View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Hertfordshire Research Archive

University of Hertfordshire UH 25 Research Archive

**Citation for published version:**
Alpa Shah, Yi Sun, Rod G. Adams, Neil Davey, Simon C. Wilkinson, and Gary P. Moss, 'Support vector regression to estimate the permeability enhancement of potential transdermal enhancers', *Journal of Pharmacy and Pharmacology,* Vol. 68 (2): 170-184, February 2016.

**DOI:**
http://dx.doi.org/10.1111/jphp.12508

**Document Version:**
This is the Accepted Manuscript version.
The version in the University of Hertfordshire Research Archive may differ from the final published version. **Users should always cite the published version.**

**Copyright and Reuse:**
This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

**Enquiries**
If you believe this document infringes copyright, please contact the Research & Scholarly Communications Team at rsc@herts.ac.uk

# Support Vector Regression to Estimate the Permeability Enhancement of Potential Transdermal Enhancers.

A. Shah[1], Y. Sun[2], R.G. Adams[2], N. Davey[2], S.C. Wilkinson[3], G.P. Moss[4]*.

[1]Department of Software Engineering and IT, Ecole de Technologie Superieure, Montreal, Canada

[2]School of Computer Science, University of Hertfordshire, Hatfield, UK

[3]School of Pharmacy, Keele University, Keele, UK

[4]Medical Toxicology Centre, Wolfson Unit, Medical School, University of Newcastle-upon-Tyne, UK

***Corresponding author:**

Dr Gary Moss

The School of Pharmacy

Keele University

Hornbeam Building HNB1.16

Keele

Staffordshire

ST5 5BG

UK

Tel: +44(0)1782 734 776

Fax: +44(0)1782 733 326

Gary Moss (g.p.j.moss@keele.ac.uk)

**Abstract**

**Objectives**: Searching for chemicals that will safely enhance transdermal drug delivery is a significant challenge. This study applies Support Vector Regession (SVR) for the first time to estimating the optimal formulation design of transdermal hydrocortisone formulations.

**Methods**: The aim of this study was to apply SVR methods with two different kernels in order to estimate the enhancement ratio of chemical enhancers of permeability.

**Key Findings**: A statistically significant regression SVR model was developed. It was found that SVR with a nonlinear kernel provided the best estimate of the enhancement ratio for a chemical enhancer.

**Conclusions**: SVR is a viable method to develop predictive models of biological processes, demonstrating improvements over other methods. In addition, the results of this study suggest that a global approach to modelling a biological process may not necessarily be the best method and that a "mixed methods" approach may be best in optimising predictive models.

**Key words:** Support Vector Regression; Support Vector Machine; Gaussian Processes; Transdermal Enhancer; Hydrocortisone.

**Conflict of interest**

The authors of this study have no conflicts of interest to disclose.

**Acknowledgements**

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Introduction**

Considerable success has been achieved in the delivery of drugs into and across the skin in the last thirty years. The advantages of this route of administration are well documented, and significant examples of successful therapies include transdermal patches for smoking cessation, pain management and hormone replacement therapy[1]. Nevertheless, the skin – and the *stratum corneum* in particular – remains a challenging barrier to the permeation of exogenous chemicals. This has resulted in a range of physical and chemical strategies to enhance drug delivery.

One of the most widely investigated methods of skin permeation enhancement involves the incorporation of chemicals into a topical formulation with the aim of reducing the skin barrier function – usually by altering the conformation of skin lipids – thus facilitating greater permeability. Such chemical penetration enhancers are often classified by their mechanism and site of action[2]. While ideally one would wish such chemical enhancers to increase permeation without exhibiting any negative side effects – including irritation and poor patient compliance, such profiles are rare and, in practice very few significant chemical enhancers have found common use in topical pharmaceutical products[3].

Mathematical models of skin permeation have been widely researched but uncommonly applied to relevant endpoints for the last twenty years. Statistically-derived relationships between chemical transport across the skin (usually characterised as either permeability, $k_p$, or steady-state flux, $J_{ss}$) and the physicochemical properties of a penetrant – usually presented in the form of an easily understood algorithm – have found utility in this field[4]. However, such models have limitations, including their lack of relevance to formulation issues as they are predominately derived from permeability studies where permeation was determined from simple solutions, inferring that such models do not consider the influence of formulation on absorption[4].

Several approaches have been used to model formulation effects. For example, hybrid quantitative structure-permeability relationships (QSPRs) were used to examine the effect of solvent mixtures on the skin permeation of model penetrants[5]. 12 compounds and 24 mixtures were used, and this approach was able to yield improved models for the permeation of complex chemical mixtures. Finite dose systems were also considered[6] by measuring the permeability of four chemicals from a range of 24 solvent blends in a finite dose *in vitro* model using a pig skin membrane. This resulted in four quantitative structure-activity relationships (QSARs) which described permeability in terms of both

physicochemical properties and solvent blends, and suggested that compounds formulated with a small difference in the boiling point and melting point of the vehicle resulted in higher skin permeation.

Discriminant analysis was used to classify the effect of skin penetration enhancers on the percutaneous absorption of hydrocortisone. Pugh et al. employed the "enhancement ratio" (of the enhancement in hydrocortisone skin permeation observed in the presence of a particular chemical enhancer, compared to a formulation without the enhancer) to define simple classifications of enhancement as either "good" (ER>10) or "poor" (ER<10)[7]. They found that the longest carbon chain length on a molecule, the molecular weight and the number of hydrogen bonding atoms on a molecule were significant for enhanced delivery. This approach was unable to provide a reliable prediction of ER for new enhancers. Thus, a range of Machine Learning methods were applied to Pugh's dataset[8]. The Support Vector Machine (SVM) and Gaussian Process (GP, with the synthetic minority over-sampling technique (SMOTE)) methods resulted in improved classification results[7,8] and offered the additional advantages of fewer false-positive results and the ability to make predictions of an enhancer's potential ability.

The Support Vector Regression (SVR) method has not previously been applied to a pharmaceutically relevant endpoint. The SVC method previously reported does offer significant improvements in model quality compared to discriminant analysis[7]. However, SVC is limited in that it is essentially a classification method and was able to provide class membership only, as defined by the degree of enhancement benefit, rather than estimates of performance improvement[8]. Further, the novel comparison of two Machine Learning methods in this study will test the current perception that a single, "global", model should be used to model a data set. A direct comparison between different methods (Gaussian Processes and SVR) will allow us to explore whether these methods provide distinct differences in model prediction and whether certain models should be used within a particular part of the "chemical space" in order to optimise predictive power and subsequent significance of the pharmaceutically relevant endpoint.

The aims of the current study are to therefore assess the viability of the SVR method in providing improved estimates of the enhancement ratio of chemicals and whether the best approach to modelling such systems is to use a single model or a range of models which optimise predictive power in certain parts of the chemical space of the data set studied [9, 10]. This study is the first time that SVR has been employed to estimate formulations effects. In doing so, results are benchmarked against

previous studies[8]. The effect of using different numbers of physicochemical descriptors on the quality of ER estimates is also considered.

**Methods**

**Description of the data and descriptive statistical analysis**

The dataset employed in this study (Table 1) consists of seventy-one compounds, each with five commonly used and readily calculable physicochemical descriptors: the count of hydrogen bonding groups on a molecule (HB), the carbon chain length of the molecule (CC), molecular weight (MW), lipophilicity (as log P, the logarithm of the octanol-water partition coefficient); aqueous solubility (as log S) and the enhancement ratio (ER), described above. Enhancers were previously grouped into "good" (1) and "poor" (0) classes, and where ER=10 is an arbitrary threshold as an enhancer with ER$\geq$10 is considered to exert a sufficiently large effect to potentially be clinically relevant[7].

[INSERT TABLE 1 HERE]

A quantile-quantile plot of MW by class ("good" or "poor"; Figure 1) shows that the linear reference line which joins the first and third quartiles of each distribution is distinct from the enhancer data – represented with the symbol '+' – which shows a curved pattern with a slope decreasing from left to right. This suggests that the two defined subsets may have different distributions. Comparison of the distributions of the values in the two subsets by a two-sample Kolmogorov-Smirnov test suggests that the two subsets are not from the same continuous distribution. The same tests were performed on all five descriptors employed in this study and a significant difference between the "good" and "poor" enhancers was found for all descriptors except HB.

[INSERT FIGURE 1 HERE]

**Principal Component Analysis**

Visualisation of the underlying distribution of the data was achieved by principal component analysis (PCA; varimax rotation). All the data were normalised so that all five descriptors had a zero mean and unit variance. PCA was applied and mapped the data to a low-dimensional space with a linear transformation whilst maintaining as much variance in the data as possible. Thus, a plot of the first two principal components using the logER (to allow the data to be suitably scaled; the use of logER

generally results in a more symmetrical plot[7]) values is shown in Figure 2. The first principal component (PC1 = 0.05HB – 0.47CC – 0.49MW – 0.53logP + 0.51logS) accounts for 66.97% of the total variance, and the second principal component (PC2 = 0.93HB – 0.10CC + 0.29MW – 0.20logP + 0.08logS) accounts for 22.64%. It is also clear that most of the high values of logER are associated with -log PC scores. For example, many compounds with logER > 1 have scores of PC1 or PC2 lower than zero. The correlation coefficient between actual values of logER and scores of the first PC was -0.58 (and -0.34 for the second PC). The p-values, 0.0000 and 0.0040 for PC1 and PC2, respectively, indicate that both correlations were statistically significant. It suggests a negative relationship between actual values of logER and scores of the first and second principal components.

[INSERT FIGURE 2 HERE]

**Performance measures**

The indicators of performance used in this study were, in common with our previous studies, the mean squared error (MSE) and the correlation coefficient (CORR)[11 – 15]. A confusion matrix (Table 2) is also applied to further analyse the accuracy of predictions following the application of a threshold of logER equal to log10 to ensure that true negatives and true positives are both as high as possible.

**Data analysis methods**

A range of methods were used to analyse the data. These methods are based on those used previously[8] and are summarised below.

**Fitted linear regression and simple linear regression**

Prior to the application of SVR modelling methods, results from two fitted linear regression methods reported previously[7, 8] are used to benchmark Machine Learning results:

$$\text{logER} = 0.318 - 0.0770\text{HB} + 0.668\text{CC} \quad (1)$$

$$\text{logER} = 0.326 - 0.0756\text{HB} + 0.0677\text{CC} - 0.000072\text{MW} \quad (2)$$

Equation (2) produces the smallest MSE on the complete dataset. Simple linear regression considers the output, y, as the weighted sum of the components of an input vector, x [16]:

$$y = \sum_{i=1}^{d} w_i x_i + w_0 \quad (3)$$

where d is the dimensionality of the input space

w = ($w_1$,…,$w_d$.$w_0$) is the weight vector, where weights are set so that the sum squared error function is minimised on a training set

$w_0$ is the bias.

**Support vector regression (SVR)**

Both ε-SVR and ν-SVR are applied to this study. Given a training dataset, {($x_1$, $y_1$),…… ($x_n$, $y_n$)}, the aim of ε-SVR is to fit a function, $f_x$,using the training data so that the difference between the target value, $y_i$, and the estimated value, $\hat{y}_i$, is no larger than ε for the i[th] training example. This is given by:

$$\min_{\omega,b,\xi,\xi*} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i + C \sum_{i=1}^{n} \xi_i^* \quad (4)$$

$$subject\ to \begin{cases} w^T \emptyset(x_i) + b - y_i \leq \epsilon + \xi_i, \\ y_i - w^T \emptyset(x_i) - b \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, i = 1, \dots, n. \end{cases}$$

where w is a weight vector

b is the bias and $\phi(x_i)$ maps $x_i$ to a higher-dimensional space

C is a constant (where C $\geq$ 0 in all cases) which is referred to as the regularisation parameter and determines the trade-off between the soft margin, which is described by the constraints, and the amount up to which differences larger than $\epsilon$ are tolerated.

$\xi_i$ and $\xi_i^*$ are slack variables which are used to relax the constraints slightly to allow for bad estimations.

Lagrange multiples ($\alpha$) are applied to produce predictions for new chemicals. The solution for the estimation at each new point, $x_*$, is determined by:

$$\widehat{y_*} = \sum_{i=1}^{n} (-\alpha_i + \alpha_i^*) \emptyset(x_i)^T \emptyset(x_*) + b \quad (5)$$

Application of (5) avoids the need to calculate $\phi(x_i)$ directly. Thus, kernel functions are used as:

$$k(x_i, x_j) = \emptyset(x_i)^T \emptyset(x_j) \quad (6)$$

The solution for the estimation at the new point, $x_*$, is equivalent to:

$$\widehat{y_*} = \sum_{i=1}^{n} (-\alpha_i + \alpha_i^*) \mathrm{k}(x_i, x_*) + b \quad (7)$$

Equation (7) is that which is normally used in practice, where a linear and radial basis function (RBF) kernels are applied to the data. Further information on the use of this approach, and the calculation of *b*, can be found elsewhere[17]. Scholkopf et al. proposed the use of a parameter, where $v \in [0; 1]$, to control the number of support vectors[18]:

$$\min_{\omega, b, \xi, \xi*} \frac{1}{2} w^T w + C(v\epsilon + \frac{1}{n} \sum_{i=1}^{n} (\xi_i + \xi_i^*)) \quad (8)$$

$$subject\ to \begin{cases} w^T \emptyset(x_i) + b - y_i \leq \epsilon + \xi_i, \\ y_i - w^T \emptyset(x_i) - b \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, i = 1, \dots, n, \epsilon \geq 0 \end{cases}$$

$\nu$-SVR represents an upper bound on the fraction of training samples which are errors (badly predicted) and a lower bound on the fraction of samples which are support vectors[18]. Thus $\in$ or $\nu$ are

different versions of the penalty parameter and the same optimisation problem is solved in either case, with the optimal solution set of $\nu$-SVR being part of the optimal solution set of $\in$-SVR [17].

**Gaussian process regression (GPR)**

Gaussian process modelling is a non-parametric method and has been used extensively in skin permeability and is described in detail elsewhere[11-15]. To make a prediction in GP, $y_*$, at a new input, $x_*$, the conditional distribution ($p(y_*|y_1,\ldots\ldots y_{Ntrn})$, where $N_{trn}$ denotes the number of training examples) on the observed vector should be calculated. The mean at $x_*$ is given by:

$$E[y_*] = \mathrm{k}_*^T (K + \sigma_n^2 I)^{-1} \mathrm{y} \quad (9)$$

where $k_*$ denotes the vector of covariances between the test point and the $N_{trn}$ training data

K denotes the covariance matrix of the training data

$\sigma_n^2$ is the variance of an independent identically distributed Gaussian noise, which infers that observations are noisy

*y* is the vector of training targets

The predictive variance at $x_*$ is given by:

$$var[y_*] = \mathrm{k}(x_*, x_*) - \mathrm{k}_*^T (K + \sigma_n^2 I)^{-1} k_* \quad (10)$$

The mean and variance represent the prediction and its associated error. In this study the results of GPR reported previously[8] are used compared with the new results.

**Experiments and Results**

*Overview*

For each analysis of the dataset (Table 1, using either 3 or 5 descriptors) the leave-one-out technique is applied and performance metrics are computed in terms of all predictions. Linear and RBF kernels are applied for regression analysis on the datasets using SVR. The SVR experiments were completed using ***libsvm***, which is available from http://www.csie.ntu.edu.tw/~cjlin/libsvm.

*Experiment 1 – analysis of a dataset with three molecular features*

Two types of SVR with linear kernels were tested on the dataset with three features (HB, CC, MW) (Table 3). This table (and, similarly in Table 4) also includes, for comparison, the results from the FLR and SLR reported previously[8]. FLR gives the best statistical performance on both measures used (MSE and CORR). It is notable that the other methods used exhibit a similar performance and are nearly as accurate in all cases in providing predictions on data that they have not previously seen.

[INSERT TABLE 3 HERE]

The first principal component is plotted against the experimentally measured logER value (Figure 3; it should be noted that this is a different PCA analysis from that presented in Figure 2, since it is based only on three descriptors). The dots are estimations from the FLR analysis and circles are experimentally derived logER values. There are two estimated logER values just above 1.0, while the corresponding experimental values are lower than 1.0. There are approximately 11 chemicals where logER>1.0 while their estimated logER values from the FLR analysis are all lower than 1.0. Since 59 of 71 chemical compounds have low logER values (logER$\leq$log (10)), the fitted linear hyperplane needed to match the majority of data points well. This suggests that the FLR method is unsuitable for finding good potential transdermal enhancers. Interestingly, when considering the use of Pugh's arbitrary threshold[7] to classify enhancer effects as "good" or "poor", a confusion matrix produced by the FLR is given by TN=57, TP=0, FP=2 and FN=12, while a confusion matrix produced by the $\varepsilon$-SVR is given by TN=56, TP=4, FP=3 and FN=8. It shows that this SVR method may not only provide a better true estimation but that it also reduces false negatives. The confusion matrix produced by the simple linear regression model is similar to that produced by the FLR (where TN=56, TP=0, FP=3 and FN=12).

[INSERT FIGURE 3 HERE]

*Experiment 2 – analysis of a dataset with five molecular features*

FLR was not applied to this study as it was produced using three features. Previously reported SLR analyses are used for comparison to the SVR methods[8]. The results (Table 4) show that SLR and ν-SVR have the same performance and that both are slightly better than ε-SVR on both performance metrics. When compared with Table 3, it can be seen that the ν-SVR method works better with five features than with three, but that the ε-SVR method, irrespective of the number of features used, has the same performance on MSE and only a slight improvement on CORR when five descriptors are used. Greater clarity on the significance of these results is not apparent and may be due to variance associated with both the source data and, potentially, variance associated with the use of predictive methods for estimating parameters such as log P and log S.

[INSERT TABLE 4 HERE]

*Experiment 3 – application of the RBF kernel (3 molecular features)*

The RBF kernel was applied to both types of SVR and tested on the dataset with three features (Table 3). For comparison, Table 4 also shows the results of the GPR with the *Matern3* kernel reported previously[8]. The results show that the GPR method gives the best results. Comparing the results in Table 3 and Table 4 shows that the ε-SVR method has the same performance on both linear and RBF kernels with three molecular features. Further, the ν-SVR method gives a slight improvement on the CORR measurement using the RBF kernel while the MSE measurement remains the same for both kernels with three features.

*Experiment 4 – application of the RBF kernel (five molecular descriptors)*

The RBF kernel was then applied to the dataset with five molecular features, and the results are shown in Table 4. The best results were found with the ν-SVR method, which also provided the overall best result for any experiment in this study (or in previous studies). Comparing the results in Tables 3 and 4 indicates that both types of SVR perform better when using five molecular features rather than three, although the differences between both SVR methods and the GPR method is quite small. However, use of the ν-SVR method results in a 23.15% decrease in MSE and improves CORR by 6.8% when applied to a five-feature model. Comparison of Tables 3 and 4 indicate that both types of SVR with the RBF kernel outperform those with the linear kernel; e.g., the ε-SVR decreases the MSE by 16.7% and increases CORR by 4% when using a RBF kernel.

*Experiment 5 – removal of the hydrogen bonding molecular feature from models*

The final experiment removed the hydrogen bonding term (HB) from consideration for modelling in both 3- and 5-feature datasets. This was due to the quantile-quantile plot indicating that "good" and "poor" enhancers could not be discriminated from each other on the basis of hydrogen bonding. It is noteworthy therefore that PC1, which correlates well with logER, is only weakly influenced by hydrogen bonding but quite strongly influenced by log P and log S. PC2, however, is very strongly influenced by HB but it does not correlate well with logER. Table 4 (the last two rows) show that the result of removing hydrogen bonding from the models is to lessen their statistical quality slightly. This suggests that a variable which, by itself, is without merit but may, when combined with others, enhance the overall performance of the model[12]. In the case of the QSAR models parameters such as log P and MW, by themselves, produce little in the way of significant models but model quality improves when such features are combined. It should also be noted that the SVR method does not currently provide mechanistic information on the permeability enhancement process. This may be methodological, as no clear relationships were observed between chemical structure and enhancement effects, or it may be that any such trends are masked by the biological variation present in the data. It also does not currently allow us to discern now the enhancers function by altering $K_{sc}$ or $D_{sc}$, for example. However, Machine Learning methods have previously been modified to provide significant mechanistic information (for example by using the Feature Selection methods described previously[12]) and it is possible that such methods may soon also be applied to SVR methods[15].

*Experiment 6 – Statistical comparison of SVR and GP methods*

Errors in prediction for the GP and SVM methods were compared with a two-tailed binomial test. GP predictions were not significantly better than SVM predictions (P = 0.34). Further, the members of the dataset (n = 71) were ranked in terms of the absolute difference of prediction accuracy, using both Spearman and Kendall methods for rank correlation. The Spearman test yielded a correlation coefficient of 0.12 (P = 0.33) and the Kendall test gave a correlation coefficient value of 0.09 (P = 0.29). Thus, the null hypothesis (that the correlation is not zero) is not rejected, suggesting that GP and SVM predictions may have a certain degree of correlation. The Wilcoxon signed rank test returned a P-value of 0.3425, indicating that the test fails to reject the null hypothesis, suggesting that the GP and SVM predictions may be from a distribution with the same median.

**Discussion**

Identification and treatment of outliers can be a contentious issue in the field of quantitative predictive model development. In particular the definition of an outlier – whether being a chemical atypical to the data set used, or an output following modelling that is itself atypical – is itself often unclear. The former may be dealt with using criteria described previosuly[24] and can be dealt with by the construction of a data set which avoids skew, bias and outliers. The latter, however, is more problematic and is usually associated with, at best, semi-quantitative interpretation of the output of a model based on related criteria, such as the experimental conditions and context under which the data was obtained. In this, and similar, studies outliers are commonly identified by principal component analysis, followed by a comparison between the targets and predictions[8, 11, 13].

Equations 5 and 9 show that both estimates of permeation enhancement can be considered as a linear combination of kernels applied to both the best point and each training data point. In SVR, the corresponding coefficients are $\alpha_i + \alpha_i^*$. Since $\alpha_i + \alpha_i^*$ are zeros inside the margin not all training examples are needed to contribute to the estimate; because of the zeros only about 50 of the 70 training examples are used to estimate the logER values of the last test point in the dataset with five features. By comparison, in the GP method the covariance matrix (kernel) over the whole training set is used as weights, making more use of all the available data.

Figure 4 shows the absolute differences between the actual logER values and the estimates from two models, the $\nu$-SVR with RBF kernel and the GPR model. The two models produce a similar error trend over the whole range of scores for the first principal component. $\nu$-SVR provides a better estimate on 40 out of 71 potential chemical enhancers, while GPR has better performance on certain specific parts of the first two principal component spaces. For example, GPR has a better performance around PC1≥3 and GPR gives a more reliable estimate on 8 out of 12 chemicals that have a PC1≥2.3. In addition, GPR has a better performance at PC1 $\in$ [-1.17, -0.94] on all six chemicals. Figure 4 also shows that the two biggest errors produced by these two models correspond to the same scores of the first principal component, which belong to S,S-Dimethyl-N-(4-bromobenzoyl)iminosulfurane and S,S-Dimethyl-N-(5-nitro-2-pyridyl)iminosulfurane (Table 1). Figure 2 indicates that S,S-Dimethyl-N-(4-bromobenzoyl)iminosulfurane corresponds to the point which has a large value of logER (approximately 1.36) and that the score of its PC is approximately 0.73, and S,S-Dimethyl-N-(5-nitro-2-pyridyl)iminosulfurane corresponds to the point having a logER value at about 1.0 and the score of the first PC at about 1.23. Both chemical compounds are far away from other compounds in having

similar large values of logER. Most of the compounds having a score for the first principal component greater than zero have smaller values of logER (usually smaller than 0.75). Thus, it would appear that S,S-Dimethyl-N-(4-bromobenzoy)iminosulfurane and S,S-Dimethyl-N-(5-nitro-2-pyridyl)iminosulfurane are outliers in the range of PC1 $\in$ [0, 4] as defined by the five molecular features employed in this analysis. This suggests that it may be difficult for current SVR and GPR methods to produce useful estimates for chemicals considered as outliers.

[INSERT FIGURE 4 HERE]

Clearly, the methods discussed herein sit with those described previously[8], offering an incremental improvement to previously published methods. However, the novelty of the current study is that it offers substantially better estimates of skin permeability enhancement than those reported by Pugh et al. as well as a predictive model for new chemicals[7]. Thus, it again indicates the benefits – and, still, the enormous potential – in applying Machine Learning methods to biological systems[15].

Further, as the findings of this study show the validation of a new method is significant but its longer-term potential may relate to the development of non-universal models – models that work best with small sub-sets of data and which may suit different methods of analysis in order to optimise specific data sets[4]. Certain chemicals perform better when one particular method is applied (Figure 4), but that the opposite may apply to other chemicals. Such an outcome may align itself to the considerations given by Flynn in proposing a series of models based on differing physicochemical properties, rather than a single 'global' model, and therefore estimating biological effects for chemicals with the most appropriate modelling method and not with a generic model[19].

No discernable trend in structure and predictive ability was found in this study, particularly when statistical ranking tests were applied (Table 1). While this might be surprising, given the structural diversity present in the enhancer dataset, but may be due to the inherent biological variation associated with results of this, and similar, experimental studies. This implies that small differences in estimated ER, in terms of statistical measures of performance, might not yield a significant difference between two enhancers. It should be noted that the underlying mechanistic understanding of the skin penetration enhancement was discussed previously[7]. In this study the authors determined that the key physicochemical descriptors for optimising the enhancement ratio of potential skin penetration

enhancers of hydrocortisone were carbon chain length, molecular weight and the total number of hydrogen bonding groups on a molecule.

Moreover, Figure 5 shows contour plots of the absolute differences between predictions and targets for the GPR and $\nu$-SVR methods. The absolute differences between predictions and targets are treated as heights above the PCA plane. While Figure 5 shows that both planes are superficially similar it is interesting to focus in detail on the differences. For example, for the area where PC1$\leq$-1.5 and PC2$\leq$-0.5 it is apparent that the GPR method is producing better estimates than the SVR method. In looking at the first two principal components, shown earlier, they suggest that, if the value of PC1 is low, the values of the descriptors HB, CC, MW and logP should be high; if the value of PC2 is low it means that the descriptors CC and logP should be high and that HB, MW and logS should have low values. Thus, for the area where PC1$\leq$-1.5 and PC2$\leq$-0.5 we should have larger values of CC and logP and a relatively smaller value for logS. It also appears that, for MW and HB, there is a compromise between PC1 and PC2. This is significant in considering the application of a global model to this problem domain.

[INSERT FIGURE 5 HERE]

Thus, the SVR method offered advantages over other methods, either as a complementary approach to solving the chosen problem domain but also in offering a different method to those currently used. It was also observed that the SVR method is more robust when dealing with chemical compounds whose properties have similar values, while Gaussian Process Regression does not work as well in such situations due to the numerical manipulations involved with the inversion operation (Equation 9). This is significant as it has implications for inherently variable biological data, including multiple or repeated experiments for the same chemical, where variability will inevitably be observed. This applies also to the generation of the physicochemical descriptors used in the modelling process, such as log P (or log $K_{o/w}$) which may be subject to variation when measured for subsequent use in models. This is also significant due to the other main finding of this study – that different models, as shown in Figure 5 and Table 1, appear to provide better estimates in different parts of the same "chemical space" for a large dataset, which is significantly larger than a substantial number of those reviewed previously. This allows significant developments in estimate quality to be made, particularly in the context of hyperparameter studies which show significant improvements in estimate quality in small datasets[4, 22].

The use of Machine Learning methods in biological domains has been limited, mostly due to issues of technology transfer and model transparency – the latter is a particular issue as most Machine Learning methods work on a "black box" basis and, unlike more rudimentary statistical methods they do not yield a distinct descriptive algorithm, or equation, describing the biological process in the context of relevant physicochemical descriptors[15]. Thus, models often get used briefly and then fail to sustain their relevance in these fields. Several methods have been used to alleviate these shortcomings, including Feature Selection, which allows specific mechanistic information to be determined for each method. Indeed, it was shown previously that Feature Selection demonstrated that some physicochemical descriptors were interchangeable and that by changing these descriptors predictions of the same quality could be produced from a range of different descriptors[12]. One implication of this is that, if descriptors are used which are not subject to experimental variation (i.e. the use of MW, the count of hydrogen bonding groups) then models might yield less variance than those which use descriptors which are subject to such variation (i.e. log P, log S, melting point).

However, it is important to note that the algorithms produced following Flynn's work are predominately global models which are simple, to easily interpret and readily applied to relevant endpoints. In providing a single expression for skin permeability across a wide range (commonly, for the datasets discussed, this range is usually $-3 < \log P < 6$ and $150 < MW < 750$) such models are consistent with comments on transport across lipid lamellae[23]. They are also simplified to the point that any separate consideration of the aqueous pore pathway described by Flynn[19] is absent, and is often being considered a matter that can be addressed by consideration of molecular volume rather than a discrete aqueous pathway[21]. For a toxicological endpoint, holistic modelling of the full biological process is preferred and models should not use individual steps for different parts of the process, such as transport or reactivity[24, 25]. Such comments clearly sit at odds with the inferences drawn from the current study.

An increase in the number of descriptors, from three to five, improved model quality and utility. However, while this must be considered in the context of potential over-fitting the number of descriptors used in this study is neither excessive nor irrelevant, being based on those molecular features found previously to be significant[24, 26]. It might therefore be appropriate to consider a similar study using an expanded set of molecular features which have been optimised and whose relevance has been determined by the application of methods such as feature selection[12, 27].

**Conclusions**

The SVR method demonstrated comparable performance to previously reported Machine Learning methods (i.e. SVC) in the same field, but with the added significant advantage of being able to provide estimates of the effects of new chemicals; in this case, the enhancement ratio of new chemical enhancers. Performance of all Machine Learning methods was significantly better than discriminant analysis[7]. It is also apparent that different Machine Learning methods (SVR and GP) provided different outputs, suggesting that the use of a global model may not optimise model predictivity in all cases. This suggests that specific regression models, possibly following the use of a classification method, may be applied to different parts of the chemical space in order to improve predictivity. Given the potential of these methods to provide statistically significant estimates of enhancement it is clear that substantial potential exists to explore this field more deeply. In doing so it may be suggested that a combination of classification (i.e. SVC[8]) and regression methods (i.e. SVR), possibly combined with Feature Selection[12], will allow further exploration of the links between mathematical methods and chemical structure which lies at the heart of such studies, and also allow improved modelling within local regions of a larger data set.

**References**


1. Wiedersberg, S.; Guy, R.H. Transdermal drug delivery: 30+ years of war and still fighting! J. Cont. Rel. 2014, 190, 150-156.

2. Ghosh, T.K.; Pfister, W.; Yum, S.I. *Transdermal and topical delivery systems*. Interpharm Press: Buffalo Grove, IL, USA, 1997.

3. Katz, M.; Poulsen, B.J. Corticoid, vehicle and skin interactions in percutaneous absorption. J. Soc. Cos. Chem. 1972, 23, 565-590.

4. Moss, G.P.; Dearden, J.C.; Patel, H; Cronin, M.T.D. Quantitative structure-permeability relationships (QSPRs) for percutaneous absorption. Tox. In Vitro. 2002, 16, 299-317.

5. Riviere, J.E.; Brooks, J.D. Prediction of dermal absorption from complex chemical mixtures: incorporation of vehicle effects and interactions into a QSAR framework. SAR and QSAR Environ. Res. 2007, 18, 31-44.

6. Ghafourian, T.; Samaras, E.G.; Brooks, J.D.; Riviere, J.E. Validated models for predicting skin penetration from different vehicles. Eur. J. Pharm. Sci. 2010, 41, 612-616.

7. Pugh, W.J.; Wong, R.; Falson, F.; Michniak, B.B; Moss, G.P. Discriminant analysis as a tool to identify compounds with potential as transdermal enhancers. J. Pharm. Pharmacol. 2005, 57, 1389-1396.

8. Moss, G.P.; Shah, A.J.; Adams, R.G.; Davey, N.; Wilkinson, S.C.; Pugh, W.J.; Sun, Y. The application of discriminant analysis and Machine Learning methods as tools to identify and classify compounds with potential as transdermal enhancers. Eur. J. Pharm. Sci. 2012, 45, 116–127.

9. Demiriz, A.; Bennett, K.P.; Breneman, C.M.; Embrechts, M.J. *Support vector machine regression in chemometrics.* In: Computing Science and Statistics: Proceedings of Interface, Volume 33. 2001.

10. Jung, E.; Choi, S.H.; Lee, N.K.; Kang, S.K.; Choi, Y.J.; Shin, J.M.; Choi, K.; Hung, D.H. Machine learning study for the prediction of transdermal peptide. J. Comp. Aid. Mol. Des. 2011, 25, 339-347.

11. Moss, G.P.; Sun, Y.; Prapopoulou, M.; Davey, N.; Adams, R.; Pugh, W.J.; Brown, M.B. The application of Gaussian processes in the prediction of percutaneous absorption. J Pharm Pharmacol. 2009, 61, 1147-1153.

12. Lam, L.T.; Sun, Y.; Davey, N.; Adams, R.G.; Prapopoulou, M.; Brown, M.B.; Moss, G.P. The application of feature selection to the development of Gaussian process models for percutaneous absorption. J. Pharm. Pharmacol. 2010, 62, 738–749.

13. Sun, Y.; Moss, G.P.; Davey, N.; Adams, R.; Brown, M.B. The application of stochastic Machine Learning methods in the prediction of skin penetration. App. Soft Comp. 2011, 11, 2367-2375.

14. Moss, G.P.; Sun, Y.; Wilkinson, S.C.; Davey, N.; Adams, R.; Martin, G.P.; Prapopoulou, M.; Brown, M.B. The application and limitations of mathematical models across mammalian skin and poldimethylsiloxane membranes. J. Pharm. Pharmacol. 2011, 63, 1411-1427.

15. Ashrafi, P.; Moss, G.P.; Wilkinson, S.C.; Davey, N.; Sun, Y. The Application of Machine Learning to the Modelling of Percutaneous Absorption: An Overview and Guide. SAR & QSAR Environ. Res. 2015, 26, 181-204.

16. Bishop, C.M. *Neural Networks For Pattern Recognition*. Oxford University Press: Oxford, 1995.

17. Chang, C.C.; Lin, C.J. LibSVM: a library for support vector machines. ACM Trans. on Intell. Sys. Tech. 2011. 2, 1-27.

18. Scholkopf, B.; Smola, A.; Williamson, R.C.; Bartlett, P.L. New support vector algorithms. Neural Comp. 2000, 12, 1207-1245.

19. Flynn, G.L. *Physicochemical determinants of skin absorption*. In: *Principles of Route-to-Route Extrapolation for Risk Assessment*; Gerrity, T.R.; Henry, C.J., Eds.; Elsevier: New York, 1992; pp 93-127.

20. Magnusson, B.M.; Anissimov, Y.G.; Cross, S.E.; Roberts, M.S. Molecular size as the main determinant of solute maximum flux across the skin. J. Invest. Dermatol. 2004, 122, 993 – 999.

21. Mitragotri, S.; Anissimov, Y.G.; Bunge, A.L.; Frasch, H.F.; Guy, R.H.; Hadgraft, J.; Kasting, G.B.; Lane, M.E.; Roberts, M.S. Mathematical models of skin permeability: An overview. Int. J. Pharm. 2011, 418, 115-129.

22. Ashrafi, P.; Sun, Y.; Davey, N.; Adams, R.; Brown, M.B.; Prapopoulou, M.; Moss, G.P. The importance of hyperparameters selection within small datasets. International Joint Conference on Neural Networks, Killarney, Ireland, July 2015, pp139 [#15532]. [Available online at: http://www.ijcnn.org/assets/docs/ijcnn2015-program-v3.pdf; last accessed 16th June 2015], 2015.

23. Lieb, W.R.; Stein, W.D. Implications of two different types of diffusion for biological membranes. Nature. 1971, 243, 219-222.

24. Cronin, M.T.D.; Schultz, W.T. Pitfalls in QSAR. J. Theoret. Chem. (Theochem). 2003, 622, 39-51.

25. OECD Principles for the Validation of (Q)SARs. [Available at: http://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm]. (last accessed 24th June 2015).

26. Baert, B.; Deconinck, E.; van Gele, M.; Slodicka, M.;, Stoppie, P.; Bode, S.; Slegers, G.; van der Heyden, Y.; Lambert, J.; Beetens, J.; de Spiegeleer, B. Transdermal penetration behaviour of drugs: CART-clustering, QSPR and selection of model compounds. Bioorg. Med. Chem. 2007, 15, 6943–6955.

27. Moss, G.P.; Sun, Y.; Davey, N.; Adams, R.G.; Wilkinson, S.C.; Gullick, D.R. *The application of Gaussian Processes in the prediction of permeability across a polydimethylsiloxane membrane*. In: *Advances in the Dermatological Sciences*; Brain, K.R.; Chilcott, R., Eds.; Royal Society of Chemistry: Cambridge, 2013, pp 376-383.

Table 1. Dataset used in this study, with outputs including rankings for GP and SVR methods and principal components. **Note**: HB represents the number of hydrogen bonding groups on a molecule; CC is the length of the longest continuous carbon chain present on a molecule; MW is the molecular weight of a chemical; log P is the mean of the octanol-water partition coefficient for a chemical; log S is the mean of the aqueous solubility for a chemical; ER Q is the enhancement ratio for a chemical, where permeation of hydrocortisone is compared for each chemical penetration enhancer to the same formulation without the chemical penetration enhancer; Formula is the chemical formula of each enhancer; SMILES a notation representing the Simplified Molecular In-Line Entry System for each chemical; GP and SVR ranking and error values indicate the ranking, in terms of model accuracy (i.e. predicted vs. measured values) for the Gaussian Process and Support Vector Regression methods with the best-ranked in each list representing the least difference between target and prediction; PC1 and PC2 are the outputs from principal component analysis.

| Compound Names | H bonds (HB) | Carbon chain (CC) | Molecular weight (MW) | Mean logP | Mean logS | ER Q | Formula | SMILES | GP rank | GP error | SVR rank | SVR error | PC1 | PC2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Urea | 7 | 0 | 60.06 | -1.692 | 0.565 | 1.5 | $C_1H_4N_2O_1$ | O=C(N)N | 30 | 0.1285 | 32 | 0.1466 | 3.4842 | 1.9057 |
| 2-Pyrrolidinone | 3 | 0 | 85.11 | -0.658 | 0.595 | 1.2 | $C_4H_7N_1O_1$ | O=C1CCCN1 | 13 | 0.0603 | 25 | 0.1184 | 3.2793 | -0.3222 |
| 1-Methylpyrrolidine | 1 | 1 | 85.15 | 0.72 | 0.385 | 1.4 | $C_5H_{11}N_1$ | C1CN(C)CC1 | 5 | 0.0114 | 47 | 0.2166 | 2.9275 | -1.5707 |
| 1-Methyl-2-pyrrolidinone | 2 | 1 | 99.13 | -0.328 | 0.685 | 1 | $C_5H_9N_1O_1$ | CN1C(CCC1)=O | 26 | 0.1087 | 29 | 0.1384 | 3.1022 | -0.8751 |
| 5-Methyl-2-pyrrolidinone | 3 | 1 | 99.13 | -0.164 | 0.34 | 1.3 | $C_5H_9N_1O_1$ | C1C(=O)NC(C)C1 | 11 | 0.0554 | 13 | 0.0347 | 2.9506 | -0.3428 |
| 1-Methylsuccinimide | 3 | 1 | 113.12 | -0.688 | 0.465 | 1.4 | $C_5H_7N_1O_2$ | C1C(=O)N(C)C(=O)C1 | 2 | 0.0003 | 15 | 0.0364 | 3.01 | -0.2522 |
| 1-Ethyl-2-pyrrolidinone | 2 | 2 | 113.16 | 0.228 | 0.445 | 1.1 | $C_6H_{11}N_1O_1$ | C1C(=O)N(CC)CC1 | 21 | 0.0909 | 5 | 0.0195 | 2.7647 | -0.8998 |
| 2-Pyrrolidinone-5-carboxylic acid | 6 | 0 | 129.12 | -1.102 | -0.065 | 1.1 | $C_5H_7N_1O_3$ | C1C(=O)NC(C(=O)O)C1 | 31 | 0.1293 | 21 | 0.0988 | 2.8639 | 1.5057 |
| 4-Acetylmorpholine | 3 | 1 | 129.16 | -0.598 | 0.645 | 1.3 | $C_6H_{11}N_1O_2$ | N1(C(C)=O)CCOCC1 | 19 | 0.0752 | 4 | 0.0137 | 2.9485 | -0.1986 |
| N-acetylcaprolactam | 3 | 1 | 155.2 | 0.58 | -0.285 | 4.6 | $C_8H_{13}N_1O_2$ | C1CN(C(C)=O)C(=O)CCC1 | 66 | 0.5904 | 69 | 0.6733 | 2.3389 | -0.2386 |
| Ethyl (R)-( )-2-pyrrolidinone-5-carboxylate | 5 | 2 | 157.17 | -0.226 | -0.13 | 1.1 | $C_7H_{11}N_1O_3$ | C1C(=O)NC(C(=O)OCC)C1 | 28 | 0.1158 | 42 | 0.1973 | 2.3724 | 0.9323 |
| (R,R)-( )-2,5-bis(methoxymethyl)pyrrolidinone | 4 | 1 | 159.23 | -0.214 | 0.013 | 2 | $C_8H_{17}N_1O_2$ | C1C(COC)NC(COC)C1 | 25 | 0.1066 | 56 | 0.2818 | 2.4245 | 0.3718 |
| 1-Cyclohexyl-2-pyrrolidinone | 2 | 2 | 167.25 | 1.756 | -1.02 | 1.2 | $C_{10}H_{17}N_1O_1$ | C2C(=O)N(C1CCCCC1)CC2 | 17 | 0.0703 | 10 | 0.028 | 1.7984 | -0.8949 |
| 1-Hexyl-2-pyrrolidinone | 2 | 6 | 169.27 | 2.276 | -1.365 | 1.2 | $C_{10}H_{19}N_1O_1$ | C1C(=O)N(CCCCCC)CC1 | 48 | 0.2363 | 57 | 0.29 | 1.2394 | -1.0195 |
| S,S-dimethyl-N-(benzoyl)iminosulfurane | 2 | 1 | 181.26 | 2.263 | -2.54 | 0.74 | $C_9H_{11}N_1O_1S_1$ | S(=NC(=O)c1ccccc1)(C)C | 57 | 0.3393 | 52 | 0.2638 | 1.3326 | -0.9286 |
| S,S-dimethyl-N-(4-nitrophenyl)iminosulfurane | 3 | 1 | 198.25 | 2.52 | -3.32 | 1.47 | $C_8H_{10}N_2O_2S_1$ | S(C)(C)=Nc1ccc(N(=O)=O)cc1 | 23 | 0.0912 | 35 | 0.1589 | 0.9614 | -0.3647 |
| S,S-dimethyl-N-(5-nitro-2-pyridyl)iminosulfurane | 4 | 1 | 199.23 | 1.54 | -2.86 | 9.03 | $C_7H_9N_3O_2S_1$ | S(C)(C)=Nc1ccc(N(=O)=O)cn1 | 70 | 1.083 | 70 | 1.0391 | 1.2334 | 0.2893 |
| 2-Nonyl-1,3-dioxolane | 2 | 9 | 200.32 | 4.23 | -3.64 | 8 | $C_{12}H_{24}O_2$ | C(CCCCCCC1OCCO1)CC | 34 | 0.1507 | 23 | 0.1081 | -0.1552 | -1.2143 |
| 1-Methyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam | 4 | 1 | 210.28 | 0.242 | -0.335 | 0.8 | $C_{11}H_{18}N_2O_2$ | N2(C1C(=O)N(C)CCCC1)C(=O)CCC2 | 59 | 0.356 | 24 | 0.1117 | 2.059 | 0.5254 |
| 1,3-Diphenylurea | 5 | 2 | 212.25 | 2.842 | -3.37 | 2 | $C_{13}H_{12}N_2O_1$ | O=C(Nc2ccccc2)Nc1ccccc1 | 4 | 0.009 | 11 | 0.0313 | 0.6518 | 0.753 |
| S,S-dimethyl-N-(2-methyl-4-nitrophenyl)iminosulfurane | 3 | 1 | 212.27 | 2.88 | -3.49 | 1.17 | $C_9H_{12}N_2O_2S_1$ | S(C)(C)=Nc1ccc(N(=O)=O)cc1C | 15 | 0.0611 | 8 | 0.0273 | 0.7701 | -0.3522 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S,S-dimethyl-N-(4-chlorobenzenesulfonyl)iminosulfurane | 2 | 1 | 215.7 | 2.75 | -3.05 | 1.44 | $C_9H_{10}N_1O_1S_1Cl_1$ | S(=NC(=O)c1ccc(Cl)cc1)(C)C | 51 | 0.2521 | 40 | 0.1769 | 0.9196 | -0.8719 |
| S,S-dimethyl-N-benzenesulfonyliminosulfurane | 3 | 1 | 217.31 | 2.007 | -2.18 | 0.48 | $C_8H_{11}N_1O_2S_2$ | S(C)(C)=NS(=O)(=O)c1ccccc1 | 63 | 0.4394 | 60 | 0.3329 | 1.2422 | -0.2171 |
| 2-Decylcyclopentanone | 1 | 10 | 224.39 | 5.77 | -5.23 | 6.7 | $C_{15}H_{28}O_1$ | C1(CCCCCCCCCC)CCCC1=O | 8 | 0.0281 | 30 | 0.1424 | -1.0466 | -1.8928 |
| S,S-dimethyl-N-(4-nitrobenzoyl)iminosulfurane | 4 | 1 | 226.26 | 2.175 | -3.61 | 0.77 | $C_9H_{10}N_2O_3S_1$ | S(C)(C)=NC(=O)c1ccc(N(=O)=O)cc1 | 62 | 0.4343 | 41 | 0.1948 | 0.7719 | 0.3007 |
| 1,3-Diphenylthiourea | 5 | 2 | 228.32 | 2.752 | -4.33 | 3.7 | $C_{13}H_{12}N_2S_1$ | S=C(Nc1ccccc1)Nc2ccccc2 | 65 | 0.4887 | 59 | 0.3166 | 0.3437 | 0.774 |
| S,S-dimethyl-N-(4-cyano-1-naphthyl)iminosulfurane | 2 | 1 | 228.32 | 3.49 | -5.69 | 1.16 | $C_{13}H_{12}N_2S_1$ | S(C)(C)=Nc1ccc(C#N)c2c1cccc2 | 53 | 0.2584 | 37 | 0.1697 | 0.0476 | -0.9925 |
| 1-Dodecylurea | 6 | 12 | 228.38 | 4.466 | -3.3 | 2.8 | $C_{13}H_{28}N_2O_1$ | O=C(NCCCCCCCCCCCC)N | 18 | 0.0712 | 39 | 0.1701 | -0.6733 | 1.0501 |
| S,S-dimethyl-N-(o-tolylsulfonyl)iminosulfurane | 3 | 1 | 231.34 | 2.492 | -2.46 | 0.93 | $C_9H_{13}N_1O_2S_2$ | S(C)(C)=NS(=O)(=O)c1ccc(C)cc1 | 22 | 0.0912 | 2 | 0.0119 | 0.9986 | -0.2184 |
| 2-Decylcyclohexanone | 1 | 10 | 238.41 | 6.058 | -5.475 | 7.9 | $C_{16}H_{30}O_1$ | C1C(CCCCCCCCCC)C(=O)CCC1 | 24 | 0.0916 | 16 | 0.0368 | -1.242 | -1.878 |
| N-dodecyl-pyrrolidine | 1 | 12 | 239.44 | 6.492 | -5.45 | 5.2 | $C_{16}H_{33}N_1$ | C1CN(CCCCCCCCCCCC)CC1 | 47 | 0.2356 | 61 | 0.3452 | -1.5075 | -1.9453 |
| 1-Dodecyl-3-methylurea | 5 | 12 | 242.4 | 4.908 | -3.295 | 1.8 | $C_{14}H_{30}N_2O_1$ | O=C(NCCCCCCCCCCCC)NC | 60 | 0.3784 | 55 | 0.2796 | -0.8045 | 0.5048 |
| S,S-dimethyl-N-(4-nitro-1-naphthyl)iminosulfurane | 3 | 1 | 248.31 | 3.575 | -4.92 | 1.28 | $C_{12}H_{12}N_2O_2S_1$ | S(C)(C)=Nc1ccc(N(=O)=O)c2c1cccc2 | 36 | 0.1809 | 33 | 0.1483 | 0.0785 | -0.3431 |
| S,S-dimethyl-N-(4-chlorobenzenesulfonyl)iminosulfurane | 3 | 1 | 251.76 | 2.6 | -2.9 | 0.4 | $C_8H_{10}N_1O_2S_2Cl_1$ | S(C)(C)=NS(=O)(=O)c1ccc(Cl)cc1 | 67 | 0.6389 | 64 | 0.4408 | 0.7558 | -0.1768 |
| 1-Dodecanoylpyrrolidinone | 2 | 11 | 253.43 | 5.19 | -4.325 | 15.6 | $C_{16}H_{31}N_1O_1$ | N1(C(CCCCCCCCCCC)=O)CCCC1 | 3 | 0.0042 | 12 | 0.0345 | -0.9879 | -1.1776 |
| N-dodecylpyrrolidinone | 2 | 12 | 253.43 | 5.494 | -4.41 | 23 | $C_{16}H_{31}N_1O_1$ | C1C(=O)N(CCCCCCCCCCCC)CC1 | 44 | 0.2276 | 51 | 0.2367 | -1.1595 | -1.2235 |
| S,S-dimethyl-N-(2,4,6-trichlorophenyl)iminosulfurane | 1 | 1 | 256.58 | 4.292 | -4.69 | 2.21 | $C_8H_8N_1S_1Cl_3$ | S(C)(C)=Nc1c(Cl)cc(Cl)cc1Cl | 40 | 0.2087 | 19 | 0.0573 | 0.0127 | -1.4774 |
| S,S-dimethyl-N-(4-phenylazophenyl)iminosulfurane | 3 | 2 | 257.36 | 3.775 | -4.06 | 2.81 | $C_{14}H_{15}N_3S_1$ | S(C)(C)=Nc2ccc(N=Nc1ccccc1)cc2 | 50 | 0.2461 | 62 | 0.3518 | 0.1124 | -0.313 |
| 1-Dodecyl-3-methylthiourea | 5 | 12 | 258.47 | 5.18 | -4.31 | 5.3 | $C_{14}H_{30}N_2S_1$ | S=C(NCCCCCCCCCCCC)NC | 41 | 0.2138 | 31 | 0.1445 | -1.1988 | 0.4965 |
| S,S-dimethyl-N-(4-bromobenzoyl)iminosulfurane | 2 | 1 | 260.15 | 2.738 | -2.83 | 23.12 | $C_9H_{10}N_1O_1S_1Br_1$ | S(C)(C)=NC(=O)c1ccc(Br)cc1 | 71 | 1.4111 | 71 | 1.4953 | 0.7326 | -0.7152 |
| S,S-dimethyl-N-(4-nitrobenzenesulfonyl)iminosulfurane | 5 | 1 | 262.31 | 1.968 | -3.33 | 0.68 | $C_8H_{10}N_2O_4S_2$ | S(C)(C)=NS(=O)(=O)c1ccc(N(=O)=O)cc1 | 27 | 0.111 | 28 | 0.1332 | 0.6519 | 1.0053 |
| 1-Dodecanoyl-2-pyrrolidinone | 3 | 11 | 267.41 | 4.946 | -3.895 | 10.1 | $C_{16}H_{29}N_1O_2$ | N1(C(CCCCCCCCCCC)=O)CCCC1=O | 16 | 0.0626 | 67 | 0.5169 | -0.9423 | -0.5372 |
| 1-Dodecanoylpiperidine | 2 | 11 | 267.46 | 5.65 | -4.56 | 14.7 | $C_{17}H_{33}N_1O_1$ | C1CCCCN1C(CCCCCCCCCCC)=O | 20 | 0.0767 | 3 | 0.0122 | -1.2154 | -1.1753 |
| N-dodecyl-2-piperidinone | 2 | 12 | 267.46 | 5.82 | -4.62 | 22.2 | $C_{17}H_{33}N_1O_1$ | C(CCCCCCCCCC)CN1CCCCC1=O | 39 | 0.1945 | 44 | 0.205 | -1.3539 | -1.2101 |
| Methyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam-1-acetate | 6 | 1 | 268.31 | 0.156 | -0.89 | 0.6 | $C_{13}H_{20}N_2O_4$ | N2(C1C(=O)N(CC(OC)=O)CCCC1)CCCC2=O | 54 | 0.2684 | 27 | 0.1324 | 1.554 | 1.8184 |
| N-(1-oxododecyl)morpholine | 3 | 11 | 269.43 | 4.488 | -3.62 | 15.1 | $C_{16}H_{31}N_1O_2$ | N1(C(CCCCCCCCCCC)=O)CCOCC1 | 45 | 0.2298 | 38 | 0.1701 | -0.7933 | -0.485 |
| 2-(Dodecyl-(2-hydroxyethyl)amino)ethanol | 5 | 12 | 273.46 | 4.332 | -3.62 | 6.4 | $C_{16}H_{35}N_1O_2$ | N(CCCCCCCCCCCC)(CCO)CCO | 52 | 0.2576 | 26 | 0.1285 | -0.9399 | 0.6376 |
| S,S-dimethyl-N-(2-methoxycarbonylbenzenesulfonyl)iminosulfurane | 5 | 1 | 275.35 | 2.117 | -2.94 | 0.19 | $C_{10}H_{13}N_1O_4S_2$ | S(C)(C)=NS(=O)(=O)c1ccccc1C(=O)OC | 69 | 0.7408 | 68 | 0.5671 | 0.6471 | 1.0529 |
| 1-Hexyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam | 4 | 6 | 280.41 | 2.69 | -2.29 | 2.1 | $C_{16}H_{28}N_2O_2$ | N2(C1C(=O)N(CCCCCC)CCCC1)C(=O)CCC2 | 32 | 0.1373 | 43 | 0.2 | 0.2511 | 0.3965 |
| 1-Dodecanoyl-2-piperidinone | 3 | 11 | 281.44 | 5.36 | -4.165 | 7.7 | $C_{17}H_{31}N_1O_2$ | N1(C(CCCCCCCCCCC)=O)CCCCC1=O | 43 | 0.2173 | 48 | 0.2246 | -1.1692 | -0.5328 |
| Azone, laurocapram | 2 | 12 | 281.48 | 6.254 | -4.85 | 22.1 | $C_{18}H_{35}N_1O_1$ | C(CCCCCCCCCC)CN1CCCCCC1=O | 49 | 0.2411 | 45 | 0.2074 | -1.5749 | -1.2056 |
| N-Dodecyl-N-(2-methoxyethyl)acetamide | 3 | 12 | 285.47 | 5.22 | -4.375 | 7.8 | $C_{17}H_{35}N_1O_2$ | N(CCCCCCCCCCCC)(CCOC)C(=O)C | 46 | 0.2307 | 58 | 0.3092 | -1.3048 | -0.537 |
| 4-(Dodecanoyl)-thiomorpholine | 3 | 11 | 285.49 | 5.422 | -4.49 | 21 | $C_{16}H_{31}N_1O_1S_1$ | N1(C(CCCCCCCCCCC)=O)CCSCC1 | 64 | 0.4557 | 54 | 0.2794 | -1.2844 | -0.5372 |
| 1-Dodecyl-3-phenylurea | 5 | 12 | 304.48 | 6.675 | -3.94 | 1.1 | $C_{19}H_{32}N_2O_1$ | O=C(Nc1ccccc1)NCCCCCCCCCCCC | 68 | 0.6672 | 65 | 0.4593 | -1.6594 | 0.5513 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2-Ppyrrolidinone-1-acetic acid dodecyl ester | 4 | 12 | 311.46 | 5.308 | -4.105 | 11 | $C_{18}H_{33}N_1O_3$ | C1C(=O)N(CC(=O)OCCCCCCCCCCCC)CC1 | 42 | 0.2143 | 6 | 0.0255 | -1.4312 | 0.1118 |
| N-Dodecyl-N-(2-methoxyethyl)butanamide | 3 | 12 | 313.52 | 6.002 | -5.045 | 6.1 | $C_{19}H_{39}N_1O_2$ | N(CCCCCCCCCCCC)(CCOC)C(=O)CCC | 58 | 0.3471 | 66 | 0.5081 | -1.7816 | -0.5301 |
| 1-Dodecyl-3-phenylthiourea | 5 | 12 | 320.54 | 6.614 | -5.085 | 3.4 | $C_{19}H_{32}N_2S_1$ | S=C(Nc1ccccc1)NCCCCCCCCCCCC | 6 | 0.0135 | 1 | 0.0104 | -2.019 | 0.5627 |
| 1-Decyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam | 4 | 10 | 336.52 | 4.682 | -3.945 | 8.8 | $C_{20}H_{36}N_2O_2$ | N2(C1C(=O)N(CCCCCCCCCC)CCCC1)C(=O)CCC2 | 14 | 0.0609 | 7 | 0.026 | -1.2245 | 0.2872 |
| Hexyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam-1-acetate | 6 | 6 | 338.45 | 2.706 | -2.605 | 1 | $C_{18}H_{30}N_2O_4$ | N2(C1C(=O)N(CC(OCCCCCC)=O)CCCC1)CCCC2=O | 29 | 0.1247 | 9 | 0.0279 | -0.2149 | 1.6916 |
| S,S-dimethyl-N-[4-[(4,6-dimethylpyrimidine-2yl)aminosulfonyl]phenyl]iminosulfurane | 7 | 1 | 338.45 | 1.777 | -3.83 | 0.83 | $C_{14}H_{18}N_4O_2S_2$ | S(C)(C)=Nc2ccc(S(=O)(=O)Nc1nc(C)cc(C)n1)cc2 | 1 | 0 | 49 | 0.2255 | 0.0823 | 2.3682 |
| S,S-dimethyl-N-[4-[(5-methoxypyrimidin-2yl)aminosulfonyl]phenyl]iminosulfurane | 8 | 1 | 340.43 | 1.495 | -4.3 | 0.72 | $C_{13}H_{16}N_4O_3S_2$ | S(C)(C)=Nc2ccc(S(=O)(=O)Nc1ncc(OC)cn1)cc2 | 56 | 0.3335 | 53 | 0.2785 | -0.0217 | 2.9355 |
| 1-Dodecyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam | 4 | 12 | 364.57 | 5.688 | -4.46 | 11 | $C_{22}H_{40}N_2O_2$ | N2(C1C(=O)N(CCCCCCCCCCC)CCCC1)C(=O)CCC2 | 9 | 0.052 | 20 | 0.0986 | -1.8868 | 0.2444 |
| Octyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam-1-acetate | 6 | 8 | 366.5 | 3.704 | -3.505 | 2.2 | $C_{20}H_{34}N_2O_4$ | N2(C1C(=O)N(CC(OCCCCCCCC)=O)CCCC1)CCCC2=O | 7 | 0.0161 | 14 | 0.0361 | -0.971 | 1.6338 |
| 1-Tetradecyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam | 4 | 14 | 392.63 | 6.432 | -4.97 | 18.6 | $C_{24}H_{44}N_2O_2$ | N2(C1C(=O)N(CCCCCCCCCCCCCC)CCCC1)C(=O)CCC2 | 55 | 0.3323 | 50 | 0.2279 | -2.4954 | 0.2216 |
| Decyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam-1-acetate | 6 | 10 | 394.55 | 4.648 | -4.16 | 4 | $C_{22}H_{38}N_2O_4$ | N2(C1C(=O)N(CC(OCCCCCCCCCC)=O)CCCC1)CCCC2=O | 10 | 0.0528 | 22 | 0.1078 | -1.6556 | 1.59 |
| N, N-didodecylacetamide | 2 | 12 | 395.71 | 9.684 | -6.55 | 8.9 | $C_{26}H_{53}N_1O_1$ | C(CCCCCCCCCC)CN(CCCCCCCCCCC)C(=O)C | 61 | 0.4104 | 63 | 0.4042 | -3.3109 | -1.1543 |
| 1,3-Didodecylurea | 5 | 12 | 396.7 | 9.484 | -6.85 | 1.9 | $C_{25}H_{52}N_2O_1$ | O=C(NCCCCCCCCCCC)NCCCCCCCCCCC | 33 | 0.1442 | 17 | 0.0395 | -3.45 | 0.5275 |
| 1,3-Didodecylthiourea | 5 | 12 | 412.77 | 9.702 | -7.81 | 1.6 | $C_{25}H_{52}N_2S_1$ | S=C(NCCCCCCCCCCC)NCCCCCCCCCCC | 38 | 0.1945 | 46 | 0.2163 | -3.8198 | 0.5255 |
| 1-Hexadecyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam | 4 | 16 | 420.68 | 7.382 | -5.455 | 9.6 | $C_{26}H_{48}N_2O_2$ | N2(C1C(=O)N(CCCCCCCCCCCCCCC)CCCC1)C(=O)CCC2 | 12 | 0.0588 | 34 | 0.1554 | -3.1391 | 0.1842 |
| Dodecyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam-1-acetate | 6 | 12 | 422.61 | 5.652 | -4.72 | 9.1 | $C_{24}H_{42}N_2O_4$ | N2(C1C(=O)N(CC(OCCCCCCCCCCC)=O)CCCC1)CCCC2=O | 35 | 0.1738 | 36 | 0.1594 | -2.3287 | 1.5456 |
| Tetradecyl-3-(2-oxo-1-pyrrolidine)-E-caprolactam-1-acetate | 6 | 14 | 450.66 | 6.394 | -5.29 | 9.6 | $C_{26}H_{46}N_2O_4$ | N2(C1C(=O)N(CC(OCCCCCCCCCCCCC)=O)CCCC1)CCCC2=O | 37 | 0.1853 | 18 | 0.0423 | -2.9517 | 1.5205 |

Table 2. An Example Confusion Matrix.

| | **Predicted negatives** | **Predicted positives** |
|---|---|---|
| **Actual negatives** | True Negatives (TN) | False Positives (FP) |
| **Actual positives** | False Negatives (FN) | True Positives (TP) |

Table 3. Comparison of statistical measures of model quality for fitted linear regression (FLR), simple linear regression (SLR) and $\nu$-and $\varepsilon$-Support Vector Regression models using three (MW, CC and HB) or five descriptors MW, CC, HB, log P and log S). **Note**: MSE represents the Mean Square Error and CORR the linear regression coefficient.

| **MODEL** | **MSE** | **CORR** |
|---|---|---|
| FLR (three features) | 0.11 | 0.76 |
| SLR (three features) | 0.12 | 0.73 |
| SLR (five features) | 0.11 | 0.76 |
| $\nu$-SVR with linear kernel (three features) | 0.13 | 0.72 |
| $\nu$-SVR with linear kernel (five features) | 0.11 | 0.76 |
| $\varepsilon$-SVR with linear kernel (three features) | 0.12 | 0.74 |
| $\varepsilon$-SVR with linear kernel (five features) | 0.12 | 0.75 |

Table 4. Comparison of statistical measures of model quality for Gaussian Process Regression (GPR) and ν-and ε-Support Vector Regression models for models with Radial Basis Function (RBF) kernels using different numbers of descriptors. **Note**: MSE represents the Mean Square Error and CORR the linear regression coefficient.

| **MODEL** | **MSE** | **CORR** |
|---|---|---|
| GPR (three features) | 0.11 | 0.77 |
| GPR (three features) | 0.11 | 0.77 |
| ν-SVR with RBF kernel (three features) | 0.13 | 0.74 |
| ν-SVR with RBF kernel (three features) | 0.10 | 0.79 |
| ε-SVR with RBF kernel (three features) | 0.12 | 0.74 |
| ε-SVR with RBF kernel (three features) | 0.10 | 0.78 |
| GPR with four features | 0.14 | 0.68 |
| GPR with two features | 0.14 | 0.68 |

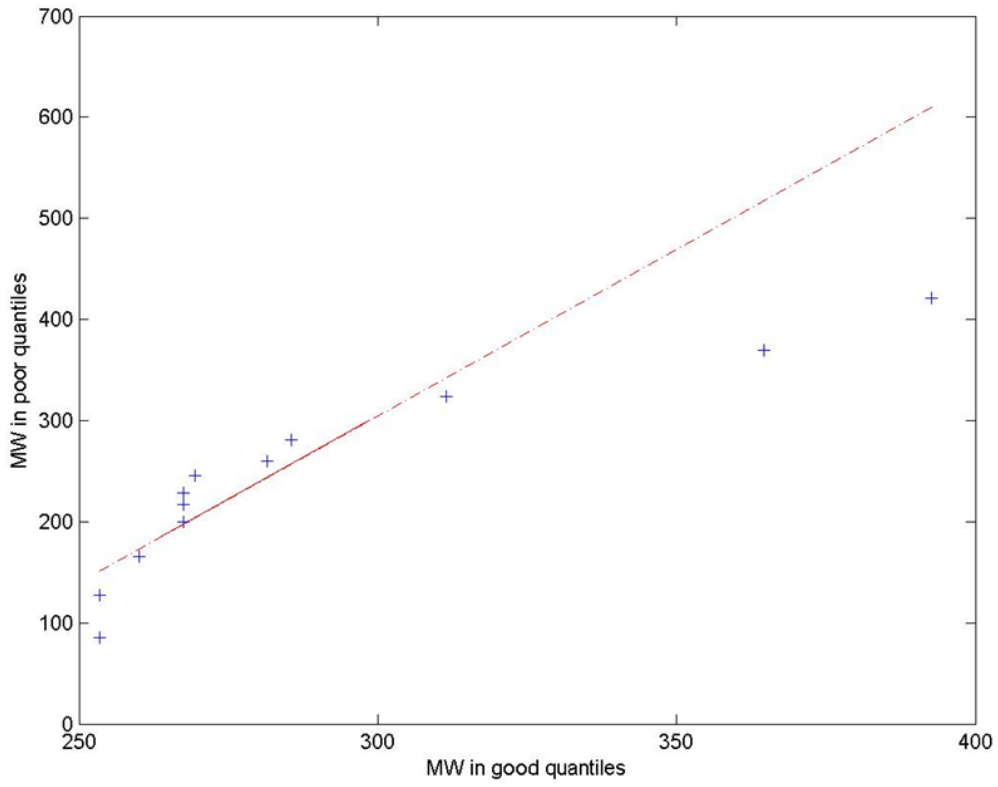


Figure 1: A quantile-quantile plot of MW for those enhancers classified as “good” and “poor”.

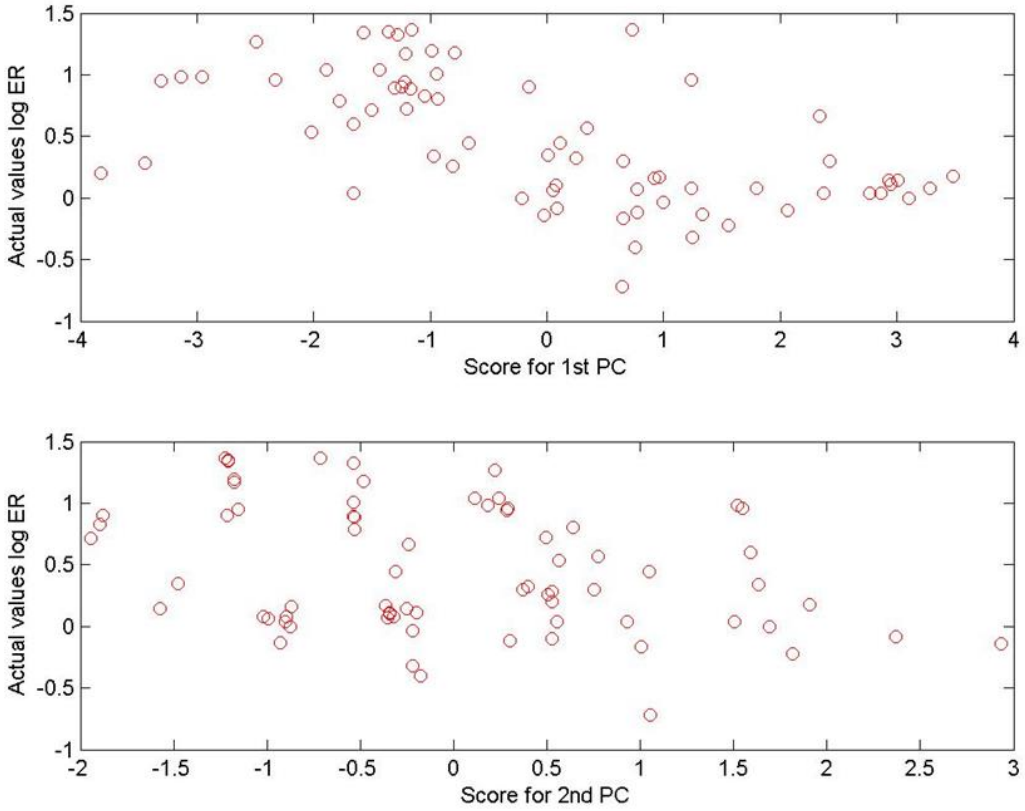


Figure 2: Plot of the target ER against the first two principal components.

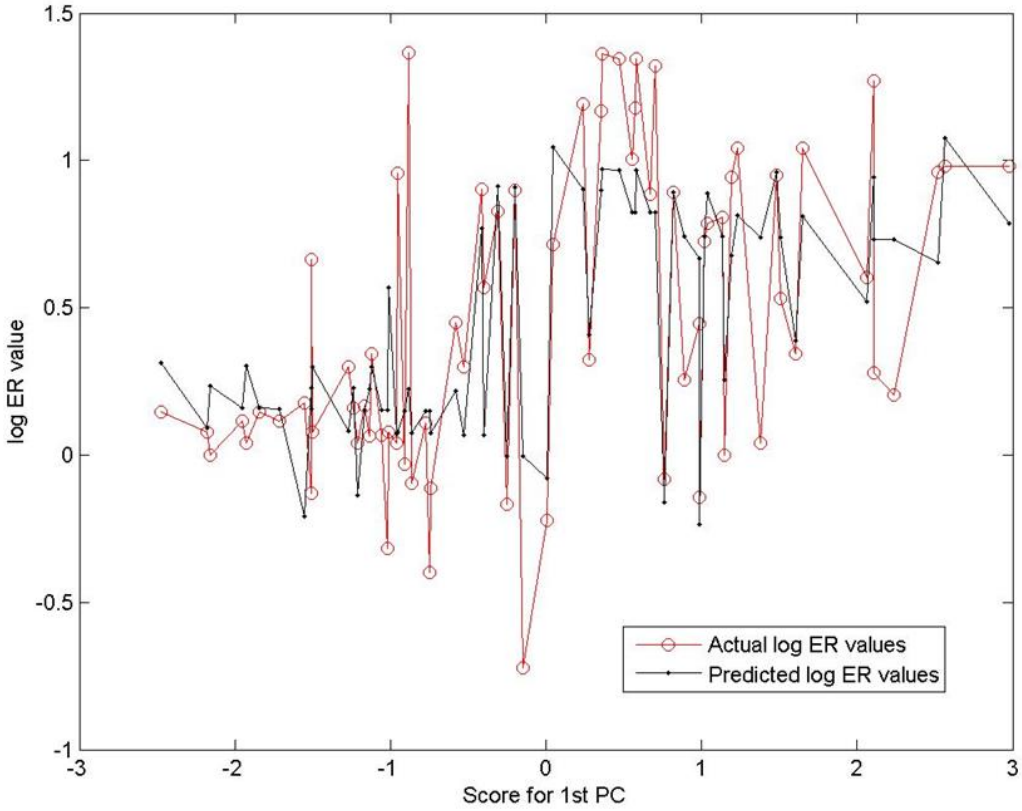


Figure 3: Plot of the first principal component against the target enhancement ratio (logER) and the predicted logER from FLR analysis.

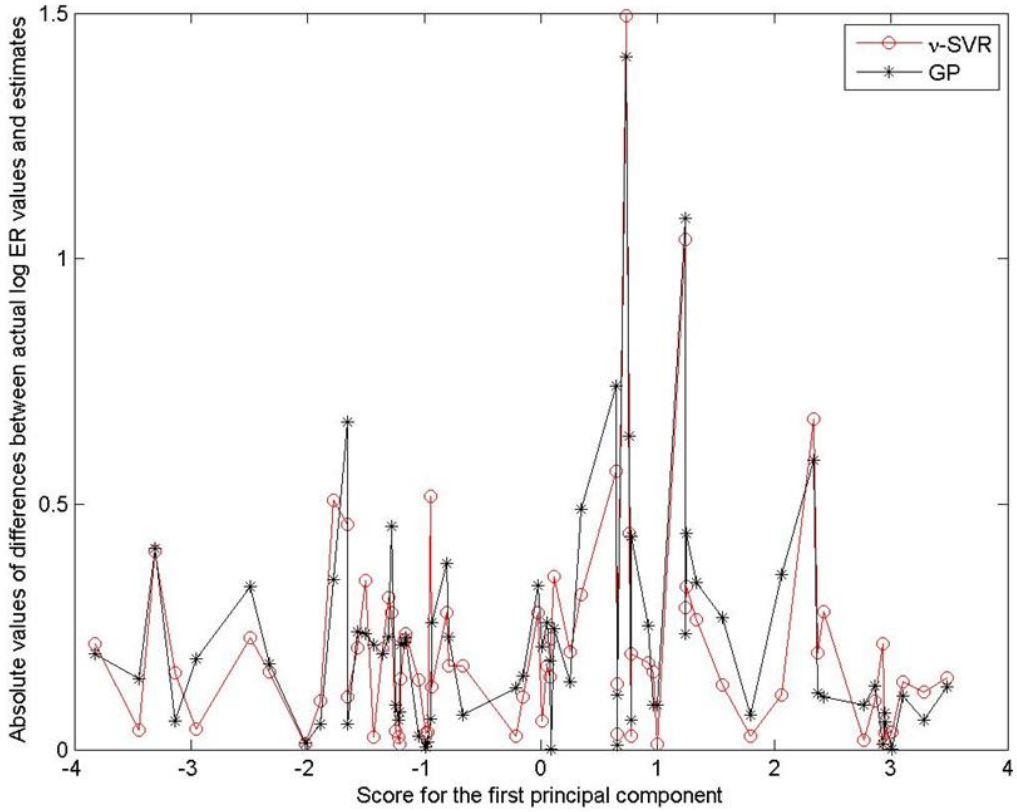


Figure 4: Plot of the relative values of differences of actual logER values and estimates against the first principal component.

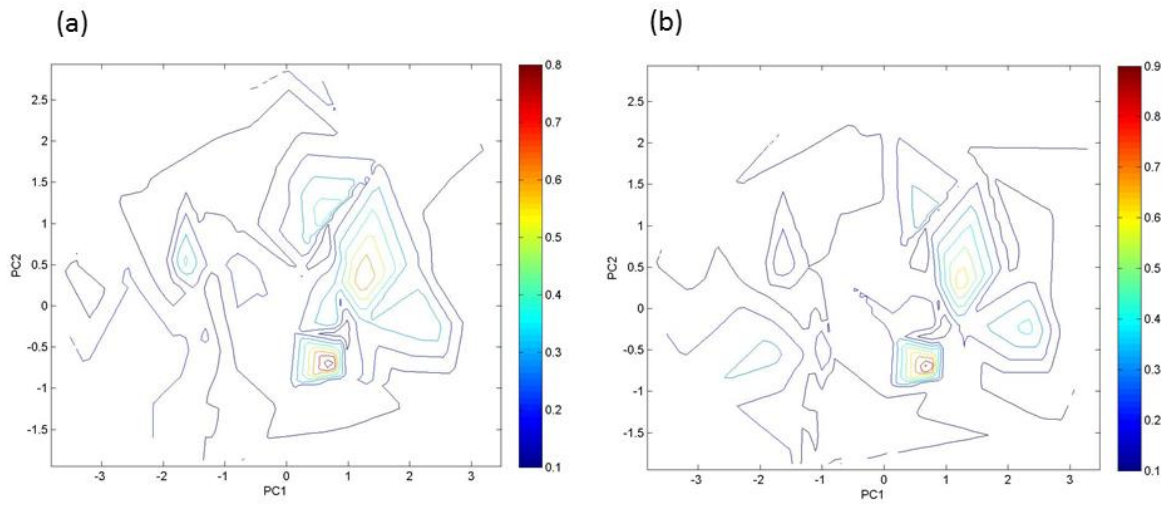


Figure 5: Contour plots of the principal components for (a) Gaussian Process Regression (GPR) and (b) Support Vector Machine (SVR) methods.