Model fitting for small skin permeability datasets: hyperparameter optimisation in Gaussian Processes regression

Ashrafi P[1], Sun Y[1], Davey N[1], Adams RG[1], Wilkinson SC[2], Moss GP[3]*

[1]School of Computer Science, University of Hertfordshire, Hatfield, UK;
[2]Medical Toxicology Centre, Wolfson Unit, Medical School, University of Newcastle-upon-Tyne, UK;
[3]The School of Pharmacy, Keele University, Keele, UK;

*Corresponding author: g.p.j.moss@keele.ac.uk
+44(0)1782 734 776
The School of Pharmacy
Keele University
Keele, Staffordshire, UK
ST5 5BG

Declarations of interest

The authors have no conflicts of interest to report.

Submission declaration; acknowledgements and funding

The authors confirm that this submission conforms to the journal's requirements.

The authors would like to thank the University of Hertfordshire and Keele University for supporting this study.

ABSTRACT

Objectives

The aim of the current study is to investigate how to improve predictions from Gaussian Process models by optimising the model hyperparameters.

Methods

Optimisation methods, including Grid Search, Conjugate Gradient, Random Search, Evolutionary Algorithm and Hyper-prior, were evaluated and applied to previously published data. Data sets were also altered in a structured manner to reduce their size, which retained the range, or 'chemical space' of the key descriptors in order to assess the effect of the data range on model quality.

Key findings

The Smoothbox Hyper-prior kernel results in the best models for the majority of data sets and they exhibited significantly better performance than benchmark QSPR models. When the data sets were systematically reduced in size the different optimisation methods generally retained their statistical quality whereas benchmark QSPR models performed poorly.

Conclusions

The design of the data set, and possibly also the approach to validation of the model, are critical in the development of improved models. The size of the data set, if carefully controlled, was not generally a significant factor for these models and that models of excellent statistical quality could be produced from substantially smaller data sets.

Key words:

Gaussian Process

Machine Learning

Skin Permeability

Hyperparameters

Quantitative structure-permeability relationship (QSPR)

INTRODUCTION

Measurement of the percutaneous absorption of exogenous chemicals has become increasingly important over the last 25 years for a variety of reasons, including pharmaceutical efficacy and, in a number of fields, toxicity. The current 'gold standard' for initial assessment of *in vitro* percutaneous absorption is an experiment using excised human or porcine skin and which follows the protocol presented in OECD 428 [1].

Since the publication of the Flynn data set [2] there has been considerable interest in the development of mathematical models that relate the percutaneous absorption of exogenous chemicals to the physicochemical properties of permeants. This began with the work of El Tayer [3] and has grown into a distinct area of research, mostly based on the use of a range of methods to interrogate the Flynn data set, or variations thereon. The early work in this field was predominately based on quantitative structure-permeability relationships (QSPRs) and has been comprehensively reviewed previously [4].

However, in the context of percutaneous absorption many QSPR models have been shown to be significantly limited in their predictive ability, for example where some of the most commonly used QSPR models were shown to poorly correlate with experimental data which covered the stated range of applicability of these models [5, 6]. Despite their advantages QSPRs have therefore gained little widespread use or credibility in the broader field of percutaneous absorption.

More recently, a range of novel methods has been applied to this problem domain. Such methods, including the use of non-linear models [55, 56], parallel artificial membrane permeability assay (PAMPA) methods [56] and Machine Learning methods such as Gaussian Process Regression [7], offer significant improvements in predictive ability over QSPR models. However, they are often criticised as non-linear methods are perceived to over-fit in many situations and Machine Learning methods are limited by their lack of transparency as they are predominately based on 'black-box' methods, which mean that they are seldom represented by a discrete algorithm. Despite studies which in different ways address this issue [8, 9] the uptake of such methods in the field of percutaneous absorption has been limited and is due mostly to the lack of ease of use of what can often be quite advanced computation techniques by non-specialists. Nevertheless, despite their more rudimentary

nature when compared to Machine Learning methods, and previous studies highlighting comparatively poor performance for QSPR methods compare to Machine Learning methods [8, 20, 24], QSPRs are still considered by many researchers in this field to be the benchmark predictive method and are used in this study in that regard.

Another significant limitation in using computational methods in estimating percutaneous absorption is the construction of the model and, implicitly, the need for a high-quality and consistent data set to underpin this development. The necessary amounts of reliable and consistent data have been discussed previously [10]. From the Machine Learning point of view, there is considerable difficulty in using Flynn's original dataset and other datasets derived from it in that the reported value of skin permeability for the same chemical varies considerably. This may be due to experimental artefacts, such as the anatomical location from which skin was excised for each experiment, or experimental temperature, which may affect the accuracy of resultant models [11]. This presents a significant challenge in the production of a new data set from a single source, which may be expected to yield more accurate models with reduced variance.

Nevertheless, one of the key issues in the development of improved models is the difficulty of developing new data sets. For example, a contract research organisation will commonly charge a significant sum to produce absorption data for one chemical (i.e. one data point) and the production of approximately 100 data points using the same method to construct a viable model is therefore, in purely financial terms, very costly and in all probability unrealistic. Thus, generation of new datasets may not reflect the needs of model development which sits apart from a specific study. In particular, industrially-focused studies may be targeted to a specific group of chemicals and this may not fit the needs of a model. In addition, data quality may be affected by variable methodological approaches or by the collation of data from a range of studies.

The aims of this study are two-fold. Firstly, to investigate how model optimisation can take place with relatively small data sets. In particular, we investigate how the three hyper-parameters control the Matérn kernel function involved in the Gaussian Process Regression methods. These include $\theta = \{\sigma_f^2, \sigma_n^2, \iota^2\}$, where ι is the characteristic length-scale, σ_f^2 is the signal variance, and σ_n^2 is the noise variance. And secondly, this study aims to investigate how the nature of data will affect the viability of the resulting model. Thus, this study will

empirically demonstrate that the optimisation of hyperparameters can be used with small datasets to produce highly predictive models and that dataset generation is also central to model quality and predictivity.

METHODS

1. Data Sets

Nine human and animal skin datasets collated from various sources have been used in this study. All data has been taken from previously published literature studies and does not require ethical approval for its subsequent use. The sizes of the datasets vary from 14 to 85 after refining the data by, for example, removing ambiguous data or values which are listed as 'greater than' or 'less than' a fixed value, rather than a discrete number. Other refinement processes include removing all the repetitions and obtaining the mean value of the targets for the same chemicals with the same molecular features and different target values [9, 12]. The number of data records in each dataset after refinement is shown in Table 1. The small size is due to the fact that gathering consistent pharmaceutical data which is generated from the same or similar protocols is difficult, time consuming, and expensive. This is usually because of the inherent biological variation of such data, and that the data is generated for other purposes and not primarily for its inclusion in predictive models. Table 2 shows the whole data set, originally obtained from Magnusson's *Set A* (see Table 1), which is used for analysis of subsets.

[INSERT TABLE 1 HERE]

[INSERT TABLE 2 HERE]

2 Gaussian Process Regression (GPR)

Gaussian Process Regression (GPR) is a technique of increasing importance in the Machine Learning field, and which is finding greater utility in the physical and biological sciences [8, 13 – 16, 22]. This technique has been reported on and reviewed extensively elsewhere and the reader is directed to those sources for further information [9, 15 – 24].

It is possible that inferring the hyperparameters from the data could be particularly problematic with small datasets. To resolve this, various optimisation methods have been used to obtain the hyperparameters that minimise negative log marginal likelihood values.

The methods used include the Conjugate Gradient, Grid Search, Random Search, Hyper-Prior and Evolutionary Algorithm methods.

3. Experimental Set-Up

3.1 Software

A range of methods were used for analysis of the data. Gaussian Process methods with a range of kernels, and a range of methods to vary the model hyperparameters (the Conjugate Gradient, Grid Search, Random Search, Hyper-Prior methods and Evolutionary Algorithms) were employed. The Gaussian Process modelling methods for non-linear regression used previously were again adopted for this study [7, 8, 19, 22]. The latest version of the Hyper Prior optimisation Toolbox was also used [21]. The MatLab Genetic Algorithm (GA) optimisation toolbox was used to carry out the Evolutionary Algorithm hyperparameter optimisation. Quantitative structure-permeability relationships (QSPRs) were used as benchmarks [25, 26].

3.2 Cross-validation

The importance of model validation in constructing computational models has been discussed previously [27]. In this study, we have validated models using the cross-validation technique [28]. 5-fold cross-validation was performed. The datasets were shuffled and divided into 5 'folds'. Each time one of the folds was considered as the test set and the remaining four were considered as the training set. At this point, a validation set was removed from the training set. The hyperparameter optimisation methods were then applied to the training set and the prediction performances were gained for the validation set. This was then repeated for the other 3 possible validation sets. The best hyperparameters were chosen as those performed best over the four validation sets (the minimum average MSLL values, which are defined in Section 4). They were used to predict the permeability values of the test set.

3.3 Initialisation of experiments

The experiments were initialised as follows:

- Grid search: The hyperparameters were considered as a range [10^{-3}, 10^{3}] with 20 equidistant steps. Using a 5-fold cross validation the model was trained with all the 8,000 (20 x 20 x 20) different sets of the hyperparameters and the predictions obtained

for the test sets. On inspecting the prediction performances on the validation sets a finer search for better values of the hyperparameters was then performed with the search range limited to [0.01, 10] with 20 steps as no better results were obtained using the hyperparameters outside this range. The model was then trained with the new hyperparameters and tested on the test sets. The average values and their standard deviation among 5-folds were then reported.

- Random search: 20 values for each hyperparameter were obtained randomly within the same range [0.01, 10] considered in the grid search. Using 5-fold cross validation the model was then trained and the predictions obtained. Since, in each run of this experiment, the hyperparameters were selected randomly the experiment was repeated 5 times and the results were obtained by calculation of the mean and standard deviation of the experiment's results.
- Conjugate gradient: The hyperparameters were initialised to log (0.5) with the number of function evaluations set to 100.
- Hyper Prior methods: The mean and variance parameters of the Gaussian and Laplacian priors were set to constant values of 0.1 and 0.01, respectively and were obtained as the best prediction performances using cross-validation in each of the data sets. For the Smooth Box Prior method, *a*, *b and* η values were set to 10^{-3}, 10 and 2, respectively. Various values of η were evaluated and the value 2 was found to be the best value for the data sets used in this study.
- Evolutionary algorithm: Following an evaluation of ratios ranging from 0.1 to 1.2, the heuristic crossover function with a ratio of 0.7 was used to accelerate convergence as it was found to have the optimum performance for the data sets used. Each of the 50 generations has a population of 50 and the optimised hyperparameters were obtained from the last generation. The 'Elite' Children value was set to 4 and the mutation function was kept uniform, meaning that the children were randomly selected from a uniform distribution within the range of hyperparameters. The crossover fraction was set to 0.8 (0.8 * 50 = 40), meaning that the rest of the children in a population are 4 Elite children and 6 children were obtained from mutation. The population of the first generation was initialised randomly and was therefore similar to the Random Search. This experiment was repeated five times using the Genetic Algorithm Toolbox in MatLab.

3.4 Data set analysis

The different data sets used in this study were characterised in terms of their membership (data set size) and range (the range of physicochemical descriptors used). Data used are those published previously [29, 30] and are shown in Tables 1 and 2.

3.5 The effect of the size of the data set and the range of the physicochemical descriptor values on prediction performance

Due to their ubiquitous use in this field, and their relevance as benchmarks in this study, the effects of molecular weight and lipophilicity (as log P or log $K_{o/w}$) were considered [4]. The first experiment considered how changes to the size (membership) of the data set affected the statistical quality of the resultant models whilst maintaining the range, or 'chemical space', of each model. The data set reported previously by Magnusson [29] was used for this experiment. In separate experiments this data set was used to construct four smaller subsets that maintained the range of descriptors of the original data set (Table 2). To construct these data sets four subsets (of size 44, 33, 17 and 9) were chosen from the Magnusson data set. Chemicals were selected only to ensure that the maximum and minimum MW ranges were maintained across all the data sets. The GPR model was then trained with each data set, with the hyper-prior Smoothbox and conjugate gradient optimisation methods employed to set the best hyperparameters for the models. As a benchmark the QSPR reported previously [26] was used, with a concentration correction to adjust between k_p and J_{max}, as the Potts and Guy QSPR model [25] did not perform well in the initial analysis. This experiment was repeated with subsets of the Magnusson data set which maintained the range of log P values across all data sets whilst reducing the data set membership. Subsets in both experiments were of the same size.

The final set of experiments involved creating four training sets of the Magnusson data set where the membership again was kept constant (at n = 40) to remove any effect associated with data set size. But, in these cases, the range of the physicochemical descriptor values examined (MW and log P) were systematically reduced by the generation of random subsets from the parent data set. A fixed test set was also produced; one-fifth of the Magnusson (Set A) was considered to be the test set and the training sets (including 5-fold cross-validation) were generated from the remaining data. The range of the first training set can be obtained by adding and subtracting the standard deviation of MW to and from the median of all MW values (excluding the values in the fixed test set). To keep the size of each training set the

same (n = 40), members of the subset were picked at random from the given range. To obtain the next training sets, the standard deviation is added by larger values (for example, 40, 100 and 200, respectively), and the same process is repeated. The GPR model was then trained using the smoothbox hyper-prior and conjugate gradient methods, and the predicted log J_{max} values were reported for the same test set. As data was chosen randomly within each data range, the experiment was repeated ten times, with the mean and standard deviations being reported for both the GPR models and the QSPR benchmark. The same methods were used to analyse changes to both MW and log P.

4. PERFORMANCE MEASURES

The correlation coefficient (*r*), Mean Standard Log Loss (MSLL) [22] and improvement over the naïve model (ION, where the naïve model always predicts the mean of the target value in the training set independently of the input), were used, as in previous studies, to determine the model performance [8, 20, 24].

ION measures how much better a predictor is than the naïve predictor. ION ranges from -∞ to 1, and greater positive ION values represent better performance. MSLL will be approximately zero for simple methods and negative for better methods [22]. The correlation coefficient ranges from -1 to 1 and in this study a high positive value defines good prediction performance [24].

RESULTS AND DISCUSSION

Selection of optimum hyperparameter method

The statistical measures (MSLL, ION and r) used to assess the quality of the different hyperparameter methods are shown in Table 3. The data sets in Table 3 are listed based on size, from the largest to the smallest, taken from the dataset published previously [29]. The best results for each data set are shown in bold text, and the worst result shown in underlined text.

[INSERT TABLE 3 HERE]

The MSLL results indicate that the smooth box hyper-prior kernel works better than the other methods for the majority of the datasets. It generally shows a good performance for all datasets irrespective of size. The ION and correlation coefficient results also show that this method results in better prediction performances for four of the datasets. A benchmark analysis using the Potts and Guy QSPR model [25], and comparing the correlation coefficients only, performs significantly worse than all the other methods in all the datasets. The results in Table 3 indicate that the hyper-prior Smoothbox method produces, independently of the performance measures used, the best overall performance for the majority of data sets. The inconsistency between ION and MSLL results may be a result of small data sets as the predictive variance, which is part of MSLL but not ION, is generally so much more variable in smaller data sets. Using the Evolutionary Algorithm (EA) to optimise the hyperparameters generally works better, in terms of performance measures, for larger data sets than for smaller data sets. In this study, the worst performances from application of the EA method are found with the smallest data sets.

Table 3 also shows that the outcomes from the grid search and random search hyperparameter optimisation methods were broadly similar in their performance measures. This partially mirrors previously reported findings [31]. Interestingly, in this study whilst both methods were generally positive they were not the best methods tested to optimise the hyperparameters. This may be due to the limitations of these methods in searching a space of three hyperparameters which are limited to a number of points in that space – in this case

this is 20 x 20 x 20 = 8000 – and that a manual manipulation of these spaces may optimise model performance. It also appears that small changes in certain hyperparameter values exerts a significant impact on the results generated by these techniques. The implication for this is that, for either small data sets or sources of variable data, small differences in the analytical techniques used to generate outputs may have significant implications for the accuracy of the resultant predictions.

It is also important to note that the hyper-prior optimisation method outperforms the conjugate gradient method, even though the latter is the method most commonly used to optimise hyperparameters in GPR [17]. This is shown in Figure 1, where the comparison of MSLL values is shown for a range of optimisation methods, and the Smoothbox hyper-prior method clearly outperforms the conjugate gradient method for the majority of data sets. A smaller standard deviation of MSLL is obtained when the hyper-prior method was used compared to the conjugate gradient method.

[INSERT FIGURE 1 HERE]

[INSERT FIGURE 2 HERE]

The effect of the size of the data set and the range of the physicochemical descriptor values on prediction performance

The results from altering the data set memberships are shown in Table 4. The most significant finding is that decreasing the size of the data set (from 85 to 9 members) but maintaining the maximum range of molecular weight does not significantly affect the good performance of the model. In all cases where the statistical measure does fall – for example, with the smallest data sets, the drop in the correlation coefficient, for example, is to 0.88 or 0.83, depending on the hyperparameter optimisation method used (Table 4). Overall, similar results are obtained for the different GPR hyperparameter optimisation methods.

[INSERT TABLE 4 HERE]

When the data set membership is decreased and the range of log P values kept constant the statistical quality of the models is not substantially affected. However, the outcomes of this

study are not as clear-cut as the previous experiment. While model performance increases in some cases with decreasing data sets – for example, increases in ION from 0.93 to 0.94 are observed, model performance declines in other cases. Such decreases are shown in Table 4 and include reduction in ION of 0.93 to 0.72, and from 0.93 to 0.80, for the Smoothbox and conjugate gradient methods, respectively. This illustrates not only the importance of a correct data set design when conducting modelling experiments [11] but also the importance of transparency in model construction and use [33]. This again highlights the importance of the range of significant physicochemical descriptors and how they may affect the resultant model and its predictions of skin permeability.

That the data range should be as wide as possible also has an implication on the descriptor choice, despite previous GPR studies [8, 24] indicating that a certain degree of interchangeability between parameters due to covariance might be significant in flexibly generating models of the same statistical quality. For example, an examination of previously published data sets [26, 29, 33] indicates that the majority of chemicals present in those data sets have a small number of hydrogen bonding groups – usually from zero to three. If the implications of these studies are valid, it may be hypothesised that little improvement in GPR models would be seen even if hyperparameter optimisation is conducted.

The final set of experiments involved creating subsets of the Magnusson data set where the membership again was kept constant (at n = 40) and the range of descriptors altered. The results of these experiments are shown in Table 4. They show that keeping the size of the data sets fixed and decreasing the range of MW is directly related to the model's performance. The same effect is not observed for changes to the log P range. These results imply that if the data sets that are used for training the model cover as wide a range of physicochemical descriptor values as possible then a good prediction performance can be expected [34].

Conclusions

Using the hyper-prior Smoothbox method to optimise the GPR hyperparameters works better than other hyperparameter optimisation methods and does so independently of the data and the performance measure methods used to characterise model quality. This method optimised GPR results in models with a better statistical performance than previous

GPR models where hyperparameters are not optimised [8, 24]. Both of these approaches are significantly better than established QSPR models [25, 26].

Whilst hyperparameter optimisation improved model quality and maintained the performance measures it should not be used in isolation; even in small data sets there was variation within the chosen method of hyperparameter optimisation, with the Smoothbox method producing the best outcomes in the majority of situations. Investigation of the physicochemical descriptors used in this data set suggests that the data set range and not necessarily the population should be as wide as possible.

The nature of the analysis is also examined in this study. Comparison of data sets where the membership is kept constant whilst the range of significant chemical features is altered generally indicated that the range of test and training sets needs to be maintained, as it may be inferred that not doing so may lead to issues of variability in performance due to how the model is trained, and with which data the model is tested with.

Thus, a consistent approach to data set design is recommended. Models should not simply be constructed based on the addition of all available data to a large data set, but rather should consider the effective and accurate range of the model and whether additional data actually helps the model – in this study it is clear that additional data does not add significantly to model quality in some cases. This may be extended into considerations of which physicochemical or experimental parameters are used to construct the model and whether any parameters limit the quality of the model. This study again shows that GPR models outperform QSPR models in a 'chemical space' in which those models should be effective [6]. The most significant implication is that a high quality model can be constructed from a relatively small data set. Such a model can cover a wide 'chemical space' but, given the improvement observed by the optimisation of the hyperparameters of the GPR model the construction of high quality models with significant real-world relevance is now readily achievable with fewer data than before.

References

1. OECD Guidelines for the Testing of Chemicals, Section 4: Health Effects. Test No. 428: Skin Absorption: In Vitro Method, OECD, 2004. [Available online at: http://www.oecd-ilibrary.org/environment/test-no-428-skin-absorption-in-vitro-method_9789264071087-en; Accessed 12th April 2017].
2. Flynn GL. Physicochemical determinants of skin absorption. In: Gerrity TR, Henry CJ (eds.) Principles of Route-to-Route Extrapolation for Risk Assessment. New York: Elsevier, 1990, pp93-127.
3. El Tayar N, Tsai RS, Testa B, Carrupt PA, Hansch C, Leo A. Percutaneous penetration of drugs – a Quantitative Structure-Permeability Releationship study. J. Pharm. Sci. 1991: 80, 744-749.
4. Moss GP, Dearden JC, Patel H, Cronin MTD. Quantitative structure-permeability relationships (QSPRs) for percutaneous absorption. Tox. In Vitro. 2002: 16, 299-317.
5. Moss GP, Gullick DR, Cox PA, Alexander C, Ingram MJ, Smart JD, Pugh WJ: Design, synthesis and characterisation of captopril prodrugs for enhanced percutaneous absorption. J. Pharm. Pharmacol. 2006: 58, 167–177.
6. Mitragotri S, Anissimov YG, Bunge AL, Frasch HF, Guy RH, Hadgraft J, Kasting GB, Lane ME, Roberts MS. Mathematical models of skin permeability: An overview. Int. J. Pharm. 2011: 418, 115-129.
7. Moss GP, Sun Y, Wilkinson SC, Davey N, Adams R, Martin GP, Prapopoulou M, Brown MB. The application and limitations of mathematical models across mammalian skin and poldimethylsiloxane membranes. J. Pharm. Pharmacol. 2011: 63, 1411-1427.
8. Lam LT, Sun Y, Davey N, Adams RG, Prapopoulou M, Brown MB, Moss GP. The application of feature selection to the development of Gaussian process models for percutaneous absorption. J. Pharm. Pharmacol. 2010: 62, 738–749.
9. 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.
10. Grass GM, Sinko PJ. Effect of diverse datasets on predictive capability if ADME models in drug discovery. Drug Discovery Today, 2001, 6 (Suppl. 1) 54-61.
11. Moss GP, Gullick DR, Wilkinson SC. Predictive methods in percutaneous absorption. Springer: Heidelberg, 2015.
12. Moss GP, Shah AJ, Adams RG, Davey N, Wilkinson SC, Pugh WJ, 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.
13. Obrezanova O, Csanyi G, Gola JMR, Segall MD. Gaussian Processes: A method for automatic QSAR modelling of ADME properties. J. Chem. Info. Mod. 2007: 47, 1847-1857.
14. Schroeter T, Schwaighofer A, Mika S, Ter Laak A, Suelzle D, Ganzer U, Heinrich N, Muller KR. Machine Learning Models for Lipophilicity and Their Domain of Applicability. Mol. Pharmaceutics, 2007, 4(4), 524-538
15. Mellor J, Grigoras I, Carbonell P, Faulon J-L. Semi-supervised Gaussian Process for automated enzyme search. J. Chem. Info. Mod. 2016: 5, 518-528.
16. Rahman M, Previs SF, Kasumov T, Sadygov RG. Gaussian Process modelling of protein turnover. J. Proteome Res. 2016: 15, 2115-2122.
17. Blum M, Riedmiller MA. Optimization of Gaussian Process hyperparameters using *rprop*. ESANN, 2013.

18. Brown MB, Lau C-H, Lim ST, Sun Y, Davey N, Moss GP, Yoo S-H, de Muynck C. An evaluation of the potential of linear and nonlinear skin permeation models for the prediction of experimentally measured percutaneous drug absorption. J. Pharm. Pharmacol. 2012: 64, 566-577.
19. MacLaurin D, Duvenaud D, Adams RP. Gradient-based hyper-parameter optimization through reversible learning. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: JMLR: W&CP volume 37. [Available at: http://hips.seas.harvard.edu/files/maclaurin-hypergrad-icml-2015.pdf; Accessed 10 April 2017].
20. Moss GP, Sun Y, Davey N, Adams R, Pugh WJ, Brown MB. The application of Gaussian Processes to the prediction of percutaneous absorption. J. Pharm. Pharmacol. 2009: 61, 1147-1153.
21. Rasmussen CE, Nickish H. The GPML Toolbox Version 3.5. [Available at: http://mlg.eng.cam.ac.uk/carl/gpml/doc/oldcode.html]; Accessed 10 April 2017].
22. Rasmussen CE, Williams KI. Gaussian Processes for Machine Learning. 2006, Boston, The MIT Press. [Available online at: http://www.gaussianprocess.org/gpml/chapters/RW.pdf; Accessed 10 April 2017].
23. Snelson EL. Flexible and efficient Gaussian Process models for Machine Learning. University College London, PhD Thesis, 2007.
24. Sun Y, Adams R, Davey N, Moss GP, Prapopopolou M, Brown MB. The application of Gaussian processes in the predictions of permeability across mammalian and polydimethylsiloxane membranes. Art. Int. Res. 2012: 1, 86-98.
25. Potts RO, Guy RH. Predicting skin permeability. Pharm. Res. 1992: 9, 663-669.
26. Moss GP, Cronin MTD. Quantitative structure-permeability relationships for percutaneous absorption: re-analysis of steroid data. Int. J. Pharm. 2002: 238, 105-109.
27. Tropsha A. Best Practices for QSAR Model Development, Validation and Exploitation. Mol. Informatics, 2010, 29(6-7), 476-488
28. Bishop CM. Neural Networks for Pattern Recognition. Clarendon Press, Oxford. 1995
29. Magnusson BM, Anissimov YG, Cross SE, Roberts MS. Molecular size as the main determinant of solute maximum flux across the skin. J. Invest. Dermatol. 2004: 122, 993-999.
30. Prapopoulou M. The development of a computation / mathematical model to predict drug absorption across the skin. King's College London, PhD Thesis, 2012.
31. Bergstra J, Benigo Y. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 2012: 13, 281-305.
32. Cronin MTD, Schultz TW. Pitfalls in QSAR. J. Mol. Struct. 2003: 622, 39–51.
33. Cronin MTD, Dearden JC, Moss GP, Murray-Dickson G. Investigation of the mechanism of flux across human skin in vitro by quantitative structure-permeability relationships. Eur. J. Pharm. Sci. 1999: 7, 325–330.
34. Zhang Q, Li P, Liu D, Roberts MS. Effect of vehicles on the maximum transepidermal flux of similar size phenolic compounds. Pharm. Res. 2013: 30, 32-40.
35. Anderson B, Higuchi W, Raykar P. Heterogeneity effects on permeability–partition coefficient relationships in human stratum corneum. Pharm. Res 1988: 5,566–573.
36. Anderson B, Raykar P. Solute structure-permeability relationship in human stratum corneum. J. Invest. Dermatol. 1989: 93, 280–286.
37. Barber E, Teetsel N, Kolberg K, Guest D. A comparative study of the rates of in vitro percutaneous absorption of eight chemicals using rat and human skin. Fundam. Appl. Toxicol. 1992: 19, 493–497.
38. Blank I, McAuliffe D: Penetration of benzene through human skin. J. Invest. Dermatol. 1985: 85:522–526.

39. Blank I, Scheuplein R, MacFarlane D. Mechanism of percutaneous absorption. III. The effect of temperature on the transport of non-electrolytes across the skin. J. Invest. Dermatol. 1967: 49, 582–589.
40. Bronaugh R, Congdon E. Percutaneous absorption of hair dyes: Correlation with partition coefficients. J. Invest. Dermatol. 1984: 83, 124–127.
41. Bronaugh R, Stewart R, Simon M. Methods for in vitro percutaneous absorption studies. VII. Use of excised human skin. J. Pharm. Sci. 1986: 75, 1094–1097.
42. DalPozzo A, Donzelli G, Liggeri E, Rodriguez L. Percutaneous absorption of nicotinic acid derivatives in vitro. J. Pharm. Sci. 1991: 80, 54–57.
43. Dick I, Scott R. Pig ear skin as an in vitro model for human skin permeability. J. Pharm. Pharmacol. 1992: 44, 640–645.
44. Liu P, Higuchi W, Ghanem A, Good W. Transport of beta-estradiol in freshly excised human skin in vitro: Diffusion and metabolism in each skin layer. Pharm. Res. 1994: 11, 1777–1784.
45. Parry G, Bunge A, Silcox G, Pershing L, Pershing D. Percutaneous absorption of benzoic acid across human skin. I. In vitro experiments and mathematical modeling. Pharm. Res. 1990: 7, 230–236.
46. Peck K, Ghanem A, Higuchi W. The effect of temperature upon the permeation of polar and ionic solutes through human epidermal membranes. J. Pharm. Sci. 1995: 84, 975–982.
47. Roberts M. Percutaneous absorption of phenolic compounds; PhD Thesis, University of Sydney, Sydney, 1976.
48. Roberts M, Anderson R, Swarbrick J. Permeability of human epidermis to phenolic compounds. J. Pharm. Pharmacol. 1977: 29, 677–683.
49. Scheuplein R, Blank I, Brauner G, MacFarlane D. Percutaneous absorption of steroids. J. Invest. Dermatol. 1969: 52, 63–70.
50. Siddiqui O, Roberts M, Polack A. Percutaneous absorption of steroids: Relative contributions of epidermal penetration and dermal clearance. J. Pharmacokinet. Biopharm. 1989: 17, 405–424.
51. Singh P, Roberts M. Dermal and underlying tissue pharmacokinetics of lidocaine after topical application. J. Pharm. Sci. 1994: 83, 774–781.
52. Southwell D, Barry B, Woodford R. Variations in permeability of human skin within and between specimens. Int. J. Pharm. 1984: 18, 299–309.
53. Williams A, Barry B. Terpene and the lipid–protein-partitioning theory of skin penetration enhancement. Pharm. Res. 1991: 8, 17–24.
54. Williams A, Cornwell P, Barry B. On the non-Gaussian distribution of human skin permeabilities. Int. J. Pharm. 1992: 86, 69–77.
55. Khajeh A, Modarress H. Linear and nonlinear quantitative structure-property relationship modelling of skin permeability. SAR QSAR Environ. Res. 2014: 25, 35-50.
56. Neely BJ, Madihally SV, Robinson RL, Gasem KA. Nonlinear quantitative structure-property relationship modeling of skin permeation coefficient. J. Pharm. Sci. 2009: 98, 4069-4084.
57. Dobricic V, Markvoic B, Nikolic K, Savic V, Vladimirov S, Cudina O. 17b-carboxamide steroids – in vitro prediction of human skin permeability and retention using PAMPA technique. Eur. J. Pharm. Sci. 2014: 14, 52-95.

Captions for Figures and Tables

Figure 1.
Comparison of MSLL performances for the Conjugate Gradient and Hyper-prior Smoothbox methods for each dataset.

Figure 2. Range of physicochemical descriptors in the datasets used in this study.

Table 1.
Summary of the data sets used in this study.

Table 2.
Dataset used for analysis of subsets. Data is taken from Magnusson's study [29, 35 – 54] and subdivisions of this data are shown for studies where the systematic reduction of dataset size was undertaken whilst retaining the range of key parameters (log K_{ow}, MW).

Note: Where the log K_{ow} (log P) range is maximised, the range is from -4.67 to 4.52 for datasets of all sizes and MW ranges are: dataset (n = 9), 46 to 316.5; dataset (n = 17), 18 to 434.5; dataset (n = 33), 32 to 434.5; dataset (n = 44), 32 to 476.6.
Where the MW range is maximised, the range is from 18 to 476.6 for all sizes and log K_{ow} (log P) ranges are: dataset (n = 9), -4.6 to 4.04; dataset (n = 17), -4.67 to 4.04; dataset (n = 33), -4.67 to 4.04; dataset (n = 44), -4.67 to 4.52.

Table 3.
Statistical performance measures (MSLL, correlation coefficient and ION) used to determine the performance of each method for the range of tests evaluated in this study. Note: for each optimization method or test the best performing models are shown in bold text, and those with the worst performance are shown in underline. Note: [1]Taken from [29]; [2]Taken from [30].

Table 4.
Statistical performance of the test data set [29, 30] and various subsets based on altering the range and size of data in each subset.

Table 1.

Dataset	Number of data points	Number of descriptors used	Descriptors used	Target	Reference
Human A	21	5	log P, MW, HA, HD, SP	log k_p	[7, 30]
Human B	84	5	log P, MW, HA, HD, SP	log k_p	[7, 30]
Rat	26	5	log P, MW, HA, HD, SP	log k_p	[7, 30]
Mouse	46	5	log P, MW, HA, HD, SP	log k_p	[7, 30]
Pig	14	5	log P, MW, HA, HD, SP	log k_p	[7, 30]
Magnusson Set A (t)	85	6	log P, MPt, MW, HA, HD, T_{exp}	J_{max}	[29]
Magnusson Set B (V_s)	50	6	log P, MPt, MW, HA, HD, T_{exp}	J_{max}	[29]
Magnusson Set C (V_p)	27	6	log P, MPt, MW, HA, HD, T_{exp}	J_{max}	[29]
Magnusson Set D (V_f)	45	6	log P, MPt, MW, HA, HD, T_{exp}	J_{max}	[29]

Where log P is the octanol-water partition coefficient; HA and HD represent the number of hydrogen bond acceptor and donor groups on a molecule, respectively; MW is the molecular weight; SP is the Fedor's solubility parameter; MPt is the melting point; T_{exp} is the experimental temperature. For the Magnusson datasets [29] the text in brackets at the end of each dataset is the notation used in the original paper, e.g. Magnusson Set A (t) is the dataset listed as 't' in the original study.

Table 2.

Chemical Number (from [29])	Chemical name	Experimental temperature (K)	MW	log K_{ow} (log P)	MPt (K)	HD	HA	log J_{max}	Dataset maintaining the range of MW values but reducing, from subset 1 to subset 4, the size of the dataset.				Dataset maintaining the range of logP values but reducing, from subset 1 to subset 4, the size of the dataset.				References (source of data)
									Included in data subset 1	Included in data subset 2	Included in data subset 3	Included in data subset 4	Included in data subset 1	Included in data subset 2	Included in data subset 3	Included in data subset 4	
1	Water	303	18	-1.38	273	2	1	-4.19	✓	✓	✓	✓			✓		[37]
2	Water	305	18	-1.38	273	2	1	-4.07	✓								[41]
3	Water	298	18	-1.38	273	2	1	-4.56	✓								[49]
4	Methanol	298	32	-0.72	175	1	1	-4.81	✓	✓			✓		✓		[49]
5	Methanol	303	32	-0.72	175	1	1	-4.3	✓					✓			[52]
6	Ethanol	298	46	-0.19	159	1	1	-4.87	✓	✓	✓		✓			✓	[49]
7	Propanol	303	60	0.34	147	1	1	-4.65	✓						✓		[38]
8	Propanol	298	60	0.34	147	1	1	-4.8	✓	✓							[49]
9	Urea	310	60.1	-2.11	406	4	3	-5.87					✓	✓	✓		[37]
10	Urea	300	60.1	-2.11	406	4	3	-5.76					✓				[46]
11	Urea	312	60.1	-2.11	406	4	3	-5.6		✓	✓	✓	✓				[46]
12	2-Butanone	303	72.1	0.37	187	0	1	-4.86						✓			[39]
13	Ethyl ether	303	74.1	0.98	157	0	1	-4.88					✓		✓	✓	[39]
14	Butanol	303	74.1	0.88	184	1	1	-5.59									[38]
15	Butanol	298	74.1	0.88	184	1	1	-5.67		✓	✓						[49]
16	Benzene	304	78.1	2.22	279	0	0	-5.61	✓	✓			✓	✓	✓		[38]
17	Pentanol	298	88.2	1.41	194	1	1	-5.82	✓				✓				[49]
18	2-Ethoxy ethanol	303	90.1	-0.27	183	1	2	-5.58	✓				✓				[39]
19	2,3-Butanediol	303	90.1	-0.99	298	2	2	-6.25	✓					✓		✓	[39]
20	Toluene	310	92.1	2.68	178	0	0	-5.32	✓	✓							[35]
21	Phenol	310	94.1	1.48	314	1	1	-4.77		✓	✓	✓			✓		[51]
22	Phenol	295	94.1	1.48	314	1	1	-6.88									[52]
23	Phenol	298	94.1	1.48	314	1	1	-5.17						✓			[47]

24	Hexanol	298	102.2	1.94	228	1	1	-6.13		✓								[49]
25	p-Cresol	298	108.1	1.94	309	1	1	-5.47	✓	✓	✓			✓				[48]
26	Benzyl alcohol	298	108.1	1.04	258	1	1	-5.62	✓				✓					[47]
27	o-Phenylenediamine	305	108.1	0.05	377	4	2	-6.74	✓				✓	✓				[40]
28	p-Cresol	298	108.1	1.94	285	1	1	-5.45	✓									[48]
29	p-Cresol	310	108.1	1.94	309	1	1	-4.62	✓						✓			[36]
30	p-Cresol	298	108.1	1.94	303	1	1	-5.44	✓									[48]
31	p-Phenylenediamine	305	108.1	-0.85	419	4	2	-7.09	✓									[40]
32	Resorcinol	298	110.1	0.76	384	2	2	-5.81		✓	✓	✓						[48]
33	Heptanol	303	116.2	2.47	238	1	1	-6.27		✓								[39]
34	Heptanol	298	116.2	2.47	238	1	1	-6.34					✓	✓	✓			[49]
35	Benzoic acid	308	122.1	1.9	395	1	2	-5.9					✓	✓				[45]
36	p-Ethylphenol	298	122.2	2.47	318	1	1	-5.85					✓					[48]
37	3,4-Xylenol	298	122.2	2.4	334	1	1	-5.83					✓	✓				[48]
38	2-Phenylethanol	298	122.2	1.36	259	1	1	-5.86		✓			✓					[47]
39	4-Hydroxybenzyl alcohol	310	124.1	0.3	393	2	2	-6.97		✓	✓		✓					[35]
40	p-Chlorophenol	298	128.6	2.43	317	1	1	-5.17					✓					[48]
41	o-Chlorophenol	298	128.6	2.04	282	1	1	-5.25									✓	[48]
42	5-Fluorouracil (+ - + -)	305	130.1	-0.78	556	2	4	-8.57	✓									[53]
43	Octanol	303	130.2	3	258	1	1	-6.6	✓				✓	✓				[52]
44	Octanol	298	130.2	3	258	1	1	-6.67	✓	✓	✓	✓	✓					[49]
45	Nicotinate, methyl	310	137.1	0.88	316	0	3	-5.97		✓				✓				[42]
46	m-Nitrophenol	298	139.1	1.93	370	1	4	-6.28					✓	✓				[48]
47	p-Nitrophenol	298	139.1	1.57	387	1	4	-6.25									✓	[48]
48	Chlorocresol	298	142.6	2.89	340	1	1	-5.72						✓				[48]
49	Nonanol	298	144	3.53	268	1	1	-7.23										[49]
50	beta-Naphthol	298	144.2	2.71	396	1	1	-6.71										[48]
51	Thymol	298	150.2	3.28	325	1	1	-6.45					✓					[48]
52	Nicotinate, ethyl	310	151.2	1.41	282	0	3	-5.65					✓					[42]
53	alfa-(4-Hydroxyphenyl)	310	151.2	-0.29	450	3	3	-7.37					✓	✓				[36]

	acetamide																
54	Methyl-4-hydroxy benzoate	298	152.1	1.87	401	1	3	-6.92		✓	✓		✓				[48]
55	Chloroxylenol	298	156.6	3.35	389	1	1	-6.95		✓							[48]
56	Decanol	298	158.3	4.06	279	1	1	-7.73					✓				[49]
57	2,4-Dichlorophenol	298	163	3	318	1	1	-5.73	✓				✓		✓		[48]
58	p-Bromophenol	298	173	2.49	337	1	1	-5.5	✓	✓				✓			[48]
59	Mannitol	303	182.2	-4.67	440	6	6	-6.93	✓	✓	✓	✓	✓				[43]
60	Mannitol	312	182.2	-4.67	440	6	6	-7.05	✓				✓	✓			[46]
61	Mannitol	300	182.2	-4.67	440	6	6	-7.26	✓				✓	✓	✓	✓	[46]
62	2,4,6-Trichlorophenol	298	197.5	3.58	342	1	1	-6.57	✓	✓				✓		✓	[48]
63	Estrone	299	270.4	3.69	528	1	2	-10.76		✓	✓			✓	✓		[49]
64	beta-Estradiol	310	272.4	4.13	449	2	2	-9.89					✓				[44]
65	beta-Estradiol	305	272.4	4.13	449	2	2	-10.2					✓	✓			[54]
66	beta-Estradiol	299	272.4	4.13	449	2	2	-11.88					✓				[49]
67	Estriol	299	288.4	2.94	555	3	3	-11.23	✓				✓	✓		✓	[49]
68	Testosterone	298	288.4	3.48	428	1	2	-10.16	✓								[50]
69	Testosterone	299	288.4	3.48	428	1	2	-10.46	✓	✓					✓		[49]
70	Progesterone	299	314.5	4.04	394	0	2	-10.37	✓	✓	✓	✓	✓	✓			[49]
71	Pregnenolone	299	316.5	4.52	466	1	2	-10.09	✓				✓	✓	✓	✓	[49]
72	Cortexone	299	330.5	3.41	415	1	3	-9.87	✓					✓			[49]
73	17-alfa-Hydroxyprogesterone	299	330.5	2.89	496	1	3	-10.77	✓	✓							[49]
74	Sucrose	310	342.3	-3.85	459	8	11	-7.24		✓	✓		✓	✓			[35]
75	Corticosterone	298	346.5	1.76	454	2	4	-10.89					✓		✓		[50]
76	Corticosterone	299	346.5	1.76	454	2	4	-10.54					✓	✓			[49]
77	Corticosterone	312	346.5	1.76	454	2	4	-8.83					✓				[46]
78	Corticosterone	300	346.5	1.76	454	2	4	-9.51	✓	✓			✓				[46]

79	Prednisolone	298	360.4	1.69	514	3	5	-10.56	✓	✓	✓	✓		✓			[50]
80	Cortisone	299	360.5	1.24	495	2	5	-11.19	✓				✓	✓			[49]
81	Hydrocortisone (HC)	298	362.5	1.43	493	3	5	-11.6	✓								[49]
82	Hydrocortisone (HC)	299	362.5	1.43	493	3	5	-11.64	✓	✓				✓	✓		[49]
83	Triamcinolone	298	394.5	1.03	543	4	6	-12.09	✓	✓	✓		✓	✓			[50]
84	Triamcinolone acetonide	298	434.5	2.6	566	2	6	-12.01	✓	✓				✓	✓		[50]
85	Betamethasone-17-valerate	298	476.6	3.98	457	2	6	-10.65	✓	✓	✓	✓	✓				[50]

Note: Where log P is the octanol-water partition coefficient, represented by log K_{ow} in the original paper [29]; HA and HD represent the number of hydrogen bond acceptor and donor groups on a molecule, respectively; MW is the molecular weight; MPt is the melting point; T_{exp} is the experimental temperature.

Where the lipophilicity (log K_{ow}, or log P) range is maximised, the range is from -4.67 to 4.52 for all four datasets of different sizes, and MW ranges are: 46 to 316.5 (for dataset where n=9); 18 to 434.5 (n=17); 32 to 434.5 (n=33); 32 to 476.6 (n=44). Where the MW range is maximised, MW range is from 18 to 476.6 for datasets of all sizes and logKow ranges are: -4.6 to 4.04 (for dataset where n=9); -4.67 to 4.04 (n=17); -4.67 to 4.04 (n=33); -4.67 to 4.52 (n=44).

Table 3.

Dataset	Grid search	Random search	Conjugate Gradient	Hyper-prior (Gaussian)	Hyper-prior (Laplace)	Hyper-prior (Smoothbox)	Evolutionary algorithm	QSPR (correlation (r) only)
				Correlation coefficient, r				
Magnusson Set A[1]	0.96 ± 0.01	0.96 ± 0.01	**0.97 ± 0.01**	0.96 ± 0.02	0.96 ± 0.02	**0.97 ± 0.02**	**0.97 ± 0.02**	0.10±0.14
Human B[2]	0.59 ± 0.15	0.60 ± 0.15	0.60 ± 0.14	**0.64 ± 0.11**	**0.64 ± 0.11**	0.63 ± 0.11	**0.64 ± 0.11**	0.08±0.20
Magnusson Set B[1]	0.93 ± 0.05	0.90 ± 0.05	0.94 ± 0.05	0.85 ± 0.11	0.83 ± 0.12	**0.96 ± 0.03**	**0.95 ± 0.03**	0.38±0.16
Mouse[2]	0.52 ± 0.40	0.52 ± 0.40	0.50 ± 0.44	0.51 ± 0.39	**0.53 ± 0.37**	0.51 ± 0.35	0.50 ± 0.39	-0.38±0.37
Magnusson Set D[1]	0.59 ± 0.31	0.56 ± 0.31	0.59 ± 0.31	0.60 ± 0.25	**0.62 ± 0.28**	0.55 ± 0.27	0.54 ± 0.25	-0.18±0.48
Magnusson Set C[1]	**0.83 ± 0.13**	**0.83 ± 0.13**	0.81 ± 0.11	0.63 ± 0.24	0.65 ± 0.23	0.80 ± 0.15	0.75 ± 0.03	-0.77±0.23
Rat[2]	0.15 ± 0.72	0.18 ± 0.71	0.19 ± 0.68	0.53 ± 0.56	**0.56 ± 0.49**	**0.56 ± 0.49**	0.08 ± 0.81	0.30±0.64
Human A[2]	0.74 ± 0.17	0.74 ± 0.17	0.73 ± 0.19	**0.77 ± 0.15**	**0.77 ± 0.17**	**0.77 ± 0.16**	0.70 ± 0.16	0.37±0.45
Pig[2]	0.84 ± 0.01	**0.92 ± 0.18**	0.87 ± 0.01	0.85 ± 0.01	0.85 ± 0.01	0.87 ± 0.01	0.65 ± 0.36	0.29±0.31
				MSLL				
Magnusson Set A[1]	-1.33 ± 0.21	-1.32 ± 0.02	**-1.35 ± 0.14**	-0.97 ± 0.06	-0.99 ± 0.04	**-1.35 ± 0.10**	**-1.10 ± 0.02**	-
Human B[2]	-0.22 ± 0.35	-0.15 ± 0.07	1.17 ± 2.90	-0.16 ± 0.07	-0.15 ± 0.07	**-0.27 ± 0.10**	**-0.20 ± 0.01**	-
Magnusson Set B[1]	-0.95 ± 0.28	-0.98 ± 0.02	-0.98 ± 0.21	-0.56 ± 0.14	-0.62 ± 0.11	**-0.99 ± 0.18**	-0.80 ± 0.06	-

Mouse[2]	0.07 ± 0.56	0.74 ± 0.48	0.72 ± 0.86	-0.02 ± 0.07	-0.06 ± 0.11	**-0.13 ± 0.28**	-0.10 ± 0.01	-
Magnusson Set D[1]	-0.22 ± 0.22	-0.18 ± 0.02	-0.18 ± 0.19	-0.12 ± 0.12	**-0.23 ± 0.21**	-0.15 ± 0.15	-0.10 ± 0.01	-
Magnusson Set C[1]	-0.20 ± 0.80	-0.17 ± 0.17	-0.20 ± 0.82	-0.15 ± 0.06	-0.12 ± 0.43	**-0.40 ± 0.32**	-0.30 ± 0.04	-
Rat[2]	-0.04 ± 0.76	-0.10 ± 0.14	-0.31 ± 0.30	-0.11 ± 0.07	-0.37 ± 0.29	**-0.43 ± 0.29**	0.16 ± 0.15	-
Human A[2]	-0.22 ± 0.27	-0.14 ± 0.10	-0.23 ± 0.26	-0.13 ± 0.15	**-0.32 ± 0.27**	-0.16 ± 0.14	-0.10 ± 0.06	-
Pig[2]	-0.98 ± 0.37	**-1.01 ± 0.09**	-0.90 ± 0.36	-0.50 ± 0.15	-0.93 ± 0.43	-0.72 ± 0.31	-0.00 ± 0.42	-

ION

Magnusson Set A[1]	0.91 ± 0.02	0.91 ± 0.00	**0.93 ± 0.02**	0.89 ± 0.03	0.91 ± 0.02	**0.93 ± 0.02**	**0.93 ± 0.00**	-
Human B[2]	0.34 ± 0.21	0.32 ± 0.01	0.36 ± 0.17	**0.41 ± 0.13**	**0.41 ± 0.14**	**0.41 ± 0.16**	**0.41 ± 0.01**	-
Magnusson Set B[1]	0.82 ± 0.08	0.77 ± 0.02	0.84 ± 0.08	0.67 ± 0.17	0.69 ± 0.18	**0.85 ± 0.07**	0.82 ± 0.02	-
Mouse[2]	0.28 ± 0.31	0.27 ± 0.06	0.24 ± 0.38	0.29 ± 0.29	**0.32 ± 0.27**	0.28 ± 0.32	0.23 ± 0.01	-
Magnusson Set D[1]	0.24 ± 0.27	0.20 ± 0.03	0.24 ± 0.28	0.21 ± 0.14	**0.30 ± 0.22**	0.22 ± 0.18	0.23 ± 0.02	-
Magnusson Set C[1]	**0.55 ± 0.23**	**0.55 ± 0.01**	0.47 ± 0.22	0.30 ± 0.17	0.42 ± 0.20	0.47 ± 0.21	0.39 ± 0.02	-
Rat[2]	0.10 ± 0.58	0.08 ± 0.04	0.24 ± 0.25	0.31 ± 0.21	0.29 ± 0.34	**0.40 ± 0.20**	0.00 ± 0.22	-
Human A[2]	0.27 ± 0.30	0.24 ± 0.09	0.27 ± 0.27	0.30 ± 0.14	**0.38 ± 0.24**	0.29 ± 0.13	0.14 ± 0.05	-
Pig[2]	0.77 ± 0.18	0.81 ± 0.06	0.82 ± 0.13	0.65 ± 0.16	**0.82 ± 0.14**	0.80 ± 0.13	0.45 ± 0.11	-

1. Magnusson et al., 2004
2. Moss and Cronin, 2002

Table 4.

Performance / subsets	Original dataset and subsets which maintain a full range of molecular weight, based on the original dataset[1]					Original dataset and subsets which maintain a full range of log P, based on the original dataset[1]					Original dataset and subsets[2] in which the range of molecular weight is systematically reduced, based on the original dataset[1]					Original dataset and subsets[2] in which the range of log P is systematically reduced, based on the original dataset[1]				
	Magnusson	Subset 1	Subset 2	Subset 3	Subset 4	Magnusson	Subset 1	Subset 2	Subset 3	Subset 4	Magnusson	Subset 1	Subset 2	Subset 3	Subset 4	Magnusson	Subset 1	Subset 2	Subset 3	Subset 4
Size of dataset	85	44	33	17	9	85	44	33	17	9	85	40	40	40	40	85	40	40	40	40
ION (Smoothbox hyper-prior)	0.93	0.92	0.91	0.88	0.90	0.93	0.90	0.93	0.94	0.72	0.93	0.11 ± 0.05	0.68 ± 0.29	0.81 ± 0.10	0.91 ± 0.03	0.93	0.92 ± 0.03	0.90 ± 0.02	0.89 ± 0.07	0.91 ± 0.04
ION (conjugate gradient)	0.93	0.89	0.90	0.87	0.88	0.93	0.89	0.92	0.94	0.80	0.93	0.10 ± 0.05	0.66 ± 0.30	0.80 ± 0.10	0.90 ± 0.03	0.91	0.90 ± 0.03	0.90 ± 0.02	0.88 ± 0.07	0.90 ± 0.04
MSLL (Smoothbox hyper-prior)	-1.35	-1.20	-1.06	-0.88	-0.99	-1.35	-1.04	-1.1	-1.02	-0.98	-1.35	-1.18 ± 0.64	-2.09 ± 0.73	-1.26 ± 0.19	-1.05 ± 0.27	-1.35	-1.10 ± 0.43	-1.07 ± 0.15	-1.02 ± 0.28	-1.15 ± 0.23
MSLL (conjugate gradient)	-1.35	-1.08	-1.06	-0.86	-1.06	-1.35	-1.04	-1.1	-1.02	-1.23	-1.35	-1.76 ± 0.69	-2.10 ± 0.72	-1.22 ± 0.23	-1.00 ± 0.28	-1.35	-0.66 ± 1.76	-0.90 ± 0.62	-0.94 ± 0.47	-1.10 ± 0.24
Correlation coefficient (Smoothbox hyper-prior)	0.97	0.97	0.93	0.83	0.97	0.97	0.96	0.97	0.85	0.88	0.97	0.34 ± 0.22	0.73 ± 0.09	0.89 ± 0.05	0.95 ± 0.01	0.97	0.95 ± 0.02	0.95 ± 0.01	0.94 ± 0.04	0.95 ± 0.03
Correlation coefficient (conjugate gradient)	0.97	0.97	0.93	0.83	0.89	0.97	0.96	0.97	0.85	0.90	0.97	0.32 ± 0.30	0.73 ± 0.11	0.88 ± 0.05	0.94 ± 0.01	0.97	0.95 ± 0.02	0.94 ± 0.01	0.94 ± 0.05	0.95 ± 0.03
Correlation coefficient (aqueous solubility)[2]	0.56	0.60	0.60	0.49	0.27	0.56	0.50	0.66	0.68	0.41	0.56	0.59 ± 0.16	0.59 ± 0.16	0.59 ± 0.16	0.59 ± 0.16	0.56	0.59 ± 0.16	0.59 ± 0.16	0.59 ± 0.16	0.59 ± 0.16
Correlation coefficient (aqueous solubility, adjusted to temperature)[2]	0.55	0.59	0.59	0.47	0.24	0.55	0.48	0.64	0.67	0.38	0.55	0.58 ± 016	0.58 ± 0.16	0.58 ± 0.16	0.58 ± 0.16	0.55	0.58 ± 0.16	0.58 ± 0.16	0.58 ± 0.16	0.58 ± 0.16

1. From [29]
2. The range of values reduces from Subset 4 to Subset 1

SUPPLEMENTAL MATERIAL – DETAILED METHODS AND THEORETICAL BACKGROUND.

Model fitting for small skin permeability datasets: hyperparameter optimisation in Gaussian Processes regression

Ashrafi P[1], Sun Y[1], Davey N[1], Adams RG[1], Wilkinson SC[2], Moss GP[3]*

[1]School of Computer Science, University of Hertfordshire, Hatfield, UK;
[2]Medical Toxicology Centre, Wolfson Unit, Medical School, University of Newcastle-upon-Tyne, UK;
[3]The School of Pharmacy, Keele University, Keele, UK;

*Corresponding author: g.p.j.moss@keele.ac.uk
+44(0)1782 734 776
The School of Pharmacy
Keele University
Keele, Staffordshire, UK
ST5 5BG

1. PREAMBLE

The majority of this material has been previously published in a number of our publications. It is collated here for both convenience and clarity. The full MatLab code for the GP method has been published previously [1].

2. THEORETICAL BACKGROUND

2.1 Gaussian Process Regression (GPR)

Gaussian Process Regression (GPR) is a technique of increasing importance in the Machine Learning field, and which is finding greater utility in the physical and biological sciences [2 – 6]. A GPR performs a non-linear regression optimised from the training data that consists of a number of input vectors (descriptor or features)–with a corresponding target value. The input vectors are denoted as $\boldsymbol{X}$, which includes $\boldsymbol{N}$ input vectors $\boldsymbol{x_i}$ (i= 1, …, N). The corresponding output values would be denoted by **y** and the new data point which we want to make the prediction of, $\mathbf{y}_*$, will be denoted as $\mathbf{x}_*$. Generally this is achieved by obtaining the weighted average of the **y**-values in the training set, with the weighting being the similarity of $\mathbf{x}_*$ to the vectors in the training set $\boldsymbol{X}$, e.g. the similarity of measured molecular weight and the number of hydrogen bonds of a chemical ($\mathbf{x}_*$) with the molecular weight and the number of hydrogen bonds of a chemical ($\boldsymbol{x_i}$) in the training set. Thus, the greater the similarity between these two chemicals gives a greater weighting. In a GPR, similarity is then measured using a covariance, or kernel, function; this is a function that takes two inputs and produces a single real value as its output. The prediction $\mathbf{y}_*$ is given by:

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

where $\boldsymbol{k}_*$is the vector of covariances between the test point and all the **x**-values in the training set. The term $(K + \sigma_n^2 I)^{-1}$, which is completely independent of a new data point, normalises the similarity vector $\boldsymbol{k}_*$ (just as with the weighted average, it is necessary to divide by the sum of the weights). The normalised weights are multiplied by **y**-values and the prediction achieved. Further details can be found elsewhere [7].

A variety of kernels, or covariance functions, can be used in GPR models. In our initial studies, the Matérn, Polynomial and Gaussian covariance functions have been applied to the data. However, as

the Matérn covariance function resulted in a better performance it is used as the main kernel function in this study [8, 9].

The Matérn covariance function has a positive parameter of ν. This function becomes especially simple when ν is a half-integer: $\nu = \rho + 1/2$, where ρ is a non-negative integer. The Matérn covariance function can be defined as a product of an exponential and a polynomial of order ρ. The most interesting cases for Machine Learning are $\nu = 3/2$ and $\nu = 5/2$ and they are defined [9] as:

$$K_{\nu=3/2}(r) = \left(1 + \frac{\sqrt{3}r}{l}\right)\exp(-\frac{\sqrt{3}r}{l}) \quad (2)$$

and

$$K_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right)\exp(-\frac{\sqrt{5}r}{l}) \quad (3)$$

where $r = |\boldsymbol{x}_i - \boldsymbol{x}_j|$ and l (length scale) is a positive hyperparameter. Further information on these functions can be found elsewhere. As mentioned earlier, the measured features may be noisy and this is modelled by multiplying the kernel by a signal variance parameter σ_f^2. The observed **y** values may also be noisy and this is parameterised by the noise variance coefficient, σ_n^2 that appears in Equation (1).

2.2 The effect of hyperparameters on model performance

The length scale (l), signal variance, σ_f^2, and noise variance, σ_n^2, are called the hyperparameters (θ) of the model as they are parameters of a prior distribution. The length-scale, l, defines how fast a function sampled from the GP oscillates. For example, when the length scale is large, this represents a slow change and the curve is smooth; when the length scale is small, this represents a fast change and curve oscillates heavily. If the length scale is too small then the kernel value in, for example, a training set would be very small with a weighing close to zero, meaning that the test data will infer, or 'learn', nothing from the training data. Optimised values can be inferred from the data using either the marginal likelihood maximisation or methods of cross-validation and marginal likelihood can therefore be used as an appropriate cost function. More specifically, minimisation of the negative log marginal likelihood, $\mathcal{L}$, with respect to hyperparameters ($\boldsymbol{\theta}$) of the covariance results in the following expression [10]:

$$\mathcal{L} = -\log p(\boldsymbol{y}|\boldsymbol{\theta}) \quad (4)$$

where

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) = -\frac{1}{2}\log|C(\boldsymbol{\theta})| - \frac{1}{2}\boldsymbol{y}^{\boldsymbol{T}}C(\boldsymbol{\theta})^{-1}\boldsymbol{y} - \frac{N}{2}\log(2\pi)$$

where N is the number of data, $\boldsymbol{y}$ are the target values and $C = K_N + \sigma_n^2 I$ [7].

It is possible that inferring the hyperparameters from the data could be particularly problematic with small datasets. To resolve this, various optimisation methods have been used to obtain the hyperparameters that minimise negative log marginal likelihood values. The methods used include the Conjugate Gradient, Grid Search, Random Search, Hyper-Prior and Evolutionary Algorithm methods.

Conjugate Gradient method

Essentially, the aim of any such analysis is to find the minimum of a cost function. The first-order method to achieving this is simply to follow the maximum gradient downwards. Usually, however, second-order methods are used to achieve this outcome. These include the Conjugate Gradient (CG) method, which is an iterative method normally used to solve a linear equation [11].

Grid Search

This method requires that the search should be conducted over a range of the parameters or hyperparameters used in a particular study. For a single parameter, a likely range of values is chosen, and the model is evaluated from values in this range. If there is more than one hyperparameter, then the search will take place in the Cartesian product defined by the range of each particular parameter (over a hyper-grid). For example, if twenty equal steps are chosen in each parameter range and there are three parameters in the model, then the total number of different parameter combinations obtained will be 8,000 (i.e. 20 x 20 x 20).

Random Search

In this method the parameters are selected at random. When granting Random Search methods the same computational budget, this approach has found better models by efficiently searching a larger but less promising configuration space [12] although this is not a major concern in a study of this nature.

Hyper-prior methods

Hierarchical model specification is commonly used to gain a joint regularisation for individual models. The first level includes parameters that may be those used in linear or non-linear models; for example, the parameters of a simple linear regression model, such as weights. At the second level of the model hyperparameters, θ, control the distribution of the first-level parameters. Finally, at the top-level models may feature a discrete set of possible model structures called Hyper-Priors ($\mathcal{H}$), which characterise the prior distributions of the hyperparameters. The Prior 'over-models' ($\mathcal{H}$) is often taken to be flat, so that one particular model is not favoured over another [7]. The 'Prior Models' in this study are the Gaussian, Laplacian, and non-linear Smoothbox Prior methods. Univariate smoothed-box prior distributions are defined with quadratic decay in the log domain and it supports the whole real axis. The model is constructed by cutting a Gaussian in two parts and inserting a uniform distribution from the lower bound parameter (a) to the upper bound parameter (b), which are the parameters of the Smoothbox Prior. The Hyper-Prior method also balances the probability mass between the constituents of the model; this is described by η such that $\eta / (\eta + 1)$ represents the box and $1 / \ \eta + 1$ represents the Gaussian sides. Larger values of η tend to make the model more box-like, and Prior Smooth box distribution is given as:

$$\mathcal{H}(\theta) = \frac{1}{\mathcal{W} \cdot \left(\frac{1}{\eta+1}\right)} \cdot \begin{cases} N(\theta|a, s^2), & t \leq a \\ 1, & t \in [a, b] \\ N(\theta|b, s^2) & b \leq t \end{cases} \quad (5)$$

and:

$$\mathcal{W} = |b - a|, s = \frac{\mathcal{W}}{\eta\sqrt{2\pi}} \quad (6)$$

where a is the lower-bound parameter, b is the upper-bound parameter, $\eta>0$ is the slope parameter, t is zero by default (this can be chosen manually) and $\theta_{(1*N)}$ contains query hyperparameters for Prior evaluation [13]. The mean and variance parameters of the Gaussian and Laplacian priors are normally initialised based on the nature of the data and these values can be obtained using cross-validation in each of the data sets analysed in this study.

Evolutionary Algorithms

They are genetic population-based meta-heuristic techniques that aim to optimise the results in each 'generation' (iteration) of analysis. To evaluate the populations (e.g. possible values of hyperparameters) one of the Evolutionary Algorithm (EA) methods – reproduction, mutation, recombination or selection, or a mixture – can be applied. A 'fitness' function is then defined to determine the quality of the solution for each range of methods used. For example, the fitness function used in this study is the Negative Log Likelihood loss function (NLL), which should be minimised when compared to different combinations of hyperparameters. It is important to note that the fitness function results are static and only depends on the current information at a particular point. However, with the use of a heuristic crossover function the fitness function can be modified to vary dynamically based on current and previous results states. If this is the case the population in the first generation is initialised randomly, and the next generation's population can be obtained using the following heuristic crossover function:

$$child = parent2 + ratio \times (parent1 - parent2) \quad (7)$$

The obtained child is closer to the parent with a better fitness value. To specify the amount by which the child is far from each parent a ratio parameter is defined. This process continues until the required number of populations in the current generation is achieved.

Further, in each generation a number of children are obtained using mutation. In this case, this means that the variables (hyperparameters) are chosen randomly from the variable ranges from a Gaussian or uniform distribution. To keep the best results of the last generation a fixed number of best children from the last generation can be added to the next generation. Adding 'Elite' children guarantees that the performance of the model does not diminish over generations [14, 15].

3. EXPERIMENTAL SET-UP

3.1 Software

The Gaussian Process modelling methods for non-linear regression used previously were again adopted for this study [4, 7, 17, 17]. The latest version of the Hyper Prior optimisation Toolbox was also used [13]. The MatLab Genetic Algorithm (GA) optimisation toolbox was used to carry out the Evolutionary Algorithm hyperparameter optimisation.

3.2 Cross-validation

The importance of model validation in constructing computational models has been discussed previously [17]. In this study, we have validated models mainly using the cross-validation technique [8]. 5-fold cross-validation was performed. The datasets were shuffled and divided into 5 'folds'. Each time one of the folds was considered as the test set and the remaining four were considered as the training set. At this point, a validation set was removed from the training set. The hyperparameter optimisation methods were then applied to the training set and the prediction performances were gained for the validation set. This was then repeated for the other 3 possible validation sets. The best hyperparameters were chosen as those performed best over the four validation sets (the minimum average of negative log probability loss (NLL) values, which are defined in section 4). They were used to predict the permeability values of the test set.

3.3 Initialisation of experiments

The experiments were initialised as follows:

- Grid search: To conduct the manual search through the hyperparameter space, the hyperparameters were considered as a range $[10^{-3}, 10^{3}]$ with 20 equidistant steps. Using a 5-fold cross validation the model was trained with all the 8,000 (20 x 20 x 20) different sets of the hyperparameters and the predictions obtained for the test sets. On inspecting the prediction performances on the validation sets a finer search for better values of the hyperparameters was then performed with the search range limited to [0.01, 10] with 20 steps as no better results were obtained using the hyperparameters out of this range. The model was then trained with the new hyperparameters and tested on the test sets. The average values and their standard deviation among 5-folds were then reported.
- Random search: 20 values for each hyperparameter were obtained randomly within the same range [0.01, 10] considered in the grid search. Using 5-fold cross validation the model was then trained and the predictions obtained. Since, in each run of this experiment, the hyperparameters were selected randomly the experiment was repeated 5 times and the results were obtained by calculation of the mean and standard deviation of the experiment's results.
- Conjugate gradient: The hyperparameters were initialised to log(0.5) with the number of function evaluations set to 100.
- Hyper Prior methods: The mean and variance parameters of the Gaussian and Laplacian priors were set to constant values of 0.1 and 0.01, respectively and were obtained as the best prediction performances using cross-validation in each of the data sets. For the Smooth Box Prior method, a, b *and* η values are set to 10^{-3}, 10 and 2, respectively. Various values of η were evaluated and the value 2 was found to be the best value for the data sets used in this study.

- Evolutionary algorithm: Following an evaluation of ratios ranging from 0.1 to 1.2, the heuristic crossover function with a ratio of 0.7 was used to accelerate convergence as it was found to have the optimum performance for the data sets used. Each of the 50 generations has a population of 50 and the optimised hyperparameters were obtained from the last generation. The 'Elite' Children value was set to 4 and the mutation function was kept uniform, meaning that the children were randomly selected from a uniform distribution within the range of hyperparameters. The crossover fraction was set to 0.8 (0.8 * 50 = 40), meaning that the rest of the children in a population are 4 Elite children and 6 children were obtained from mutation. The population of the first generation was initialised randomly and was therefore similar to the Random Search. This experiment was repeated five times using the Genetic Algorithm Toolbox in MatLab.

3.4 Data set analysis

The different data sets used in this study were characterised in terms of their membership (data set size) and range (the range of physicochemical descriptors used). Data used are those published previously [18, 19] and are shown or described in Tables 1 and 2 (Main paper).

3.5 The effect of the size of the data set and the range of the physicochemical descriptor values on prediction performance

The effect of datasets sizes and molecular features ranges on the prediction performances was examined. Due to their ubiquitous use in this field, and their relevance as benchmarks in this study, the effects of molecular weight and lipophilicity (as log P or log $K_{o/w}$) were considered [20 - 22].

The first experiment considered how changes to the size (membership) of the data set affected the statistical quality of the resultant models whilst maintaining the range, or 'chemical space', of each model. The data set reported by Magnusson and co-workers [18] was used for this experiment. In separate experiments this data set was used to construct four smaller subsets that maintained the range of descriptors of the original data set (Table 2, main paper). To construct these data sets four subsets were chosen from the Magnusson data set – the data set sizes were 44, 33, 17 and 9. Chemicals were selected only to ensure that the maximum and minimum MW ranges were maintained across all the data sets. The GPR model described above was then trained with each data set with the hyper-prior Smoothbox and conjugate gradient optimisation methods used to set the best hyperparameters for the models. As a benchmark the QSAR reported previously [21] was used,

with a concentration correction to adjust between k_p and J_{max}, as the Potts and Guy QSAR model [20] did not perform well in the initial analysis. This experiment was repeated with subsets of the Magnusson data set which maintained the range of log P values across all data sets whilst reducing the data set membership. Subsets in both experiments were of the same size.

The final set of experiments involved creating four training sets of the Magnusson data set where the membership again was kept constant (at $n = 40$) to remove any effect associated with data set size. But, in these cases, the range of the physicochemical descriptor values examined (MW and log P) were systematically reduced by the generation of random subsets from the parent data set. We produced a fixed test set. One-fifth of the Magnusson (set A) was considered to be the test set and the training sets (including 5-fold cross-validation) were generated from the remaining data. The range of the first training set can be obtained by adding and subtracting the standard deviation of MW to and from the median of all MW values (excluding the values in the fixed test set). To keep the size of each training set the same (n = 40), members of the subset were picked at random from the given range. To obtain the next training sets, the standard deviation is added by larger values (for example, 40, 100 and 200, respectively), and the same process is repeated. The GPR model was then trained using the smoothbox hyper-prior and conjugate gradient methods, and the predicted log J_{max} values were reported for the same test set. As data was chosen randomly within each data range, the experiment was repeated ten times, with the mean and standard deviations being reported for both the GPR models and the QSAR benchmark. The same methods were used to analyse changes to both MW and log P.

4. PERFORMANCE MEASURES

The correlation coefficient (*r*), Negative Log Likelihood (NLL) and improvement over the naïve model (ION, where the naïve model always predicts the mean of the target value in the training set independently of the input), were used to determine the model performance [23].

If a predictive distribution is produced at each test input, $\boldsymbol{x}_*$, in the chosen dataset ($\mathcal{D}$), the NLL of the target under the model can be evaluated; if μ is considered as the mean prediction where the GPR produces a Gaussian predictive density, the NLL can be defined as:

$$NLL = -\log\, p(y_*|\mathcal{D}, x_*) = \frac{1}{2}\log(2\pi\sigma_*^2) + \frac{(y_*-\mu)^2}{2\sigma_*^2} \qquad (8)$$

where the predictive variance, σ_*^2 for GPR is determined to be $\sigma_*^2 = V(f_*) + \sigma_n^2$, where $V(\mathrm{f}_*)=k\left(\mathbf{x}_{*,}\,\mathbf{x}_*\right)\text{-}\mathbf{k}_*^{\mathrm{T}}(\mathrm{K}+\sigma_{\mathrm{n}}^2\mathrm{I})^{-1}\mathbf{k}_*$As the noisy target, y_*, is being produced, the noise variance, σ_n^2 must be added. This loss may be standardised by subtracting it from the obtained NLL using Equation (8) and considering the mean and variance of the training data; this is characterised as the standardised log loss (SLL) or, as reported previously, the MSLL [7].

By contrast, *ION* measures how much better a predictor is than the *naïve* predictor, and is given by:

$$ION = \frac{MSE_{Naive} - MSE_{GP}}{MSE_{Naive}}, \quad (9)$$

where *MSE* denotes the mean squared error.

The MSLL will be approximately zero for simple methods and negative for better methods. ION ranges from -∞ to 1, and greater positive ION values represent better performance. The correlation coefficient ranges from -1 to 1 and in this study a high positive value defines good prediction performance [7].

5. REFERENCES

1. 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.
2. Obrezanova O, Csanyi G, Gola JMR, Segall MD. Gaussian Processes: A method for automatic QSAR modelling of ADME properties. J. Chem. Info. Mod. 2007: 47, 1847-1857.
3. Schroeter T, Schwaighofer A, Mika S, Ter Laak A, Suelzle D, Ganzer U, Heinrich N, Mu1ller KR. Machine Learning Models for Lipophilicity and Their Domain of Applicability. Mol. Pharmaceutics, 2007, 4(4), 524-538.
4. Lam LT, Sun Y, Davey N, Adams RG, Prapopoulou M, Brown MB, Moss GP. The application of feature selection to the development of Gaussian process models for percutaneous absorption. J. Pharm. Pharmacol. 2010: 62, 738–749.
5. Mellor J, Grigoras I, Carbonell P, Faulon J-L. Semi-supervised Gaussian Process for automated enzyme search. J. Chem. Info. Mod. 2016: 5, 518-528.
6. Rahman M, Previs SF, Kasumov T, Sadygov RG. Gaussian Process modelling of protein turnover. J. Proteome Res. 2016: 15, 2115-2122.
7. Rasmussen CE, Williams KI. Gaussian Processes for Machine Learning. 2006, Boston, The MIT Press. [Available online at: http://www.gaussianprocess.org/gpml/chapters/RW.pdf; Accessed 10 April 2017].
8. Bishop CM. Neural Networks for Pattern Recognition. Clarendon Press, Oxford. 1995.

9. Sun Y, Adams R, Davey N, Moss GP, Prapopopolou M, Brown MB. The application of Gaussian processes in the predictions of permeability across mammalian and polydimethylsiloxane membranes. Art. Int. Res. 2012: 1, 86-98.
10. Snelson EL. Flexible and efficient Gaussian Process models for Machine Learning. University College London, PhD Thesis, 2007.
11. Shewchuk JR. An introduction to the conjugate gradient method without the agonizing pain. 1994 [Available online at: https://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf; Accessed 10 April 2017].
12. Bergstra J, Benigo Y. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 2012: 13, 281-305.
13. Rasmussen CE, Nickish H. The GPML Toolbox Version 3.5. [Available at: http://mlg.eng.cam.ac.uk/carl/gpml/doc/oldcode.html]; Accessed 10 April 2017].
14. Mitchell M. An introduction to genetic algorithms. 1998, Boston, The MIT Press.
15. Winter G, Periaux J, Galan M, Cuesta P. Genetic algorithms in engineering and computer science.
16. John Wiley & Sons, Inc., 1996, Chichester.
17. MacLaurin D, Duvenaud D, Adams RP. Gradient-based hyper-parameter optimization through reversible learning. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: JMLR: W&CP volume 37. [Available at: http://hips.seas.harvard.edu/files/maclaurin-hypergrad-icml-2015.pdf; Accessed 10 April 2017].
18. Tropsha A. Best Practices for QSAR Model Development, Validation and Exploitation. Mol. Informatics, 2010, 29(6-7), 476-488
19. Magnusson BM, Anissimov YG, Cross SE, Roberts MS. Molecular size as the main determinant of solute maximum flux across the skin. J. Invest. Dermatol. 2004: 122, 993-999.
20. Prapopoulou M. The development of a computation/ mathematical model to predict drug absorption across the skin. King's College London, PhD Thesis, 2012.
21. Potts RO, Guy RH. Predicting skin permeability. Pharm. Res. 1992: 9, 663-669.
22. Moss GP, Cronin MTD. Quantitative structure-permeability relationships for percutaneous absorption: re-analysis of steroid data. Int. J. Pharm. 2002: 238, 105-109.
23. Moss GP, Dearden JC, Patel H, Cronin MTD. Quantitative structure-permeability relationships (QSPRs) for percutaneous absorption. Tox. In Vitro. 2002: 16, 299-317.
24. Moss GP, Shah AJ, Adams RG, Davey N, Wilkinson SC, Pugh WJ, 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.
25. Moss GP, Sun Y, Wilkinson SC, Davey N, Adams R, Martin GP, Prapopoulou M, Brown MB. The application and limitations of mathematical models across mammalian skin and poldimethylsiloxane membranes. J. Pharm. Pharmacol. 2011: 63, 1411-1427.