substantial effect on its permeability, and as such most predictive methods have relied on a qualitative or quantitative appraisal of such properties, usually as discrete entities within a mathematical representation of permeation, in order to understand the mechanisms of absorption and to allow prediction of the penetration of a range of exogenous chemicals. In particular, the effects of lipophilicity (most commonly expressed as log P, the octanol-water partition coefficient), hydrogen bonding, molecular weight (or size) and melting point were considered highly significant in their influence, and therefore predicting in permeability (Scheuplein & Blank, 1971; Michaels et al., 1975). Subsequently, several researchers determined that molecular size was more significant than previously suggested (i.e. Potts & Guy, 1992; Magnusson et al., 2004).

It is interesting to consider the nature of descriptors returned by different analyses of datasets. This is clearly highlighted by Potts and Guy (1992; 1995). In these two studies the authors determined that the relationship between K_p and physicochemical descriptors differed as the nature of the dataset (from Flynn, 1990) was, in the later study, qualitatively examined and abbreviated to 37 compounds. This subset was shown to be dependant on lipophilicity and hydrogen-bonding,

whereas an analysis of the whole dataset (Potts and Guy, 1992) demonstrated that lipophilicity and molecular weight were the key determinants in percutaneous absorption.

Hydrogen-bonding, despite being absent from the seminal Potts and Guy (1992) model, has been considered as a key influence in percutaneous absorption for just over thirty years (Roberts, 1976). Partition phenomena, and in particular the development of the solvatochromic theory (Kamlet et al., 1983) and developments in the understanding of epidermal permeability (Abraham et al., 1995; Roberts et al., 1995) indicated the importance of hydrogen-bonding acceptor and donor properties in percutaneous absorption.

Roberts et al. (1996) showed that the introduction of even one hydrogen-bonding group to a molecule could result in a significant decrease in its permeability, whereas the addition of further groups to the molecule results in further, smaller, non-linear decreases. They concluded that hydrogen-bonding was the key factor in diffusion across the *stratum corneum*, whereas lipophilicity was more important for partitioning and may be related to the pK_a of the penetrant.

While it is difficult to directly compare the studies discussed above with other approaches (due to, for example, differences in dataset composition or statistical methods of analysis), it may be argued that the use of methods that do not properly consider the nature of the dataset used undermines any resultant model. Moss et al. (2009) compared the statistical accuracy of Gaussian Processes, single linear networks and QSPRs by a range of statistical methods, and found that the nature of the dataset was inherently non-linear and that skin permeation (as represented by K_p) was best described, in purely statistical terms, by Gaussian Process approaches.

As this field expanded a large number of studies presented a diverse range of models based on an array of different, often complementary, datasets, and an increasing number of physicochemical properties, including hydrogen bonding and molecular size, were presented (i.e. Abraham et al., 1995; Potts & Guy, 1995; Pugh et al., 1996; Roberts et al., 1996; Cronin et al., 1999; Patel et al., 2002). Various modifications have been made to these models, some of which involve the use of non-linear modelling. For example, Wilschut et al. (1995) examined five mathematical models by non-linear multiple regression. The octanol-water partition coefficient and molecular weight were used as independent parameters. They suggested that a modified form of the Potts and Guy (1992) equation best modelled skin absorption. Finally, in order to understand the scope, limitations and context of these models, and how they should be applied, it must be emphasised that they are all based on infinite doses being delivered from aqueous vehicles.

Therefore, while non-linear modelling of skin absorption is not new, it is certainly an area which has not been extensively or systematically explored. The aim of this study is to further compare the statistical accuracy and predictive ability of linear and non-linear methods of modelling, and to also explore combinations of molecular descriptors that may influence, individually or synergistically, percutaneous absorption.

METHODS

Dataset

The dataset employed in this study was obtained from Moss et al (2009). It is, briefly, a dataset that contains 142 different chemicals and their associated physicochemical descriptors and permeability values (K_p, as cm/h), and is an extension of that published by Flynn (1990) and utilised in the study by Potts and Guy (1992). It is supplemented by the addition of data from previous publications (Moss et al., 2006; Patel et al., 2002; Wilschut et al., 1995) and from the Edetox database (available at www.ncl.ac.uk/edetox/index.html). It includes the data, obtained from the literature, for six physicochemical descriptors of each compound; namely, molecular weight (MW), melting point (MPt), solubility parameter (SP) (Fedors, 1974), the octanol-water partition coefficient (log P, used as provided in the sources listed above), hydrogen bonding acceptor groups (HA) and donor groups (HD).

Mathematical methods for model development

The mathematical methods employed herein have been described in detail elsewhere (Rasmussen and Williams, 2006; Moss et al., 2009). The modelling in this study was carried out by a combination of machine learning methods and quantitative structure–permeability relationships (QSPRs). The QSPRs employed are those by Potts and Guy (1992), Cronin et al. (1999), Moss and Cronin (2002) and Luo et al. (2007).

Machine learning methods include Single Layer Networks (SLN), which is a simple linear regression – it is the same as a linear regression method and uses iterated re-weighted least squares

training – and Gaussian Process Regression (GPR), which is a regression that calculates the relationship between variables via a non-linear processes. Further, Gaussian Process Regression with Automatic Resonance Detection (GPRARD) has been employed to calculate the relative significance of the molecular descriptors in GPR modelling (Moss et al., 2009). Performance measures of GPR, SLN, GPRARD and QSPRs were calculated by via Matlab (R2008a). This programme relies on tailored scripts (essentially, a series of commands that allow Matlab to process the required calculations) to conduct calculations for the specific tests used in this study and in previous studies (i.e. Moss et al., 2009). The scripts used were analysis by SLN/QSPR, GPR/GPRARD, GPR (improvement over the naïve model (ION), with statistical significance determined by a paired t-test), one script for GPR (normalised mean squared error (NMSE), paired t-test), GPR (correlation coefficient (CORR), r, paired t-test), SLN (ION paired t-test), SLN (NMSE paired t-test) and for SLN (CORR paired t-test). Matlab was also used to perform statistical analysis of performance measures between SLN and GPR. Statistical comparisons between QSPR and maching learning methods (GPR and SLN) was performed using SPSS® (version 16).

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 (Potts & Guy, 1992; Cronin et al., 1999; Moss et al., 2002). Further details on the nature of these models may also be found elsewhere (Cronin et al., 1999; Moss et al., 2002).

Machine Learning Methods

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, x, which can be written as follows:

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

where d is the dimensionality of the input space (i.e. the number of features used to describe a molecule) and w = (w_1; : : : ;w_d;w_0) is the weight vector. The weights are set so that the sum squared error function is minimized on a training set.

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 modeling it is assumed that the underlying function that produces the data, f(x), will remain unknown, but that the data are produced from a (infinite) set of functions, with a Gaussian distribution in the function space. This has been described in detail elsewhere (Moss et al., 2009; Rasmussen & Williams, 2006). Briefly, a Gaussian process is completely characterised by its mean and covariance function. The mean function is normally considered to be the "zero everywhere" function. The covariance function, k(x_i, x_j), is crucial to GP modeling as it expresses the expected correlation between the values of f(x) at the two points x_i, x_j. In other words, it defines nearness or similarity between data points. Since the model employed herein is a Gaussian process, this

distribution is also Gaussian and is therefore fully defined by its mean and variance. The mean at x_* is given by:

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

where k_* denotes the vector of covariances between the test point and N_{trn} training data; K denotes the covariance matrix of the training data; σ_n^2 is the variance of an independent identically distributed Gaussian noise, which means that observations are noisy, K_*^T is the transpose of K_*, and I is the identity matrix; finally, y denotes the vector of training targets. The variance, at x_*, is given by:

$$var[y_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_* \tag{3}$$

where $k(x_*,x_*)$ denotes the variance of y_*. In the present study, the mean is used as the prediction and the variance as error bars on the prediction.

GPR with automatic relevance determination (GPRARD)

To implement automatic relevance determination (Neal, 1996) in GPR, the characteristic length-scale matrix, M, is redefined as a diagonal matrix containing the elements of vector L =[l_1^{-2},…,l_D^{-2}], and l_1,…,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.

Feature Selection

The features, or molecular descriptors, most frequently used in studies of modelling percutaneous absorption were employed in this study. Parameters were also used that were readily accessible and calculable without the need for expensive, specialist software (Moss et al., 2002; 2009). The features utilised in this study are listed above.

Analysis of the dataset

Data was visualised by scatter diagrams plotted with Microsoft Excel 2007, in order to discern patterns between the features. Such visualisation has been shown previously (Moss et al., 2009).

The dataset was divided, for machine learning method development, into a training set and a test in the ratio of 75% (107 compounds) and 25% (35 compounds) respectively (Katritzky et al., 2006). The compounds were randomly allocated into the subsets automatically by Matlab R2008a via *primeSeed* code, which acts as a recorder to document the allocation of the compounds in the subsets. In total, the experiment was repeated 10 times, generating 10 different test sets. Each test set contains a unique *primeSeed* code that records the compounds allocated in the corresponding test set. The same *primeSeed* codes were included in every script for the machine learning method and QSPR to ensure identical compounds were tested by each method.

Regression modeling was employed with each combination of descriptors as input vectors. In Gaussian process modelling, the initial values of the logarithms of the characteristic length scale, the

signal variance and the noise variance were chosen using cross validation from 10 user-defined pre-sets. In addition, a 5-fold cross-validation procedure was used to select optimal parameters for each test. In such cases, each training set was divided further into training and validation sets 5 times.

To investigate which compound descriptors contribute significantly to the prediction, GPRARD methods were applied to the dataset. Experiments were again conducted on 10 randomly selected training and test sets. However, in this case the hyperparameters are optimised by maximizing the marginal likelihood using the derivative rather than selecting from pre-set hyperparameters using a cross validation procedure (Geinoz et al., 2004). In each case the logarithms of characteristic length-scale, signal variance and noise variance were initialized for each input dimension, as [0; 0; 0; 0; 0; 0; log(SQRT(0.1))]. Rasmussen and Williams' (2006) GP Toolbox was applied to the dataset to carry out Gaussian process modeling.

Performance measurements of QSPR models and machine learning methods were calculated via Matlab R2008a. The parameters employed to ascertain statistical quality of each model were percent improvement over the naïve model, (ION, %), normalised mean squared error (NMSE) and the correlation coefficient (CORR), as described above and employed previously (i.e. Moss et al., 2009).

RESULTS

Consideration of the results of this study can be divided broadly into four regions: the shape of the distribution between physiochemical properties and skin permeability, the comparison of the prediction accuracy between machine learning methods modelling and QSPR models, the comparison of the accuracy to quantify percutaneous absorption via non-linear and linear approaches and the selection of features that are significant in the mathematical quantification of skin absorption. As a measure of performance, ION, NMSE and CORR have been employed.

Distribution of the physicochemical parameters and permeability coefficients.

Visualisation of the data provides an insight into the relationship between physiochemical properties and permeability coefficients among 142 compounds employed in this study. In common with previous work in this field (Moss et al., 2009; Sun et al., 2008) the visualisation of the data shown in Figure 1 suggests that the underlying relationship between the physicochemical descriptors and permeability coefficients is inherently non-linear. For example, the response between log K_p and log P, shown in Figure 1a, has a scattered distribution which indicates that this relationship is non-linear. This is congruent with other literature evidence (i.e. Degim et al., 2003; Moss et al., 2006). Moreover, other molecule descriptors including melting point, molecular weight, solubility parameters, HA and HD also showed non-linearity with log K_p.

Further, the visualisation of the data described previously ((Moss et al., 2009; Sun et al., 2008)) also indicates that the skin permeability coefficient is not solely dependent on one molecular descriptor.

Compounds with similar properties for one particular feature can demonstrate enormous variations in log K_p. For example, if compounds with one hydrogen bond donor group are considered, log K_p is observed to vary from -1.2 to -5.0 approximately.

However, it should be noted that certain parameters, such as hydrogen bond donor and acceptor groups, may be considered as discrete rather than continuous variables, and as such a linear relationship between these parameters and descriptors that are continuous in nature (such as log P, MW or solubility parameter) should not necessarily be expected, and may be of limited statistical value.

Statistical evaluation of model quality

Figure 1 indicates that those latter models, derived from a more comprehensive extension of Flynn's (1990) dataset (for example, those that incorporate data from other studies (i.e. Johnson et al., 1995;, Kirchner et al., 1997; Degim et al., 1998)) result in improved predictions. It should also be noted that, as expected, the model proposed by Barratt (1995) performs relatively weakly due to the limitation in the number of observation included in that study, a point made previously in the literature (Genioz et al., 2004). This suggests the importance of dataset validity, particularly with regard to size, the consistency of experimental protocols, reproducibility, comprehensiveness in model developments (Moss et al., 2002). This point was highlighted by Moss and Cronin (2002) who developed a QSPR model which does not include the steroid data used by Scheuplein et al. (1969), which is collated into Flynn's dataset (1990), but instead uses the data colleted by Johnson et al. (1995). The inclusion of the model by Barratt (1995) suggests a possible limitation in the use of this data, which the results in Figure 1 would appear to substantiate.

As discussed above, log P, molecular weight and terms pertaining to hydrogen bonding have been widely identified as highly significant phenomena in developing a mechanistic understanding of percutaneous absorption. Despite this, the QSPR-type models employed in this study – and which contain most, if not all of these parameters – fail to accurately predict K_p; in most cases, they return predictions that are, in terms of the statistical tests used to compare the performance of models (i.e. measures of ION, NMSE and CORR), significantly worse than the naïve model, which is simply the average K_p value of the whole dataset. It can be seen that, by using the same parameters, Gaussian Process and Single Layer Network models provide statistically better results than QSPR models, particularly in terms of higher ION and lower NMSE values, although difference is NMSE values are not always as pronounced as those for ION. Nevertheless, the improvement is statically valid ($p < 0.05$, Table 1). For the combination of features discussed in the preceding section, the Gaussian Process demonstrates the best results, even compared to Single Layer Network, in terms of both model's prediction accuracy (ION) and stability (NMSE) as shown in Table 1. However, as discussed previously (Moss et al., 2009) the nature of the dataset and its compatibility with a particular mathematical approach should be considered. Nevertheless, Figure 2 shows a clear improvement in the predictivity of the GP model, compared to Potts and Guy (1992).

Feature Selection

Due to the large number of statistical comparisons made between each possible combination of molecular features, Table 2 shows only a condensed comparison of the statistical tests carried out comparing the combination of features in Gaussian Process models. Specifically, it includes only models that demonstrated no significant difference compared to the highest ION (%) ranked model [GPR: MPt, log P and HD] – in effect, the best performing combination of features as defined by the

results of the statistical comparison of models. Table 4 shows the Gaussian Process models that, on both ION (%) and NMSE measurements, demonstrated no significant difference compared to Gaussian Process models with better performance measures, as well as either no significant difference or significant improvement compared to Gaussian Process models with worse performance measures. The statistically "best" Gaussian models all contain the specific combination of log P and HD, coupled with either melting point or molecular weight. It appears that melting point and molecular weight are, in a purely modelling context, interchangeable in this process, and replacing one with the other does not exert a detrimental effect on a particular model. It should also be pointed out that the reduced correlation coefficient observed for the Potts and Guy equation (in Table 1) may be as a result of the application of this model to our dataset, which differs from that used originally to develop this model, and which may be potentially of limited value.Table 5 shows a summary of length scale analysis, calculated by Gaussian Process Automatic Resonance Detection for each feature in the Gaussian Process models recorded in Table 2. Essentially, a lower length scale value indicates a higher significance of the role of a particular molecular descriptor in predictions of permeability coefficients. From Table 5, it can be seen that the difference in the length scale factor between the molecular features in each model is relatively small. The only exception is the solubility parameter, which demonstrated a minimum of two significance figures difference compared to other molecular descriptors. In essence, this indicates that the solubility parameter is not a significant feature in the quantification of percutaneous absorption. This is further supported by addition of solubility parameters into the combination of descriptors, which did not lead to a significant improvement in model predictivity. For example, the Gaussian Process combinations [GPR: MW, MPt, SP, log P, HD] and [MPt, log P, HA, HD] offer equally significant predictions of log K_p. In some cases the inclusion of solubility parameters can cause significant

reductions in a model's predictivity – for example, the Gaussian Process combination [GPR: MW, MPt, log P, HA, HD] is more significant than [GPR: MW, MPt, SP, log P, HA, HD].

The results of this particular analysis suggest that, using the physicochemical descriptors of log P, the number of hydrogen bonding donor groups, and either molecular weight or melting point, results in a Gaussian Process model with optimal predictivity and that the addition of further molecular descriptors does not improve the quality of the model.

Comparison of non-linear and linear predictions of skin absorption

Lian et al., (2008) suggested that the simplicity of linear equations enhance the ability of a model to provide accurate predictions. This comment has been explored in this study, where the difference in predictivity of the permeability coefficient between Gaussian Process and Single Layer Network modelling has been explored. The results in Table 4 indicate that the Gaussian Process provides significantly better predictions of log K_p than Single Layer Networks for the overall highest ION model, as well as the best models within its categories based on specific combinations of physicochemical descriptors. The only exception was the model with two features, where the overall best Single Layer Network model (where MPt and HA are returned as the most significant parameters) demonstrated no significant difference with Gaussian Process model [GPR: MW and HD]. The results of the statistical comparisons (paired t-tests) of these models is summarised in Table 5. These results suggest that Gaussian Process modelling is, in statistical terms, the most appropriate model of those analysed to employ in predicting percutaneous absorption, with the observed differences being statistically significant. In terms of models quality (i.e. accuracy of

prediting K_p) the statistical comparisons used in this study would suggest the following rank order: GP > SLN > QSPR (all types).

DISCUSSION

An important point in this study is that the composition of the dataset (the inputs) clearly affects the nature of any model derived (the output). This may seem obvious but it is important to make such a point, given that the dataset used in this study is different from those employed to develop the established QSPR models. However, the specific composition of the dataset can clearly influence the nature of the model. Moss et al. (2009) discussed this, in terms of the breadth of the Flynn (1990) dataset which underpins so much of the work in this field. That dataset is composed predominately of molecules which, for example, have log P values less than 2.0. Moss et al. (2009) argued that this may in effect be providing only a limited picture of percutaneous absorption, limiting the applicability of the model, and this indeed has been addressed by other researchers (i.e. Wilschut et al., 1995) where non-linear modifications of the Potts and Guy (1992) equation were proposed. This suggests that a simple linear relationship between K_p and any number of molecular descriptors may not fully represent percutaneous absorption, and may result in limited or inaccurate predictivity for a particular model. In addition, data visualisation also suggests a clear non-linear relationship between physicochemical properties of molecules, suggesting no clear linear trend between any of these descriptors (Ghafourian & Fooladi, 2001). Those QSPR models that can be loosely described as being of the "Potts and Guy" type suggest that a linear response exists between, for example, K_p and log P. As discussed recently (i.e. Moss et al., 2006; Neumannet et al., 2005; Pannier et al., 2002) it should be noted that such a relationship only exists within the specific range of the models; ostensibly, this reflects the range of data employed to construct the model. It may be the case that the models are therefore limited by the range of their dataset and that this study, and

those like it, yield models that are more representative of percutaneous absorption across a wider range of physicochemical properties.

A range of non-linear methods have been employed to improve predictions of skin absorption. Artificial neural networks (ANN) have been employed (Degim et al., 2003), showing high predictive power. However, it is a limited method in that ANN's have a tendency to over-fit where large numbers of physicochemical descriptors exist, compared to the data points used. Such models are often weighted and are susceptible to over-training (Neumann et al., 2006). This results in idiosyncratic results, particularly as the output will tend to fit the noise in such cases, providing poor predictivity for new compounds (Guha & Jurs, 2005). GP methods do not alleviate all these issues, but minimise them (Rasmussen and Williams, 2006), providing better predictions of percutaneous absorption than existing models (Moss et al., 2009).

Therefore, this study employed GP methods of analysis and, in particular, Gaussian Process Automatic Resonance Detection. This measures the covariance and length scale of each feature in the combination. The inverse of the length scale determines the relevance between input and output, thereby a low length scale value implies that the input and covariance are highly dependent on each other. In other words, this can reduce the limitation of Gaussian Process caused by a "black box" approach (Moss et al., 2009) and provide an insight into the significance of specific molecular descriptors. Single layer networks (SLN) were also evaluated as they allow interpretation of the predictivity limitations in linear model at different ranges of features, providing a comparison between linear QSPR and machine learning methods (Gramatica et al., 2007).

Data visualisation (Moss et al., 2009) also indicates that K_p is not solely dependent on one molecular descriptor. Compounds with similar properties for one particular descriptor can demonstrate enormous variations in log K_p. For example, for compounds with one hydrogen bond donor group, log K_p is observed to vary from -1.2 to -5.0 approximately. Such a visualisation of data clearly demonstrates the synergic effects between the physicochemical features investigated in this study and would indicate either that more than one physicochemical descriptor is required to successfully model percutaneous absorption, or that such parameters are not independent of each other (such as the relationship between log P and molecular weight) and that the use of particular parameters may be limited in terms of gaining specific understandings of mechanisms of absorption. Further, effects such as ionisation (and therefore solubility and speciation) have not been considered by any of these studies.

Nevertheless, Figure 2 demonstrates a clear improvement in predictivity by the GP model compared to the Potts and Guy (1992) model. Figure 2 contains data points obtained from a subset of the overall dataset, due to the methods employed for the generation of tests sets, as described in the previous section. The test set shown in Figure 2 is that which results in the Potts and Guy (1992) model achieving the best performance among the ten test sets generated by this analysis. This is compared with experimental log K_p and predicted K_p from the model with the highest ION (%) value [GPR: MPt, log P and HD. It is a good example of how GP methods provide a better fit to experimental log K_p in contrast to Potts and Guy (1992). Even with such a subset, where performance is, in effect, at an optimum, the statistical performance of the Potts and Guy (1992) model results in a majority of the predicted log K_p values being distinctly different from the experimental log K_p values, as indicated by comparatively poor ION and NMSE values. Even in this case the GP model is, in statistical terms, more accurate. Further, and rather qualitatively, it may

be suggested that the scatter of the output shown in Figure 2 from the GP is substantially less linear than that from the QSPR model, and that the latter appears to be more representative of the scatter associated with the experimental data.

In the best Gaussian Process models (shown in Table 4), every combination of features contains log P and HD together with either melting point or molecular weight. This suggests that molecular weight, melting point, log P and HD are important features in permeability coefficient predictions. It also suggests at the inter-relationship and lack of independence of certain descriptors. Interestingly, in this type of combination, melting point and molecular weight are inter-exchangeable to give predictions with no significant differences; for example, the Gaussian Process combinations [MPt, Log P, HD] and [MW, Log P, HD] produce models of a similar statistical quality. Furthermore, addition of molecular weight or melting point to these models does not significantly influence log K_p predictions, such as the Gaussian Process combinations [MW, log P, HD], [MPt.log P, HD] and [MW, MPt, log P, HD], which all demonstrate no significant difference in performance measures. This inter-exchangeability implies that high level of correlation existed between melting point and molecular weight (Williams, 2003). On the other hand, it should be considered that, while molecular weight and melting point are interchangeable for modelling purposes, this does not necessarily indicate a degree of correlation between the two parameters. This is reflected in the findings of a previous Gaussian Process study (Moss et al., 2009).

HD only appears to exert its importance in skin permeability when coupled with log P. When log P is absent, inclusion of HD in the model can significantly decrease a model's predictive power. However, when a model is constructed containing log P, HD constantly demonstrated a lower length scale value than HA. In this case, addition of HA to the model does not result in improvements in

predictivity. This means that the GP models [GPR: MW, log P, HA, HD] and [GPR: MW, logP, HA] have a similar statistical performance, whereas removal of HD can significantly reduce performance measures. For example, [GPR: MPt, log P, HD, HA] is significantly better than [GPR: MPt, log P, HA] ($p = 0.0022$).

It is not just the removal of HD that impacts on the statistical quality of models. For example, in the absence of log P, the GP model [GPR: MW, MPt and HA] demonstrated no significant difference in ION (%) value compared to [GPR: MPt, Log P, HD], the latter being the model with the best overall performance measures. This might be due to the molecular weight bias of the dataset employed in this study (Poda et al., 2001), which is predominately based on Flynn (1990). Further, this also highlights that the effects of ionisation might not have been considered in the development of these models, or in previous models that employ such literature data. It should also be noted, however, that [GPR: MW, MPt and HA] performs poorly in NMSE measurements, indicating that this model is not, in a statistical sense, a stable and reliable combination.

The results presented in Table 4 indicate that the Gaussian Process provides significantly better predictions of log K_p than Single Layer Networks for the overall highest ION model, as well as the best models within its categories based on specific combinations of physicochemical descriptors. The only exception was the model with two descriptors, where the overall best Single Layer Network model (which returned MPt and HA as being the most significant parameters) demonstrated no significant difference with the GP model [MW and HD]).

The "black box" approach as presented previously (Moss et al., 2009) does not allow the elucidation of mechanistic information, only predictions of K_p for chemicals of interest. The current study, and

the use of feature selection methods, allows all combinations of molecular descriptors to be assessed for their ability to improve statistically the quality of models generated. While this has resulted in a clear understanding of the models that will improve prediction of K_p, it has demonstrated that the combination of descriptors responsible for such improvements is not always clear or consistent. This essentially demonstrates the interconnection of the parameters used. For example, an increase in lipophilicity can be achieved by increasing MW (Buchwald & Bodor, 2001) and such increases are not necessarily linear. A compound's lipophilicity is determined by its chemical structures and the position of the aromatic ring; carbons, benzene rings and amide groups can increase log P (Geinoz et al., 2002; Refsgaard et al., 2005). As molecular weight increases, the number of carbon skeletons increases and therefore the lipophilic surface of a compound increases. The increase in number of hydrophobic alkane groups is considered as the major contribution to the increase in lipophilicity and hence, to an extent, permeability (Roberts et al., 1995). Water prefers to interact with hydrogen bonding groups or ionic molecules rather than non-polar compounds (Williams, 2003). Ghasemi and Saaidpour (2007) highlighted that, as molecular weight increases, the increase in lipophilicity resulted in the compound becoming non-polar, increasing solubility in the stratum corneum and reducing solubility in the aqueous environment (dermis). This is consistent with the findings of previous studies in this field (Moss et al., 2006; 2009).

As number of hydrogen bonding groups on molecule increases, the ability of the molecule to form hydrogen bond with water increases and therefore lipophilicity decreases. Therefore, hydrogen bond can, indirectly, be an indication of log P. Fitzpatrick et al., (2004) suggested that a hydrogen bond related descriptor should be included in a model when there is the absence of a parameter directly relating to lipophilicity.

Compared to log P and molecular weight, the exact mechanistic understanding of how hydrogen bonding influences percutaneous absorption is less clear. Several authors (for example, Abraham et al., 1995; Pugh et al., 1996; Ravesky and Schaper , 1998) also demonstrated that hydrogen bonding is highly related to skin permeability. Poulin & Krishnan (2001) and Potts and Guy (1995) suggested that hydrogen bonding can significantly influence percutaneous absorption by reducing the compound's ability to penetrate the skin, and that hydrogen bond acceptor groups play a more significant role than donor groups, a suggestion also supported by Pugh et al. (1996).

However, the findings of this study suggest a different conclusion, where generally acidic hydrogen bond donor groups have been shown to be more significant than generally basic hydrogen bond acceptor groups. These findings are in agreement with those presented by El Tayar et al (1991) and Geinoz et al. (2002). This discrepancy may be due to the role ionisation plays in both the overall process of percutaneous absorption but also in the nature of the descriptors. Poulin and Krishnan (2001) suggested that the effects of lowering K_p by hydrogen bonding are particularly strong when the molecule has two or more hydrogen bonding donors or acceptors group in the compounds. According to Roberts et al., (1995) and Ghafourian & Fooladi (2001) inclusion of one hydrogen bond group (either a donor or acceptor) to the hydrocarbon skeleton would cause a substantial reduction in K_p. Addition of subsequent groups also reduce K_p, but do so in a non-linear additive manner.

Hadgraft (2004) highlighted that interactions between compound and the polar head groups of skin lipids in the intercellular channels plays a significant role in percutaneous absorption. Molecules containing hydrogen bond groups can associate with the immobilised polar head groups of the lipids. As a result, their passage across the skin may be hindered, decreasing the diffusion

coefficient and reducing their ability to diffuse across stratum corneum (Pugh et al., 2000). This may result in hydrogen bonding and ionic forces modification, which implies a change of head group domains. This complicates skin penetration as such an alteration may influence permeation of other exogenous chemicals in, for example, the same formulation. The number of hydrogen bond groups may vary during the partitioning process. For example, once a donor group donates a hydrogen bond, it has the potential to become a hydrogen bond acceptor, while groups that have not been ionised remained as hydrogen bond donor groups. Further, intermolecular hydrogen bond has a substantial influence on aqueous solubility (Yin et al., 2002) since the O-H and N-H bonds are strongly polarized and may readily facilitate donation.

Most of the permeants in the data set are either weak acids or weak bases. Hence, ionisation can occur at different pH values (Hadgraft, 2004). According to Aberg et al. (2008), the skin surface is acidic with pH ranging from 4 to 6. However, the pH of extracellular fluid in the body is approximately 7.4, and implies a large pH gradient between the stratum corneum and underlying tissues. Removal or addition of a hydrogen bond can lead to a compound becoming ionised. The extracellular stratum corneum lipid contains free fatty acids that can undergo dissociation, resulting in a negative surface charge caused by the presence of ionised carboxyl groups (Aberg et al., 2008). As the skin is a negatively charged membrane, this electrostatic interaction becomes a hindrance of ionised penetrants (Raiman et al., 2003). Thus, ionic compounds, particularly cations, have a lower ability to penetrate the skin compared to neutral compounds.

The disparity between the findings of Potts and Guy (1995) and this study also relate to ionisation. During the process of experimentally measuring K_p, the solute is placed in a solvent where it is possible for the solute to interact with the solvent and, depending on the pK_a of the solute and the

pH of the solvent, for the solute to ionise. It should also be considered that log P values are also measured with ionisable compounds under conditions that may favour more ionic species, thus influencing the log P value obtained. Thus, experimental measurements of K_p might not appropriately reflect the effects of hydrogen bonding but may instead reflect the effects of ionisation.

CONCLUSIONS

In comparing different approaches for developing predictive models of percutaneous absorption, the current study agrees with previous work (Moss et al., 2009) that suggests the inherently non-linear nature of the skin data set used. Further, Gaussian Process machine learning methods produce statistically more robust models than other approaches (SLN or QSPR-based models). The use of feature selection enables the development of a mechanistic understanding of percutaneous absorption. While this approach results in specific models that are statistically superior, it also indicates clearly the interdependence of the physicochemical descriptors employed in this, and in many other, studies. This suggests that the approach of quantifying models of skin absorption by means of a simple equation may have limited mechanistic value. While hydrogen bonding appears to play an important role in percutaneous absorption the issue of ionisation may limit the validity and accuracy of models.

REFERENCES

Aberg, C., Wennerstrom, H., Sparr, E. (2008). Transport processes in responding lipid membranes: a possible mechanism for the pH gradient in the stratum corneum. *Langmuir* **24**, 8061 – 8070.

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

Barratt, M. D. (1995). Quantitative structure-activity relationships for skin permeability. *Toxicology in Vitro*, **9**, 27 – 37.

Buchwald, P., Bodor, N. (2001). A simple, predictive, structure-based skin permeability model. *Journal of Pharmacy and Pharmacology* **53**, 1087 – 1098.

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

Degim, I. T., Hadgraft, J., Ilbasmis, S., Ozkan, Y. (2003). Prediction of skin penetration using artificial neural network (ANN) modelling. *Journal of Pharmaceutical Sciences* **92**, 656 – 664.

Degim, I. T., Pugh, J. W., Hadgraft, J. (1998). Skin permeability data: anomalous results. *International Journal of Pharmaceutics* **170**, 129 – 133.

El Tayar, N., Tsai, R. S., Testa, B., Carrupt, P. A., Hansch, C., Leo, A. (1991). Percutaneous penetration of drugs: a quantitative structure-permeability relationship study. *Journal of Pharmaceutical Sciences* **80**, 744 – 749.

Fedors, R. F. (1974) A method for estimating both the solubility parameters and molar volumes of liquids. *Poly. Eng. Sci*. **14**, 147 – 154.

Fitzpatrick, D., Corish, J. Hayes, B. (2004). Modelling skin permeability in risk assessment – the future. *Chemosphere*, **55**, 1309 – 1314.

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

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

Geinoz, S., Rey, S., Boss, G., Bunge, A., Guy, R.H., Carrupt, R.A., Reist, M. Testa, B. (2002). Quantitative structure-permeation relationships for solute transport across silicone membranes. *Pharmaceutical Research* **19**, 1622 – 1629.

Ghafourian, T., Fooladi, S. (2001). The effect of structural QSAR parameters on skin penetration. *International Journal of Pharmaceutics **217***, 1 – 11.

Gramatica, P., Giani, E., Papa, E (2006). Statistical external validation and consensus model;ing: a QSPR case study for K_{oc} prediction. Journal of Molecular Graphics and Modelling ***25*96**, 755 – 766.

Guasemi, J., Saidpour.S. (2007). Quantitative structure–property relationship study of *n*-octanol–water partition coefficients of some of diverse drugs using multiple linear regression. *Analytica Chimica Acta* **604**, 99 – 106.

Guha, R. J, P.C. (2005). Interpreting computational neural network QSAR models: a measure of descriptor importance. Journal of Chemical Information and Modelling **45**, 800 – 806.

Hadgraft, J. (2004). Skin deep. *European Journal of Pharmaceutics and Biopharmaceutics* **58**, 291 – 299.

Johnson, M. E., Blankschtein, D., & Langer, R. (1995). Permeation of steroids through human skin. *Journal of Pharmaceutical Sciences* **84**, 1144 – 1146.

Kamlet, M.J., Abboud, J.L., Abraham, M.H., Taft, R.W. (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,** 2877 – 2887.

Katritzky, A. R. Dobchev, D. A., Fara, D. C., Hur, E. Tamm, K., Kurunczi, L., Karelson, M., Varnek, A., Solovev, V. (2006). Skin permeation rate as a function of chemical structure. *Journal of Medicinal Chemistry* **49**, 3305 – 3314.

Kirchner, L. A., Moody, R. P., Doyle, E., Bose, R., Jeffery, J., Chu, I. (1997). The Prediction of Skin Permeability by Using Physicochemical Data. *ATLA Alternatives to Laboratory Animals* **25**, 359 – 370.

Lian, G., Chen, L. Han, L. (2008). An evaluation of mathematical models for predicting skin permeability. *Journal of Pharmaceutical Sciences* **97**, 584 – 598.

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

Michaelis, A.S., Chandrasekaran, S.K., Shaw, J.E. Drug permeation through human skin: theory and *in vitro* experimental measurement. *AIChE*. **21** (1975) 985 – 996.

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

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

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

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

Neumann, D., Kohlbacher, O., Merkwirth, C., Lengauer, T. (2006). A Fully Computational Model for Predicting Percutaneous Drug Absorption. Journal of Chemical Information and modelling **46**, 424 – 429.

Pannier, A. K., Brand, R. M. Jones, D. D. (2003). Fuzzy modelling of skin permeability coefficients. *Pharmaceutical Research* **20**, 143 – 148.

Patel, H., ten Berge, W., Cronin, M. T. D. (2002) Quantitative structure-activity relationships (QSARs) for the prediction of skin permeation of exogenous chemicals. *Chemosphere* **48**, 603 – 613.

Poda, G. I., Landsittel, D. P., Burmbaugh, K., Sharp, D. S., Frasch, H, . F., Demchuk. (2001). Random sampling or 'random' model in skin flux measurements? Commentary on: Investigation of the mechanism of flux across human skin in vitro by quantitative structure-permeability relationships. *European Journal of Pharmaceutical Science* **14**, 197 – 200.

Potts, R.O., Guy, R.H. (1992) Predicting skin permeability, *Pharmaceutical Research.* **12**, 663 – 669.

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

Poulin, P., Krishnan, K. (2001). Molecular structure-based prediction of human abdominal skin permeability coefficients for several organic compounds, *Journal of Toxicology and Environmental Health* **62**, 143 – 159.

Pugh, W. J., Degim, I. T., Hadgraft, J. (2000). Epidermal permeability–penetrant structure relationships: 4, QSAR of permeant diffusion across human stratum corneum in terms of molecular weight, H-bonding and electronic charge. *International Journal of Pharmaceutics.* **197**, 203 – 211.

Pugh, W.J., Roberts, M., Hadgraft, J. (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**, 149 – 165.

Raiman, J., Hanninen, K., Kontturi, K., Murtomaki, L. & Hirvonen, J. (2003). Drug adsorption in human skin: A streaming potential study. *Journal of Pharmaceutical Sciences* ***92***, 2366 – 2372.

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

Ravesky, O. A., Schaper, K. J. (1998). Quantitative estimation of hydrogen bond contribution to permeability and absorption processes of some chemicals and drugs. *European Journal of Medicinal Chemistry* **33**, 799 – 807.

Refsgaard, H. H. F., Jensen, B. F., Brockhoff, P. B., Padkjaer, S. B., Guldbrandt, M., Christensen, M. S. (2005). *In silico* prediction of membrane permeability from calculated molecular parameters. *Journal of Medicinal Chemistry* **48**, 805 – 811.

Roberts M: Percutaneous absorption of phenolic compounds; PhD Thesis, University of Sydney, Sydney, 1976.

Roberts, M., Pugh, W.J., Hadgraft, J. (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**, 23 – 32.

Roberts, M., Pugh, W.J., Hadgraft, J., Watkinson, A. (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**, 219 – 233.

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

Scheuplein, R.J., Blank, I.H., Brauner, G.I., MacFarlane, D.J., (1969). Percutaneous absorption of steroids. *J. Invest. Dermatol*. **52**, 63–70.

Sun, Y., Moss, G.P., Prapodopolou, M., Davey, N., Adams, R., Brown, M.B. (2008). Predictions of Skin Penetration Using Machine Learning Methods, In: Proceedings of 8th IEEE International Conference on Data Mining, Pisa, Italy. Edited by F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan, and X.D.Wu. ISBN 978-0-7695-3502-9.

Williams, A. C. (2003). *Transdermal and Topical Drug Delivery*. London: Pharmaceutical Press.

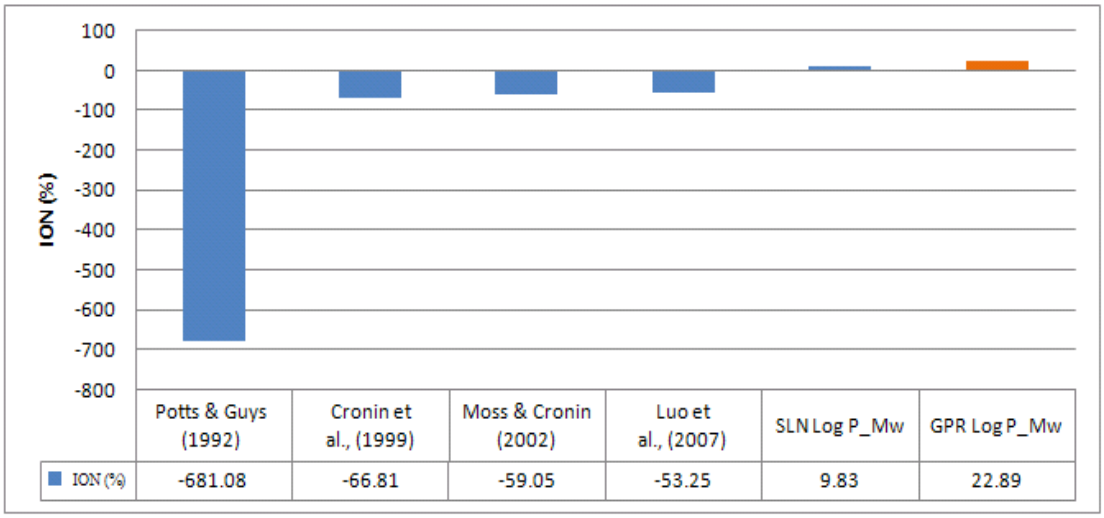
Wilschut, A., ten Berge, W.F., Robinson, P.J., McKone, T.E. (1995) Estimating skin permeation — the validation of 5 mathematical skin permeation models. *Chemosphere* **30**, 1275 – 1296.

Yin, C., Liu, X., Guo, W., Lin, T., Wang, X. (2002). Prediction and application in QSPR of aqueous solubility of sulfur-containing aromatic esters using GA-based MLR with quantum descriptors. *Water Research* **36**, 2975 – 2982.

FIGURES

Figure 1.

(a)


100
0
-100
-200
-300
-400
-500
-600
-700
-800
ION (%)
Potts & Guys (1992)
Cronin et al., (1999)
Moss & Cronin (2002)
Luo et al., (2007)
SLN Log P_Mw
GPR Log P_Mw
ION (%)
-681.08
-66.81
-59.05
-53.25
9.83
22.89

(b)

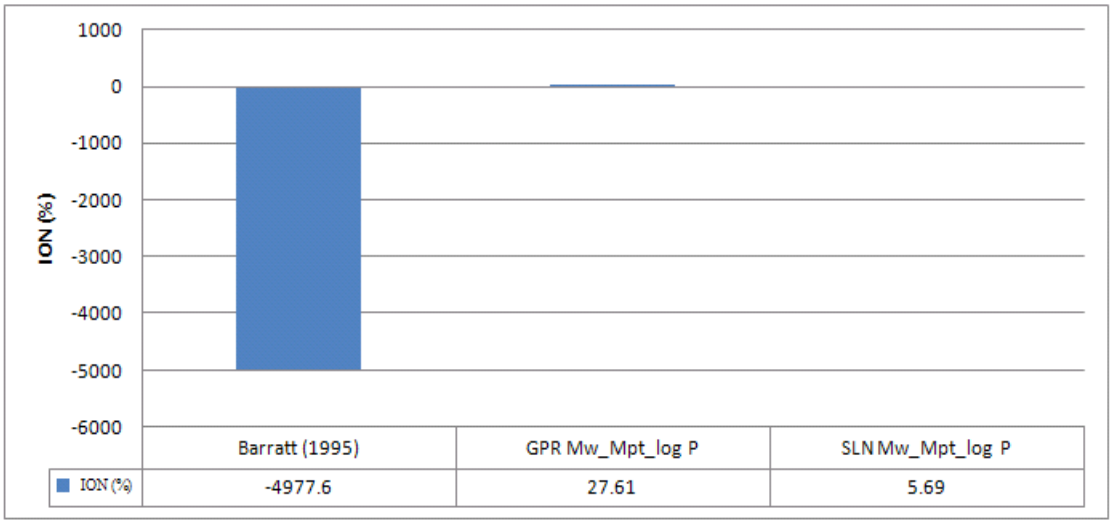
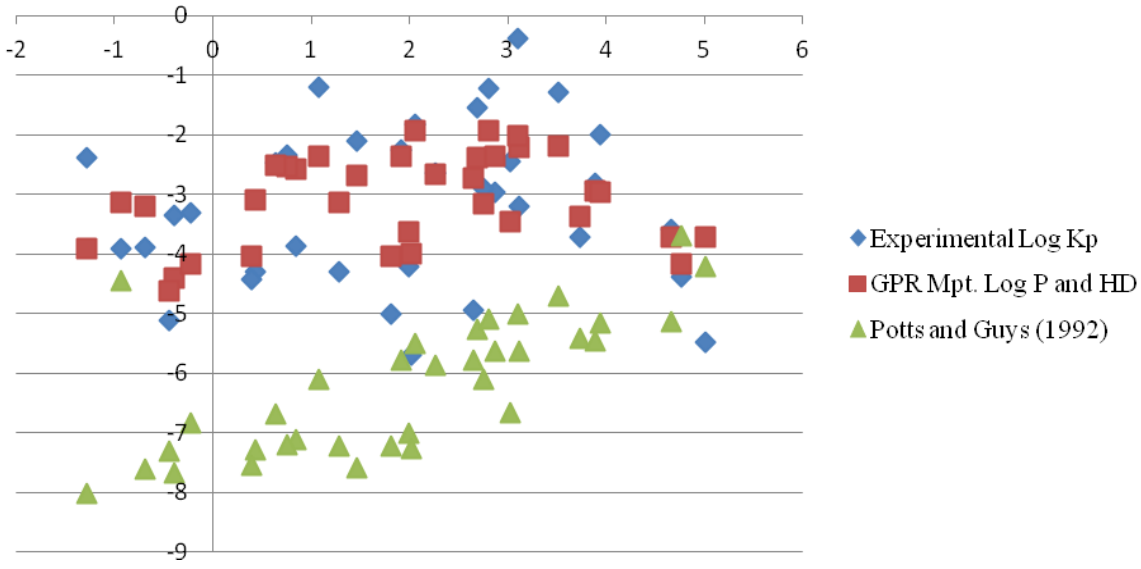

1000
0
-1000
-2000
-3000
-4000
-5000
-6000
ION (%)
Barratt (1995)
GPR Mw_Mpt_log P
SLN Mw_Mpt_log P
ION (%)
-4977.6
27.61
5.69

Figure 2.


0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-2
-1
0
1
2
3
4
5
6
Experimental Log Kp
GPR Mpt. Log P and HD
Potts and Guys (1992)

TABLES

Table 1.

Descriptors	Models or Machine Learning Method Assessed	ION (%) ± SD (higher number is best)	NMSE ± SD (smallest number is best)	CORR ± SD (highest value is best)	p-values (QSPR vs. SLN)			p-values (QSPR vs. GP)			p-values (SLN vs GP)		
					ION (%)	NMSE	CORR	ION (%)	NMSE	CORR	ION (%)	NMSE	CORR
Log P and Molecular Weight	Potts & Guy (1992)	-681.08 ± 139.93	7.52 ± 0.80	0.25 ± 0.11	0.00	0.00	0.08	0.00	0.00	0.01			
	Cronin et al., (1999)	-66.81 ± 41.90	1.70± 0.42	0.31 ± 0.12	0.00	0.00	0.36	0.00	0.00	0.00			
	Moss & Cronin (2002)	-59.05 ± 39.06	1.62 ± 0.39	0.29 ± 0.12	0.00	0.00	0.26	0.00	0.00	0.002			
	Luo et al., (2007)	-53.25 ± 36.24	1.56± 0.35	0.24 ± 0.11	0.00	0.00	0.06	0.00	0.00	0.00			
	SLN	9.83 ± 11.1	0.93 ± 0.17	0.34 ± 0.17							0.00	0.00	0.01
	GPR	22.89 ± 10.62	0.79 ± 0.17	0.49 ± 0.11									
Log P, Molecular Weight and Melting Point	Barratt (1995)	-4977.6 ± 3318.9	51.60 ± 34.32	0.30 ± 0.15	0.00	0.00	0.76	0.00	0.00	0.00			
	SLN	5.69 ± 12.32	0.97 ± 0.19	0.31 ± 0.14							0.00	0.00	0.00
	GPR	27.61 ± 9.32	0.75 ± 0.15	0.53 ± 0.11									

Table 2.

ION (%) ranking	GPR Combination of Features	MPt.Log P.HD	MW.MPt.Log P.HD	MW.MPt.SP.Log P.HD	MPt.Log P.HA.HD	MW.SP.Log P.HD	MPt.SP.Log P.HD	MW.MPt.Log P.HA.HD	MW.Log P.HD	MW.Log P.HA.HD	MW.MPt.SP.Log P.HA.HD	MW.MPt.HA	MW.MPt.HA.HD
1	GPR MPt.LogP.HD	-	X	X	X	X	X	X	X	X	X	X	X
2	GPR MW.MPt.LogP.HD	X	-	X	X	X	X	X	X	√	X	X	√
3	GPR MW.MPt.SP.LogP.HD	X	X	-	X	X	X	√	X	√	X	√	√
4	GPR MPt.LogP.Ha.HD	X	X	X	-	X	X	X	X	X	X	X	X
5	GPR MW.SP.LogP.HD	X	X	X	X	-	X	X	X	X	X	X	√
6	GPR MPt.SP.LogP.HD	X	X	X	X	X	-	X	X	X	X	X	√
7	GPR MW.MPt.LogP.Ha.HD	X	X	√	X	X	X	-	X	X	√	X	√
8	GPR MW.LogP.HD	X	X	X	X	X	X	X	-	X	X	X	X
9	GPR MW.LogP.Ha.HD	X	√	√	X	X	X	X	X	-	X	X	X
10	GPR MW.MPt.SP.LogP.Ha.HD	X	X	√	X	X	X	√	X	X	-	X	√
11	GPR	X	X	√	X	X	X	X	X	X	X	-	√

	MW.MPt.HA												
12	GPR MW.MPt.HA .HD	X	√	√	X	√	√	√	X	X	√	√	-

Note: X indicates no significant difference ($p < 0.05$) and √ indicates a significant difference ($p > 0.05$) between the two groups compared.

Table 3.

	Statistical performance measures			Length scale						
Combination of features	**ION (%) ± SD**	**NMSE ± SD**	**CORR ± SD**	**MW**	**MPt**	**SP**	**Log P**	**HA**	**HD**	**Features significance Ranking**
MPt.LogP.HD	37.59 ± 8.54	0.64 ± 0.13	0.63 ± 0.09	-	1.23	-	0.51	-	0.99	Log P > HD > MPt
MW.MPt.Log P.HD	37.40 ± 7.56	0.65 ± 0.15	0.62 ± 0.09	5.22	1.28	-	0.51	-	1.03	Log P > HD > MPt > MW
MW.MPt.SP. Log P.HD	37.35 ± 7.23	0.65 ± 0.14	0.62 ± 0.09	5.20	1.27	31.09	0.51	-	1.0	Log P > HD > MPt > MW > SP
MPt.Log P.Ha.HD	35.19 ± 10.81	0.67 ± 0.18	0.62 ± 0.10	-	1.14	-	0.85	2.51	1.11	Log P > HD > MPt > HA
MW.SP. Log P.HD	35.12 ± 7.08	0.67 ± 0.12	0.62 ± 0.08	0.77	-	83.70	0.64	-	0.62	HD > Log P > MW > SP
MPt.SP.Log P.HD	34.21 ± 11.46	0.68 ± 0.19	0.61 ± 0.10	-	1.22	24.47	0.51	-	0.98	Log P > HD > MPt > SP
MW.LogP.HD				0.77	-	-	0.64	-	0.62	HD > Log P > MW
MW.Log P.Ha.HD				0.62	-	-	0.78	0.64	0.41	HD > HA > Log P > MW
MW.MPt.SP. Log P.Ha.HD				0.90	1.31	53.92	0.86	0.70	0.39	HD > HA > Log P > MW > MPt > SP
MW.MPt.Ha				0.38	0.86	-	-	0.43	-	MW > HA > MPt
MW.MPt.Log P. Ha.HD				0.90	1.32	-	0.87	0.70	0.40	HD > HA > Log P > MW > MPt

MW.MPt.Ha.HD				0.26	1.91	-	-	0.38	0.70	MW > HA > HD > MPt

Table 4.

Highest ION (%) model	GPR models	SLN models	GPR ION(%) ± SD	GPR NMSE ± SD	SLN ION (%) ± SD	SLN NMSE ± SD	P-value (ION %)	P-value (NMSE)	Significant difference (ION %)	Significant difference (NMSE)
Overall	MPt.LogP.HD	MPt.HA	37.59 ± 8.54	0.64 ± 0.13	11.23 ± 11.29	0.91 ± 0.13	0.00	0.00	Y	Y
2 features	MW.HD	MPt.HA	25.54 ± 12.90	0.77 ± 0.19	11.23 ± 11.29	0.91 ± 0.13	0.34	0.046	N	Y
3 features	MPt.LogP.HD	MPt.SP.HA	37.59 ± 8.54	0.64 ± 0.13	10.77 ± 11.52	0.91 ± 0.14	0.00	0.00	Y	Y
4 features	MW.MPt.log P .HD	MW.MPt.SP. HA	37.40 ± 7.56	0.65 ± 0.15	9.36 ± 11.20	0.93 ± 0.18	0.00	0.00	Y	Y
5 features	MW.MPt.SP. Log P.HD	MW.MPt.SP. HA.HD	37.35 ± 7.23	0.65 ± 0.14	6.90 ± 13.33	0.96 ± 0.19	0.00	0.00	Y	Y
6 features	MW.MPt.SP. Log P.HA.HD	MW.MPt.SP. Log P.HA.HD	31.61 ± 10.70	0.71 ± 0.15	3.47 ± 14.24	0.99 ± 0.20	0.00	0.00	Y	Y

Table 5.

Model 1	Model 2	Model 1		Model 2		P-value		Significance difference
		ION (%) ± SD	NMSE ± SD	ION (%) ± SD	NMSE ± SD	ION (%)	NMSE	
GPR Mpt. LogP.HD	SLN MPt.HA	37.59 ± 8.54	0.64 ± 0.13	11.23 ± 11.29	0.91 ± 0.13	0.00	0.00	Y
GPR MPt. LogP.HD	Luo et al., (2007)	37.59 ± 8.54	0.64 ± 0.13	-53.25 ± 36.24	1.56 ± 0.35	0.00	0.00	Y
SLN MPt.HA	Luo et al., (2007)	11.23 ± 11.29	0.91 ± 0.13	-53.25 ± 36.24	1.56 ± 0.35	0.00	0.00	Y

LEGENDS FOR FIGURES AND TABLES

Table 1. Performance measures and statistical comparisons of QSPR and corresponding machine learning methods models (in all cases, 10 evaluations were carried out in order to measure standard deviations and P-values for each parameter). Highlights indicate the models with the best performance measured values within each category. Statistical analysis relates to p-values for t-tests carried out between the performance measures of each QSPR model and corresponding Single Layer Network Machine Learning models. Table 2. Summary of the statistical analysis of the comparisons of Gaussian Process models with different combinations of physicochemical descriptors.

Table 3. Statistical performance measures of the best-performing models, and significance of molecular descriptors employed in the Gaussian Process models.

Table 4. Statistical analysis (paired t-test) between the Gaussian Process and Single Layer Network models with the highest ION (%) in each number of molecular descriptor categories.

Table 5. Statistical analysis of the best models obtained by Gaussian Process, Single Layer Network and QSPR methods.

Figure 1. Comparison of the improvement over the naïve model for (a) machine learning methods (GP and SLN) compared with a range of QSPR models that relate the permeability of a penetrant to log P and MW and (b) with the QSPR model proposed by Barratt (1995).

Figure 2. Comparison of the predictive ability of Gaussian Process models with the QSPR model proposed by Potts and Guy (1992) across a wide range of lipophilicities. Data points shown are obtained from a subset of the overall dataset (as described in the Methods section). The test set shown is that which results in the Potts and Guy (1992) model achieving the best performance among the ten test sets generated by this analysis.