ORIGINAL RESEARCH

The application of Gaussian processes in the predictions of permeability across mammalian and polydimethylsiloxane membranes

Yi Sun [1], Rod Adams [1], Neil Davey [1], Gary P. Moss [2], Maria Prapopopolou [3], Marc B. Brown [4], Gary P. Martin [3], Simon C. Wilkinson [5]

1. School of Computer Science, University of Hertfordshire. UK. 2. School of Pharmacy, Keele University, UK. 3. Pharmacy Department, King's college London, UK. 4. MedPharm Ltd, Surrey research park, Guildford, UK. 5. Medical Toxicology Centre, Wolfson Unit, Medical School, University of Newcastle-upon-Tyne, UK.

Correspondence: Yi Sun. Address: School of Computer Science, University of Hertfordshire, College Lane, Hatfield, UK, AL10, 9AB. Email: comrys@herts.ac.uk.

Received: May 30, 2012 **Accepted:** August 20, 2012 **Published:** December 1, 2012
DOI: 10.5430/air.v1n2p86 **URL:** http://dx.doi.org/10.5430/air.v1n2p86

Abstract

The problem of predicting the rate of percutaneous absorption of a drug is an important issue, particular with the increasing use of the skin as a means of moderating and controlling drug delivery. One key feature of this problem domain is that human skin permeability to penetrants (often characterised by K_p, the permeability coefficient) has been shown to be inherently non-linear when mathematically related to the key physicochemical parameters of penetrants. The aims of the current study were to apply and validate Gaussian process regression methods to datasets for membranes other than human skin, and to explore how the nature of the dataset may influence its analysis. Permeability data for absorption across rodent and pig skin, and polydimethylsiloxane *Silastic*® membranes was collected from the literature. Two QSPR methods were applied to compare to the Gaussian process models. The results demonstrated that Gaussian process models with different covariance functions outperform the QSPR model for human, pig and rodent datasets, but in general are not good for *Silastic*® membranes. These results suggest that the physicochemical parameters employed in this study might not be appropriate for developing models that represent this membrane. In addition, the results show the size of the datasets, in both absolute and comparative senses, appears to influence model quality.

Key words

Gaussian processes, Skin permeability, Mammalian, Polydimethylsiloxane membranes

1 Introduction

The problem of predicting the rate at which various chemical compounds penetrate human skin is an important issue with the increasing use of skin as a means of achieving both local and systemic drug delivery. The *permeability coefficient* for human skin is measurable, but making such measurements is expensive in terms of cost and time. However it is possible to predict the permeability from other well-known characteristics of the chemical whose permeability is being investigated, such as its molecular weight or its solubility.

In Moss et al. and Sun's studies [1, 2], it is shown that advanced machine learning techniques, especially, *Gaussian Processes* (GP), outperform quantitative structure-activity relationships (QSARs), which are widely used in the pharmacy community.

Data from human skin is relatively hard to obtain, whereas data from other animals is much more available in the pharmaceutical industry. The main motivation in this paper is to find out if data from non-human skin can act as a good predictor, of the actual permeability of human skin. We investigate how the addition of non-human data to the original human data may help in the training of the predictor.

Hence in this paper, a GP is evaluated on four different datasets, namely experimentally derived drug permeation data across human skin, pig skin, rodent skin, and a synthetic *Silastic*® membrane.

One key feature of predicting percutaneous absorption accurately is that the target, the skin permeability coefficient, may have a strongly non-linear relationship with the compound physicochemical descriptors (features). This has already been shown to be the case [1, 2], using a human skin dataset. In this work, it will also be shown that a non-linear relationship exists with pig and rodent skins (see section 5.1).

In this paper it is demonstrated empirically that adding rodent data to human data as a mixed training set, can give reasonably accurate predictions, in fact, similar results to those obtained using the same sized dataset comprising human skin data alone.

2 Problem domain

Predicting percutaneous absorption accurately has proven to be a major challenge and one which has substantial implications for the pharmaceutical and cosmetic industries, as well as toxicological issues in fields such as pesticide usage and chemicals manufacture. Predictive modeling is a frequently used tool to increase the throughput of percutaneous absorption experiments. The use of animal models for percutaneous penetration is often considered essential, given the possible toxicity, cost, ethics and inconvenience of employing human skin during laboratory experiments. Human skin differs from that of many animals in numerous ways including the thickness of the stratum corneum, number of appendages per unit area and amount of skin lipids present. Despite this, it is very surprising that no quantitative mathematical models have been developed for the purpose of characterising permeation across non-human skin. This is perhaps due to the development of the Potts and Guy model [3], the first major, quantitative model for measuring percutaneous absorption, which was based on human skin data. This paper, therefore, documents the first development of quantitative models for predicting percutaneous absorption across animal skin, and should allow a more accurate comparison to be made between drug permeation across human and various animal skins.

In using a model system, the researcher must take into account the inherent differences of the various species employed and the parameters affecting percutaneous penetration in each species. The model selected must therefore resemble human skin as closely as possible. Various models have been offered by many researchers. Bartek, et al. [4] and Wester and Noonan [5] investigated several potential models, including rabbit, miniature swine and rat, and concluded that rabbit skin, and then rat skin, were the most permeable membranes, and that flux denoted as J, that is the rate of permeate transfer from one side to the other, through pig skin most resembled that of the permeation across human skin. Many other workers have also indicated the suitability of porcine (pig) skin, especially from weanling or stillborn animals, as a model for percutaneous absorption [6-10].

Further, synthetic membranes can often be chosen as a means of measuring across a lipophilic barrier. Poly (dimethylsiloxane) (Silastic®, Silescol®) is a widely employed model membrane. It has demonstrated good agreement with Fick's first law of diffusion [11, 12], which relates (in the context of skin absorption) to the passage of a substance from one

side of a membrane (skin, synthetic, etc.) to the other. It has been experimentally determined that flux (J) increases when the concentration gradient (that is, the relative amount of concentration above and below the membrane) is as large as possible.

By convention K_p denotes the permeability coefficient. K_p is a concentration corrected version of flux that allows comparison of permeation for different molecules. K_p is defined as $K_p = J / \Delta C_m$, where ΔC_m denotes the concentration difference across the membrane. 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).

These approaches have been extensively reviewed [13]. Usually, lipophilicity (*P*) and molecular weight (*MW*) appear to be the only significant features in QSAR forms, although subset analysis has shown the significance of other parameters [13]. *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 *log* scale, log *P*, is used to simplify the notation in common use. For the same reason log K_p is used for skin percutaneous absorption rather than K_p. It is important to note that log K_p is a completely different term to log *P*.

Recently, new approaches, for example, artificial neural networks and fuzzy modelling, have been applied to predict percutaneous absorption. Moss [1] has employed Gaussian Processes to predict percutaneous absorption using a human skin dataset. This study showed the underlying non-linear nature of the dataset, and provided a substantial statistical improvement over existing models.

As such the aim of the current study was to validate the Gaussian Process regression model in an attempt to better predict percutaneous absorption on experimentally derived drug permeation data across human skin, pig skin, rodent skin, and a synthetic membrane including polydimethylsiloxane (Silastic®), a membrane widely used as a substitute for skin either in preliminary studies or where skin is not available.

3 Description of datasets employed

The four datasets employed in this study have been collated with reference to a range of literature sources. The human, pig, rodent, and synthetic membrane datasets consist of 140, 15, 103 and 19 chemical compounds, respectively. Among these four datasets, there are common chemical compounds tested for different membranes. For example, *caffeine* was tested through human, rodent and the synthetic membrane. The log K_p *(cm/h)* value for human skin is -3.68, for rodent is -2.99, and for the synthetic membrane is -1.70. Table 1 shows the number of chemical compounds in each dataset that are also present in the human dataset.

Table 1. The number of chemical compounds common to human and various animal and synthetic membranes, where the total number of chemical compounds in the human skin is 140

Datasets	Common compounds	Non-common compounds
Pig	3	12
Rodent	48	55
Synthetic	7	12

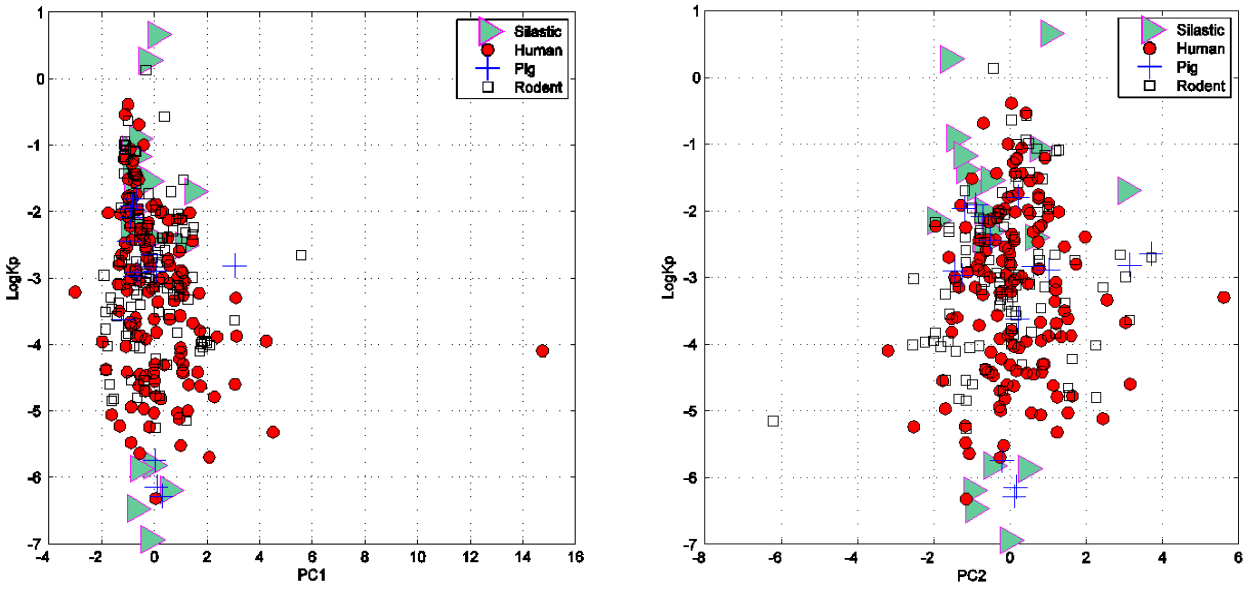
In Moss and Sun's studies [1, 2], it is shown that using the following five features: *molecular weight* (*MW*), *solubility parameter* (*SP*), *log P*, counts of the number of hydrogen bonding acceptor (*HA*) and donor groups (*HD*), can produce better predictions when compared to using only lipophilicity and molecular weight alone. Therefore, in this work, these five compound features are used.

Visualisation of the skin data

In order to visualise the underlying distribution of the data we perform a principal component analysis and plot the data in the first two principal components; this is done just to visualise the data and is not used in the actual data analysis.

First all the compounds from the four datasets were combined. The data was then normalised so that all five features had a zero mean and unit variance. The normalised data points were visualised by applying principle component analysis (PCA), which maps data to a low-dimensional space with a linear transformation.

The compounds were plotted using the corresponding log K_p values against the first two principal components to represent the variation in the five features of all chemical compounds (see Figure 1). The first principal component accounts for 43.0% of the total variance, and the second accounts for 31.4%. Figure 1 shows that there is no linear relation between log K_p and the compound features. It suggests that there may be more complex non-linear structures in the data. In addition, it can be seen that the rodent skin dataset has a similar distribution to the human skin dataset. As for the pig and synthetic datasets, it can be seen that all log K_p values are outside the range [-5.5, -3], except the one chemical from the pig skin dataset which has a log K_p value about -3.6, while the human dataset ranges between [-6.5, 0]. This has important implications for the quality of the models trained from these datasets.



(a) (b)

Figure 1. The relationship between log K_p and the PCA space of chemical compounds: a) the first principal component; b) the second principal component

4 Methods

As already stated QSARs are the standard methods for predicting permeability in the Pharmaceutical industry. Two QSAR methods were applied to the human skin data in order to provide a comparison between Gaussian Processes and previous approaches to this task. The first one, denoted as *Potts*, was proposed by Potts and Guy [3] and derived from the Flynn dataset [14]. It is given by the equation $\log K_p\,(cm/s) = 0.71 \log P - 0.0061\,MW - 6.3$. The second model, denoted as *Moss*, is represented by $\log K_p\,(cm/s) = 0.74 \log P - 0.0091\,MW - 2.39$, which was derived from a slightly larger dataset [13].

As a baseline since there are no QSAR models used for animal skin, a simple *naïve model* and a simple linear regression model were used for comparison. In the naïve model, for any input the prediction is always the same value, namely the mean of log K_p in the training set. Any model that cannot perform better than this will obviously be of little use.

Sections 4.1 – 4.3 describe the trainable regressors we have used and Section 4.4 gives the performance measures used to evaluate the models.

4.1 Simple linear regression

This simple linear regression (*LR*) considers the output y as the weighted sum of the components of an input vector $\mathbf{x}$, which can be written as follows [11]:

$$y = y(\mathbf{x}, \mathbf{w}) = \sum_{d=1}^{D} w_d x_d + w_0 \tag{1}$$

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

4.2 The Gaussian processes regression

In earlier work [2] we found that Guassian Processes performed at least as well as any other trainable model from machine learning on this data, so GPs are the principle method employed here.

A Gaussian process (GP) is defined simply as a collection of random variables which have a joint Gaussian distribution. It is completely characterised by its mean and covariance function. Usually, the mean function is considered to be the zero everywhere function. The covariance function, $k(\mathbf{x}_i, \mathbf{x}_j)$, allows for specifying a-priori knowledge from a training dataset. It defines nearness or similarity between the values of $f(\mathbf{x})$ at the two points $\mathbf{x}_i$ and $\mathbf{x}_j$.

To make a prediction y_* at a new input $\mathbf{x}_*$, the conditional distribution $p(y_* \mid y_1, ..., y_N)$ on the observed vector $[y_1, \ldots, y_N]$ is needed to be computed. Since the model is a Gaussian process, this distribution is also a Gaussian and is completely defined by its mean and variance. By applying standard linear algebra, the mean and variance at $\mathbf{x}_*$ are given by:

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

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

where $\mathbf{k}_*$ denotes the vector of covariances between the test point and the N 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; y denotes the vector of training targets; and $k(\mathbf{x}_i, \mathbf{x}_j)$ denotes the variance of y_*. As is normally the case, mean values were used as predictions, and the variance was used as error bars on the prediction.

4.3 Covariance functions

When using a GP it is necessary to choose a suitable covariance function and here we investigate those that are most widely used in machine learning and engineering. Hence the squared exponential covariance function, the neural network covariance function, the rational quadratic covariance function, and two members from the Matérn class of covariance function [15] are applied. In each case, an independent noise contribution is incorporated into the covariance function.

4.3.1 The squared exponential covariance function

The squared exponential covariance function (denoted as *SE*), as follows:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T M(\mathbf{x}_i - \mathbf{x}_j)) \tag{4}$$

where $M = l^{-2}\mathbf{I}$, l is *characteristic length-scale*, and σ_f is *signal variance*. In this work, the squared exponential covariance function is computed with isotropic distance measure.

4.3.2 The neural network covariance function

This function is so named because it was shown by Neal (1996) to be equivalent to a feed-forward neural network with a single hidden layer (denoted as NN), in the limit of an infinite number of hidden units. It can be written:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \alpha \sin^{-1}\left(\frac{\hat{\mathbf{x}}_i^T \beta \hat{\mathbf{x}}_j}{\sqrt{(1+\hat{\mathbf{x}}_i^T \beta \hat{\mathbf{x}}_i)(1+\hat{\mathbf{x}}_j^T \beta \hat{\mathbf{x}}_j)}}\right) \tag{5}$$

where α and β are scalar hyperparameters to be optimised, and $\hat{\mathbf{x}}$ is the $\mathbf{x}$ vector extended by appending an element with the value 1. In this work, the neural network covariance function with a single parameter for the distance measure.

4.3.3 The rational quadratic covariance function

The rational quadratic covariance function (*QR*) with isotropic distance measure is given by

$$k(\mathbf{x}_i, \mathbf{x}_j) = \left(1 + \frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\alpha l^2}\right)^{-\alpha} \tag{6}$$

where α and l, the characteristic length-scale, are non-negative parameters of the covariance function.

4.3.4 The Matérn class of covariance function

In this work, two simple cases of the Matérn covariance function with isotropic distance measure were used. The Matérn covariance function can be defined as a product of an exponential and a polynomial of order p are considered. $p = 1$ (Matérn3) and $p = 2$ (Matérn5) are set, separately.

4.4 Performance measures

As mentioned in Moss, et al. and Sun, et al. [1, 2], mean squared error (*MSE*), improvement over Naïve (*ION*), negative log estimated predictive density (*NLL*), and Pearson correlation coefficients (*CORR*) were all used to evaluate the performance of each model.

The *MSE* measures the average squared difference between model predictions and the corresponding targets. The *ION* measures the degree of improvement of the model over the Naïve predictor, whose value is always the same value, namely the mean of log K_p in the training set. Thus, the *ION* can be computed as:

$$ION = \frac{MSE_{naïve} - MSE}{MSE_{naïve}} \times 100\% \tag{7}$$

The *NLL* is defined as:

$$NLL = \frac{1}{N_{tst}} \sum_{n=1}^{N_{tst}} -\log p(y_n \mid \mathbf{x}_n) \tag{8}$$

where

$$-\log p(y_n \mid \mathbf{x}_n) = \frac{1}{2}\log(2\pi\sigma_*^2) + \frac{(y_n - E[y_n])^2}{2\sigma_*^2} \tag{9}$$

In which case σ_*^2 is the predictive variance. The *CORR* measures the correlation between predictions and targets. For comparison, a model should have low values of both MSE and *NLL*, as well as high values of both *ION* and the correlation coefficient (*CORR*) on a given test dataset.

5 Experiments

5.1 Experiment 1

The first experiment investigated whether it was possible to produce a good regression model for both human and non-human skin.

A linear regressor and a Gaussian Process regression model were compared for pig and rodent membranes as well as synthetic membranes. The most suitable covariance functions are investigated for predicting skin percutaneous absorption. A GP model with the different covariance functions described in Section 4.3 was applied for each single skin dataset, respectively. In addition, the QSAR models and Moss and Cronin [13] were compared with GPs for the human skin dataset. For each different dataset, the *leave-one-out* technique was applied, that is, one chemical is used for testing, and all others are employed for training. This was repeated for each compound in turn. Finally, performance metrics were computed in terms of all predictions.

Tables 2-5 show results of the human, pig, rodent and synthetic membrane datasets, separately. In each table the best result for each column is indicated in bold. It can be seen that predictions of skin percutaneous absorption obtained from GPs with different covariance functions outperform predictions from QSARs on the human dataset and outperform naïve and LR on all of the datasets. The covariance *NN* works better than the other covariances on both human and synthetic skin dataset, and *Matérn3* outperforms the others on both pig and rodent datasets. For the synthetic skin dataset, GP with the covariance *NN* had the best performance, while *RQ* gave a worse result than the naïve model, and the others had a similar result to the naïve model.

Furthermore, one can see that the simple LR has a similar result to the naïve model on both human and rodent datasets, but a worse result on the synthetic dataset, and the LR gives the worst result on the pig dataset. A further investigation was undertaken using only two features, which were *MW* and $\log P$ with LR on the pig dataset. In this case, the results are $MSE = 2.45$, $ION = 1.89$ and $CORR = 0.35$. Interestingly, if only these two features were applied with GP using *Matérn3* on the pig dataset, one can obtain $MSE = 0.46$, $ION = 81.63$, $CORR = 0.90$ and $NLL = 1.69$, which in general is better than the best results in Table 3 except for a higher NLL value. However, GP with 5 features still gives better results than GP with 2 features on human, rodent and synthetic datasets as shown in Table 6, where results for 5 features are taken from the best results in Tables 2 to 5.

Table 2. Leave-one-out results on human skin dataset

Model		MSE	ION	CORR	NLL
QSAR	Moss	20.09	-1157.60	0.14	-
	Potts	5.50	-244.53	0.10	-
Naïve		1.60	0	-1	-
LR		1.51	5.36	0.28	-
GP	NN	**1.13**	**29.14**	**0.53**	**1.48**
	SE	1.23	23.13	0.49	1.53
	RQ	1.13	29.06	0.53	9.42
	Matérn3	1.20	25.08	0.51	9.43
	Matérn5	1.21	24.26	0.50	9.43

ISSN 1927-6974 E-ISSN 1927-6982

Table 3. Leave-one-out results on pig skin dataset

Model		MSE	ION	CORR	NLL
Naïve		2.50	0	-1.00	-
LR		19.42	-678.12	0.04	-
GP	NN	0.59	76.52	0.86	1.65
	SE	0.64	74.45	0.84	20.65
	RQ	0.73	70.81	0.82	21.51
	Matérn 3	**0.51**	**79.74**	**0.88**	**1.08**
	Matérn 5	0.62	75.30	0.85	1.76

Table 4. Leave-one-out results on rodent skin dataset

Model		MSE	ION	CORR	NLL
Naïve		1.30	0	-1.00	-
LR		1.29	0.94	0.15	-
GP	NN	0.88	32.63	0.56	1.41
	SE	0.86	34.39	0.58	1.40
	RQ	0.86	34.07	0.58	1.41
	Matérn 3	**0.83**	**36.25**	**0.60**	**1.38**
	Matérn 5	0.84	35.70	0.59	1.39

Table 5. Leave-one-out results on the synthetic membrane dataset

Model		MSE	ION	CORR	NLL
Naïve		5.79	0	-1.00	-
LR		7.17	-23.92	0.35	-
GP	NN	**3.57**	**38.26**	**0.60**	**2.03**
	SE	5.45	5.84	0.23	2.52
	RQ	6.33	-9.34	-0.70	2.96
	Matérn 3	5.55	4.15	0.08	2.72
	Matérn 5	5.19	10.37	0.22	2.65

Table 6. Leave-one-out results with 5 features and 2 features

Dataset	Model		MSE	ION	CORR	NLL
Human	GP	5 features	1.13	29.14	0.53	1.48
	NN	2 features	1.25	21.72	0.45	1.54
Pig	GP	5 features	0.51	79.74	0.88	1.08
	Matérn 3	2 features	0.46	81.63	0.90	1.69
Rodent	GP	5 features	0.83	36.25	0.60	1.38
	Matérn 3	2 features	1.14	12.72	0.34	1.50
Synthetic	GP	5 features	3.57	38.26	0.60	2.03
	NN	2 features	6.48	-11.88	-0.32	2.42

5.2 Experiment 2

The second experiment investigates whether a GP model trained using just non-human skin dataset provides reasonable predictions for the human skin dataset. This is potentially important as it is obviously much easier to obtain animal tissue

than human tissue. The pig and the rodent skin datasets were used as the training set, separately, and the trained GP model was tested on the whole human skin dataset. In this experiment, the *NN* covariance function was used in the GP. Note there were only 15 compounds in the pig dataset.

Table 7. Performances on the whole human skin dataset when trained on the pig or rodent dataset

Model		MSE	ION	CORR	NLL
QSAR	Moss	20.09	-1171.00	0.14	-
	Potts	5.50	-248.22	0.10	-
Pig trained on pig	Naive	1.58	0	0.00	-
	GP: Matérn 3	1.70	-7.39	0.34	3.97
Rodent trained on rodent	Naïve	1.63	0	0.00	-
	GP: Matérn 3	**1.24**	**24.14**	**0.56**	**2.61**

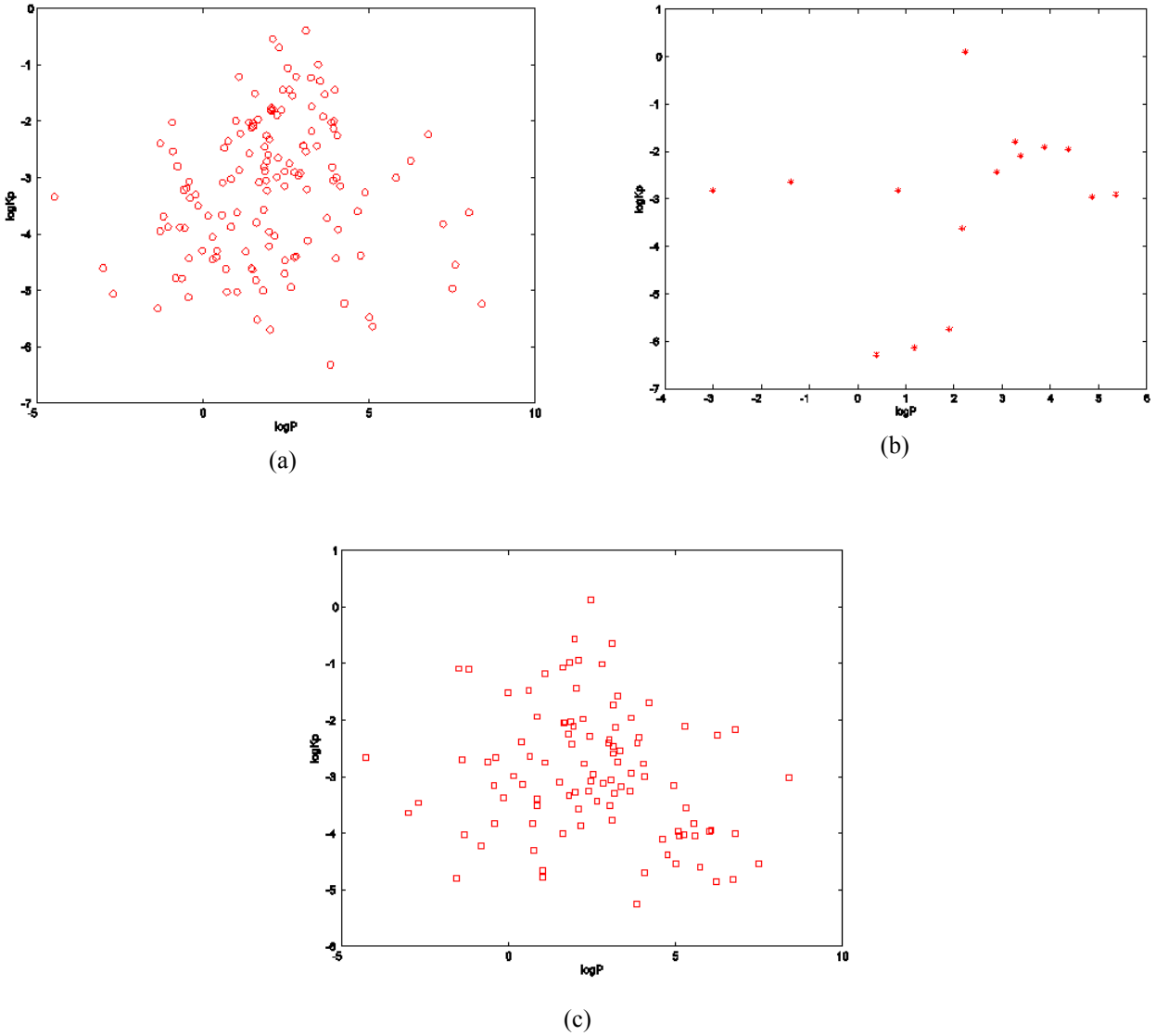


Figure 2. Permeability coefficients as a function of log P on the a) human b) pig c) rodent

Table 7 shows results for the whole human dataset. Again the best results are shown in bold. It can be seen that the GP models outperform the QSAR models, with the model trained using the rodent dataset giving the best prediction result. It shows that the GP model trained using the pig dataset has a worse result than the pig naïve model. This may be because the number of examples in the pig dataset is very small. Figure 2 ((a) to (c)) shows the skin permeability coefficient from human, pig and rodent datasets against $\log P$ respectively. It shows that examples from the pig dataset, plot (b), do not cover the whole range of feature and target values of the human skin dataset, plot (a), possibly due to the limited number of examples in this dataset. For example, for the area of values of $\log K_p$ in the range of [-5.5, -4] and values of $\log P$ in the range of [-1, 2], one can see that there are no pig examples in plot (b), while a relative large number of examples can be found in the corresponding human plot (a). Since a GP works by using a weighted average off near data points to extrapolate the prediction for new data points, areas where there are no near data points in the training set will elicit poor predictions. It can be seen that the rodent data, plot (c), has a similar data distribution to the human data, plot (a). This may be the reason why training on the rodent dataset gives the best prediction result on the human dataset. Since values of $\log P$ in the pig set are all less than 6, a couple of data points with high values (≥ 10) of $\log P$ from the human and rodent datasets are not shown in the corresponding plots.

5.3 Experiment 3

The possibility of using an animal model to predict human skin permeability was investigated in experiment 2. In order to make a direct comparison between a regressor trained on human data and one trained on non-human it was necessary to find the compounds that were in both the human and rodent data sets and use these (that is common chemical compounds) as the training set for the respective models.

Hence in this experiment, a quantitative comparison of human skin permeability predictions between trained GP models on a rodent and a human training dataset was undertaken. To make this comparison, the compounds that are common in the human and rodent dataset had to be used. A training set including 48 common chemical compounds for which a target value were known for both the human and rodent data was taken. Two trained models were then produced. Previously unseen human data were taken as a test set for both models. This used the rest of the human dataset, including 92 compounds. A GP with the Matérn3 kernel was used with the data containing all five features.

Table 8 shows the corresponding results with the best results in bold. It can be seen that both the human and rodent training sets give better predictions on the human test set than using either the naïve or the QSAR models, where the human training set has the best performance. Nevertheless the model trained on the rodent is very nearly as good and has the same *NLL* measure. Interestingly, it can be seen that the rodent naïve model is better than the human naïve model for predicting the human skin permeability. Moreover, GP predictions from the rodent model are much better than the human naïve model. This means that the rodent model could be more useful than often thought for predicting human skin permeability [5]. Note that QSAR's *ION* results were from the human training set.

Table 8. Performances on the human test set using model trained on rodent and human training sets, separately

Model		MSE	ION	CORR	NLL
QSAR	Moss	19.35	-1203.3	0.16	-
	Potts	6.01	−304.59	0.12	-
trained on human	Naïve	1.48	0	0	-
	GP	**1.05**	**29.40**	**0.52**	**1.47**
trained on rodent	Naïve	1.37	0	0	-
	GP	1.13	17.22	0.46	1.47

5.4 Experiment 4

All the previous experiments have used training data from a single species. The final experiment investigated how adding rodent examples into a human training set may affect predictions on a human test set. To avoid inconsistent training examples, that is, examples with the same features but different target values, the non-common compounds were used as

training examples. In this experiment, a model was trained using the 92 non-common compounds, denoted as *trnH*, and tested using human data with the 48 common compounds.

Results produced by the human model are shown in the first two rows of Table 9.

Table 9. Performances on the human test set using models trained on the human, rodent and mixed training set, respectively. Performance using rodent values directly is also shown (see text)

Model		MSE	ION	CORR	NLL
trained on human	Naïve	2.15	0	0	-
	GP	1.93	9.96	0.43	13.06
trained on mixed	GP	1.91 ± 0.09	11.78 ± 3.05	0.51 ± 0.05	2.10 ± 0.31
Actual rodent values		1.51	29.47	0.68	–

To generate a mixed model using human and rodent data, a mixed dataset was needed. Moreover, a training set of size 92 compounds was needed in order to compare it with *trnH* including 92 compounds. Chemical compounds which are not included in those common compounds were extracted from the rodent dataset, in total 55 chemical compounds, denoted by *trnR*. Then 50 samples from *trnH*, and 42 from *trnR* were randomly selected, and added together, denoted as *trnHR*. A GP model was trained using *trnHR*. Finally, the model was tested for the same human data as was used for the human model (that is 48 compounds). This procedure was repeated 10 times, and average results are shown in Table 9.

It can be seen that including rodent examples in the training set can produce predictions on average almost as accurate as using a human training set of the same size. In fact, the model trained on this mixed training set produces the best performance on the *NLL* measure.

Finally, the actual value of permeability for rodent skin was used as the prediction value for the human skin. This works since the test set contains only those compounds that are common to both rodent and human tissues. The final row of Table 9 shows the result of this `prediction'. Interestingly, it gives the best result.

6 Discussion

In general, the GP works considerably better than the both QSAR models and the *naïve p*redictions applied to the human and animal skin dataset. The QSARs employ linear representations of the data, where weights are fixed for each feature on different compounds.

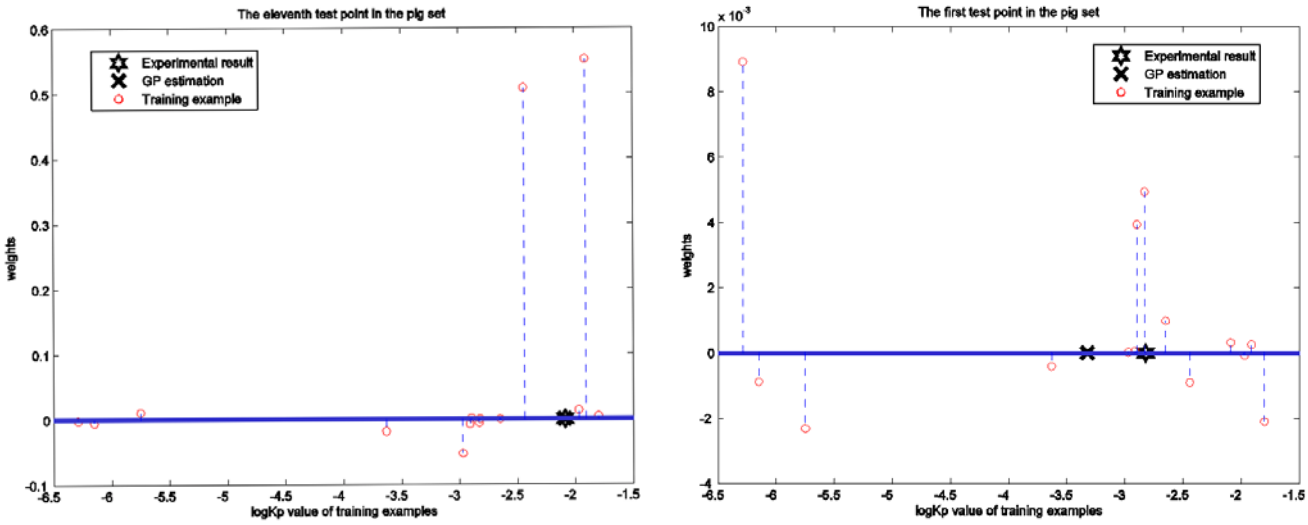
Looking at eq. (2), a GP prediction may be considered as a linear combination of observation target values, that is $E[y_*] = \sum_{i=1}^{n} w_i y_i$, where each weight ($w_i$) is given by the corresponding entry in the vector of $\mathbf{k}_*^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}$ (Rasmussen and Williams (2006)). Consider an extreme case, where $(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}$ is a diagonal matrix (K_{ii} denotes each entry on the diagonal line), as if there is no similarity between any pair of training examples. In this case, the weight is determined by the similarity between the test point $\mathbf{x}_*$ and each specific training example $\mathbf{x}_i$, divided by a real number, that is, $K_{ii} + \sigma_n^2$. Thus, the more similar a vector is to a training example, the larger is the corresponding weight. However, training examples usually have non-zero covariances among them. Therefore, each weight is a measure of similarity of the test point to a specific training example, and then adjusted by the covariance matrix over the whole training set.

Two test points are randomly selected from the pig dataset. In each case, there are 14 training examples. Figure 3 shows log K_{p} (that corresponds to y_i) of training examples against weights (w_i) for each test point. The star sign and cross sign on the zero weights line show the actual experimental measure of log K_{p} and GP estimation, respectively, for the corresponding test point.

When the training examples next to the test point (according to log K_p values) have large positive weights, the GP estimation can produce a relatively accurate prediction; for example, see Figure 3 (a). On the other hand, when the training examples far away from the test point have large positive weights, the accuracy of the GP estimation can deteriorate, for example, see Figure 3 (b).

Note that weights (similarities) are measured on those molecular features through covariance functions. Therefore, the selection of both molecular features and covariance functions is a really important issue.

Moreover, the fact that the linear model and GP have worse results with 5 features on the pig dataset may suggest more data are needed. The nature of the pig skin dataset, most notably its size, significantly impacts on not just the quality of the pig skin Gaussian process model, but it also puts the findings of the other models into context. For example, the clear `false positive' of this study is that the rodent skin dataset produces a more accurate model than pig skin. While this would suggest that rodent skin is more representative of human skin than pig skin, there is a vast body of literature, from [5] onwards, that presents experimental evidence that this is not the case. The work of Potts and Guy [3, 16] raises an important issue in the accuracy of small datasets. Their models being substantially different in that the first model [3] focused on the whole Flynn dataset [14], whereas the second [16] used only the thirty-seven non-electrolytes from that dataset. The inclusion of a deliberately small dataset in this study clearly shows the impact that the volume of data can have on the quality of the model, and presents quite clearly the possibility of developing a misleading model, which reflects the comments on model design and quality by Cronin and Schultz [17]. Further, while it may seem appropriate to collate datasets together and produce larger datasets, such collations may not necessarily result in better models, given the underlying distribution of the data and the physicochemical parameters represented. Hence, an efficiency in dataset design, including an even and representative spread of data, avoiding data redundancy, may be more important to dataset quality than simply adding all the new percutaneous absorption data as it appears in the literature.



(a) (b)

Figure 3. Weights against $\log K_p$ values of pig training examples

7 Conclusions

The results of this study show that, in general, GP methods produce better results, including better predictions of experimental targets and statistical performance measures, than QSPR models and naïve predictions when applied to the human and animal skin datasets employed in this study. They suggest that the Gaussian process models produce better

results, both in a statistical sense, and in terms of the accuracy of prediction, than the QSAR models used to benchmark them. The results produced are, as in previous studies, consistent with an underlying non-linearity in the dataset [1] and Lam, et al. [18]. This is particularly evident at the extremes of the model.

While the rodent dataset provides a better model – in these experiments – than the pig skin model, it is still not as accurate as the human skin model. Clearly, a larger pig skin dataset might improve the accuracy of that model, relative to the rodent skin model. The results of this study indicate that permeation across animal (rodent and pig) skin is, in a statistical sense, similar. It also suggests that the synthetic skin is a poor membrane to use in place of human or animal skin. However, the overriding issue raised in this study is the nature of a dataset and how it can influence the results, and interpretation, of any model produced. The size of the datasets appears to influence model quality, producing counterintuitive results that simply do not agree with a large literature of laboratory-generated experimental data. Nevertheless we have shown that data from non-human skin can provide useful information in the prediction of the permeability of human skin.

References

[1] Moss G, Sun Y, Davey N, Adams R., Pugh W, Brown M. The application of Gaussian processes to the prediction of percutaneous absorption. Journal of Pharmacy & Pharmacology. 2009; 61: 1147-1153. PMid:19703363 http://dx.doi.org/10.1211/jpp.61.09.0003

[2] Sun Y, Moss G, Prapopoulou M, Adams R, Brown M, Davey N. Prediction of skin penetration using machine learning methods, in: Proceedings of ICDM. Pisa. 2008.

[3] Potts R, Guy R. Predicting skin permeability. Pharmaceutical Research. 1992; 9: 663-669. PMid:1608900 http://dx.doi.org/10.1023/A:1015810312465

[4] Bartek M, LaBudde J, Maibach H. Skin permeability in vivo: comparison in rat, rabbit, pig and man. Journal of Investigative Dermatology. 1972; 58: 114-123.PMid: 4622425 http://dx.doi.org/10.1111/1523-1747.ep12538909

[5] Wester R, Noonan P. Relevance of animal models for percutaneous absorption, International Journal of Pharmaceutics. 1980; 7: 99-110. http://dx.doi.org/10.1016/0378-5173(80)90054-X

[6] Marzulli F, Maibach H. Animal models in dermatology. Churchill- Livingstone, Edinburgh, Ch. Relevance of animal models: the hex-achlorophene story. 1975.

[7] Chow C, Chow A, Downie R, Buttar H. Percutaneous absorption of hexachlorophene in rats, guinea pigs and pigs. Toxicology. 1978; 9: 147-154. http://dx.doi.org/10.1016/0300-483X(78)90039-2

[8] Bronaugh R, Stewart R, Congdon E, Giles A. Methods for in vitro per-cutaneous absorption studies i: Comparison with in vivo results. Applied Pharmacology. 1982; 62: 481-488. http://dx.doi.org/10.1016/0041-008X(82)90149-1

[9] Roberts M, Mueller K. Comparisons of in vitro nitroglycerin (tng) flux across yucatan pig, hairless mouse and human skins, Pharmaceutical Research. 1990; 7: 673-676. PMid:2114619 http://dx.doi.org/10.1023/A:1015842916969

[10] Sato K, Sugibayashi K, Morimoto Y. Species differences in percutaneous absorption of nicorandil, Journal of Pharmaceutical Sciences. 1991; 80: 104-107. PMid:1828835 http://dx.doi.org/10.1002/jps.2600800203

[11] Bishop C. Neural Networks for Pattern Recognition. New York: Oxford University Press; 1995.

[12] Woolfson A, McCafferty D, Moss G. Development and characterisation of a moisture-activated bioadhesive drug delivery system for percutaneous local anaesthesia, International. Journal of Pharmaceutics. 1998; 169: 83-94. http://dx.doi.org/10.1016/S0378-5173(98)00109-4

[13] Moss G, Cronin M. Quantitative structure-permeability relationships for percutaneous absorption: re-analysis of steroid data. International Journal of Pharmaceutics. 2002; 238:105-109. http://dx.doi.org/10.1016/S0378-5173(02)00057-1

[14] Flynn G L. Physicochemical determinants of skin absorption. New York: Elsevier; 1990; 93-127.

[15] Rasmussen C, Williams C. Gaussian Processes for Machine Learning. The MIT Press. 2006.

[16] Potts R, Guy R. A predictive algorithm for skin permeability: the effects of molecular size and hydrogen bond activity. Pharmaceutical Research. 1995; 12: 1628-1633. PMid:8592661 http://dx.doi.org/10.1023/A:1016236932339

[17] Cronin M T D, Schultz T W, Pitfalls in QSAR. Journal of Molecular Structure: THEOCHEM. 2003; 622: 39-51. http://dx.doi.org/10.1016/S0166-1280(02)00616-4

[18] Lam LT, Sun Y, Davey N, Adams R, Prapopoulou M, Brown M B, Moss G P. The application of feature selection to the development of Gaussian Process models for percutaneous absorption. Journal of Pharmacy and Pharmacology. 2010; 62: 738 -749. PMid:20636861