PREDICTING THE ABSORPTION RATE OF CHEMICALS THROUGH MAMMALIAN SKIN USING MACHINE LEARNING ALGORITHMS

Submitted to the University of Hertfordshire
in partial fulfilment of the requirements of the degree of
DOCTOR OF PHILOSOPHY

SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE

Parivash Ashrafi
May 2016

Do not go where the path may lead, go instead where there is no path and leave a trail.

Ralph Waldo Emerson

Abstract

Machine learning (ML) methods have been applied to the analysis of a range of biological systems. This thesis evaluates the application of these methods to the problem domain of skin permeability. ML methods offer great potential in both predictive ability and their ability to provide mechanistic insight to, in this case, the phenomena of skin permeation. Historically, refining mathematical models used to predict percutaneous drug absorption has been thought of as a key factor in this field. Quantitative Structure-Activity Relationships (QSARs) models are used extensively for this purpose. However, advanced ML methods successfully outperform the traditional linear QSAR models. In this thesis, the application of ML methods to percutaneous absorption are investigated and evaluated.

The major approach used in this thesis is Gaussian process (GP) regression method. This research seeks to enhance the prediction performance by using local non-linear models obtained from applying clustering algorithms. In addition, to increase the model's quality, a kernel is generated based on both numerical chemical variables and categorical experimental descriptors. Monte Carlo algorithm is also employed to generate reliable models from variable data which is inevitable in biological experiments. The datasets used for this study are small and it may raise the over-fitting/under-fitting problem. In this research I attempt to find optimal values of skin permeability using GP optimisation algorithms within small datasets. Although these methods are applied here to the field of percutaneous absorption, it may be applied more broadly to any biological system.

Acknowledgement

I would like to express my special appreciation to my supervisors Dr Yi Sun, Dr Neil Davey, and Dr Gary P. Moss for being patient, inspirational, and enthusiastic in guiding my research.

I want to specially thank Yi, for encouraging my research and for allowing me to grow as a research scientist; Neil, for his endless support, enlightening advice, and tremendous kindness; and, Gary, for his continuous support of my Ph.D study, the innumerable motivation, and his scientific eminence.

I would like to thank my dear friend Dr Parimala Alva for all her help and support. I have also enjoyed interaction with all members of the Biocomputation research group during my time at the University of Hertfordshire.

It is also proper to thank Professor Rod G. Adams for his support at various occasions during my research. For financial support, I am thankful to the University of Hertfordshire.

Words can not express how grateful I am to my mother, father, brother, mother-in-law, and father-in-law for all their caring support. I could never have done this without their boundless love and encouragements. Last, but not the least, I am grateful to my loving, supportive, encouraging, and patient husband Mehdi, for his honest solidarity in this journey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

In recent years, there has been an increasing interest in the need for better predictions of percutaneous (through the skin) absorption within the pharmaceutical and cosmetic industries as well as in fields relating to toxicological issues, such as pesticide usage. The skin permeability measurements are especially important because skin is a permeable membrane which works as a controllable medium to pass the chemicals through the skin, blood stream, and the underlining tissue. Therefore, this research could be very interesting in biological systems. To carry out research on *in vivo* or *in vitro* experiments, the excised skin of the hairless animals (including rats, mice and pigs) or the skin of human cut through surgery could be used, although such studies have limitations due to the time taken and the expense in carrying them out (Lien and Gaot (1995)). Historically refining mathematical models used to predict percutaneous drug absorption have been thought of as a key factor in this field to avoid issues from ethical, cost effectiveness to suitability of treatment. For example, although they do not provide accurate predictions, Quantitative Structure-Activity Relationships (QSARs) models are used extensively for having understandable mathematical equations (Flynn (1990); El Tayar et al. (1991); Potts and Guy (1992); Moss et al. (2002); Cronin and Schultz (2003)).

Visualising the datasets reveals that relationship between absorption and the physicochemical features of the compound is non-linear (Moss et al. (2009, 2011); Sun et al.

(2012); Chen et al. (2007)) and the linear QSAR methods are not beneficial to predict the chemical's absorption rate through the skin. Consequently, computational models especially machine learning methods are further employed for this purpose. Studies carried out by Lim et al. (2002) and Chen et al. (2007) show that artificial neural network (ANN) technique outperforms the linear methods for predicting percutaneous absorption of the chemicals. Similarly, ensemble approach combining KNN and linear models performs better than simple linear methods to predict the permeability (Neumann et al. (2006)). Studies performed by Moss et al. (2009); Lam et al. (2010); Xue et al. (2004) illustrate that the other ML regression techniques such as Gaussian Process (GP) and Support Vector Machine (SVM) , gave much better results than the traditional linear QSAR models. Furthermore, the importance of optimisation techniques in ML methods such as SVM to be used for QSAR/QSPR purpose is discussed in Norinder (2003). Michielan and Moro (2010) review and compare the performance of various ML regression and classification techniques in QSAR strategies. These approaches include GP, Partial Least Squares (PLS), ANN, SVM, Decision Trees (DTs), and Random Forest (RF). Reviewing the results, they concluded that that SVM has demonstrated a good performance in a large number of regression and classification problems. In addition, studies by Shah et al. (2012) show that both GP and SVM can achieve comparable overall results and each of them may outperform the other one in the various chemical property space.

The major challenge of this research is the small size of the datasets. This is due to the difficulties in measuring the permeability of the compounds in the labs. Preparation and sampling work for a single chemical in the lab can take around three days and may need 6 to 24 repeated experiments. Therefore, the cost of experimentally generating only one estimation is as much as £25,000 - £30,000, hence the expansion of research in model development. Therefore, obtaining large datasets is very demanding and expensive.

The aim of this study is to deal with small datasets to predict the target values efficiently. Each dataset contains the structural chemicals features (numerical descriptors) and the experimental conditions conditions (nominal features). For each of the compounds, there is a measured corresponding target value. Therefore, this study seeks to examine various

computational models for predicting the percutaneous absorption rate of the chemicals using machine learning techniques, mainly the Gaussian Processes algorithm. Moreover, the prediction performances are enhanced by the optimisation methods applied to the GP.

1.2 Contribution

The novel contributions of this study includes:

- This study provides a review that compares the performance of linear and non linear regression methods to predict the permeability of the compounds. It illustrates that the Gaussian Process regression method can be a thorough replacement for QSAR/QSPR methods used in the pharmaceutical science domain. Moreover, local non-linear models can be built up by using clustering algorithms. This study provides a detailed investigation on the effect of size of the data, feature ranges and experimental conditions. It shows that temperature and diffusion cell type affect the model performance and they should be considered when the models are trained.

- One of the main objectives of this research is to deal with small size datasets. As they may increase the over-fitting and under-fitting problem, finding the solutions to cope with small datasets is highly investigated. Hyper-prior techniques for hyper-parameter optimisation in the GP, shows good improvement in the prediction performance compared to the previous models.

- As the datasets contain both numerical and categorical data, a new kernel function is implemented to be used in the GP. This kernel uses a mix of numerical and categorical data and shows to result in small increase in the model performance. However, in this study only 3 categorical (nominal) data are used and the performance may improve more if more categorical descriptors are recognised and employed.

1.3 Publications on this thesis

1.3.1 Journal paper

- **The application of machine learning to the modelling of percutaneous absorption: an overview and guide** by P Ashrafi, GP Moss, SC Wilkinson, N Davey, Y Sun - SAR and QSAR in Environmental Research, 2015 (see Appendix C.1)

1.3.2 Conference paper

- **The importance of hyperparameters selection within small datasets** by P Ashrafi, Y Sun, N Davey, R Adams, MB Brown, Maria Prapopoulou, Gary Moss- International Joint conference on Neural Networks (IJCNN), 2015 (see Appendix C.2)

1.3.3 Conference poster abstracts

- **The effect of quality and consistency of data on the development of predictive machine learning models for percutaneous absorption** by P. Ashrafi, Y Sun , N Davey, RG Adams, MB Brown, SC Wilkinson, GP Moss - presented in 14th conference on Perspectives in Percutaneous Penetration. France, April 2014 (see Appendix C.3.1)

- **Investigation of inconsistency in a skin permeability dataset using the Monte Carlo method** by P. Ashrafi, Y Sun , N Davey, RG Adams, SC Wilkinson, GP Moss - presented in the 15th conference on Perspectives in Percutaneous Penetration. France, April 2016 (see Appendix C.3.2)

- **The effect of experimental conditions on the development of quantitative models of skin permeation** by P. Ashrafi, Y Sun , N Davey, RG Adams, SC Wilkinson, GP Moss - presented in the 15th conference on Perspectives in Percutaneous Penetration. France, April 2016 (see Appendix C.3.3)

- **Assessment of chemical enhancers of transdermal drug delivery by support vector regression** By A Shah, P Ashrafi, Y Sun, RG Adams, N Davey, SC Wilkinson,

GP Moss - presented in the 15th conference on Perspectives in Percutaneous Penetration. France, April 2016 (see Appendix C.3.4)

1.4 Terminology abbreviations

The words and expressions used frequently in the skin data studies and in this thesis are shown as follows:

Terminology	Abbreviation
Quantitative Structure-Activity Relationships	**QSAR**
Quantitative Structure– Property Relationships	**QSPR**

The computational methods have been used in this study include:

Terminology	Abbreviation
machine learning	**ML**
negative log likelihood	**NLL**
correlation coefficient	**CorrCoef**
standardised log loss	**SLL**
mean standardised log loss	**MSLL**
improvement over naïve	**ION**
mean squared error	**MSE**
principal component analysis	**PCA**
Gaussian Processes	**GP**
Support Vector Machines/Regression	**SVM/SVR**
conjugate gradient	**CG**
evolutionary algorithms	**EA**
Single layer network	**SLN**
k-nearest neighbour	**KNN**
growing neural gas	**GNG**

1.5 The structure of this thesis

Chapter 2 reviews the literature about the skin layers, permeability routes and traditional methods have been used to estimate the permeability of the compounds through human and animal skins.

Chapter 3 shows the datasets have been used for the purpose of this study along with their experimental and feature details. All the machine learning and computational methods employed in this study are discussed in **Chapter 4**. These methods include GP definition and application along with its various covariance functions and the ways in which the hyper-parameters are optimised. In addition, it discusses the other regression methods such as SVR, SLN and KNN. Finally, the GNG clustering method is illustrated and the chapter finishes by all the performance measurement methods that are used for evaluation and comparison of the experiments results.

Chapter 5 shows the methods have been used to display the PCA plots. These plots are useful to visualise the numerical and categorical features of the data and see their relationships. The main experiments on the datasets using GP with various number of numerical, nominal data descriptors and covariance functions are discussed in **Chapter 6**. In this chapter the performance of the other regression methods are compared with the GP. Additionally, GNG clustering technique is applied to the all human data and the performance of the data in the new clusters are compared to the original datasets. At the end of this Chapter a solution to data inconsistency issue is also revealed.

Chapter 7, various hyper-parameter optimisation methods (in GP) are applied to the 11 datasets including both human and animal sets. Additionally, the features ranges, their size and the effect of these two factors on the model performance are also examined.

Chapter 8, reviews all the results and findings of this thesis in summary. Furthermore, in this chapter the contribution that is made by this thesis and future work are discussed with details.

Chapter 2

Skin Permeability and the Traditional QSAR/QSPR Approaches

This chapter will review the physiology and structure of the skin, including the different layers of the skin, the physicochemical properties required of a chemical for successful permeation of the skin, and the conditions (both *in vitro* experimental and *in vivo* environmental) which may affect the absorption rate of the chemical compounds through the skin. In addition, this chapter will discuss the historical QSAR approaches used to model the percutaneous penetration of the skin.

2.1 Introduction

The analysis of percutaneous (across the skin) absorption of exogenous (external) chemicals, in fields as diverse as pharmaceuticals, cosmetics, pesticides and the bulk handling of industrial chemicals, has become increasingly important over the last 20-25 years. It started from Flynn (1990), El Tayar et al. (1991) and Potts and Guy (1992), although in a pharmaceutical / cosmetic / toxicological sense this work started by Moss and Cronin (2002). Quantitative Structure-Activity Relationships (QSARs) or Quantitative Structure–Property Relationships (QSPR) are widely used to relate the physicochemical properties of a penetrant to its skin permeation (Cronin and Schultz (2003)). The QSAR models in skin and other biological or environmental systems are popular because they can both

yield predictions for new chemicals and describe the mechanism of action (in this case, the mechanism of skin permeability, which is based on the important physical properties of a molecule which influence the permeability process). So, any such model has to offer both these advantages if it is to be used by physical and biological scientists with little or no experience in model development or use. That is why the improved quality of machine learning methods, which don't have an equation, is so important and why the Feature Selection is also important as it can help define issues of mechanistic relevance. In most of QSAR/ QSPR models, multiple linear regression is the method used in most publication. However, visualising the data in a large number of cases shows non-linear relationships between the data and the target to be predicted. In addition, the current approaches are limited by the nature of the models chosen and the nature of the dataset. The important point to be noted is that the collection of reliable data is difficult due to the reason that the skin can be quite variable, affected by the site where it has come from and the individual from whom it is taken. Therefore, the accuracy of the model is related to the variability of the data and variability of data depends on the biological variation, variation in methods used and in the quality of the work carried out (Chilcott et al. (2005)). It is important to emphasise that variability can not be entirely removed from a biological system, so one should consider this case to generate good models.

There are a wide range of experimental protocols for assessing the skin permeability. In order to measure the percutaneous absorption and being able to find relationship between methods, we should consider assessing the relative permeability in the same circumstances such as the same sites of the skin and the temperature in area at which the studies are performed. If the experiments performed properly, there should not be any effect on the results. In addition to the biological variation, any change in the conditions (buffers, solvents, etc.), will change the results significantly and produce different outcomes. Knowledge of this will have to be considered in the models. From this, one can see that while there are a small number of protocols (flow-though and static cells) but there are a large number of ways of changing experiments (temperature, etc) and it may also affect the results remarkably.

It is mentioned previously, and this is important to emphasises that measuring the permeability coefficient of a single chemical can take around three days of preparatory and

sampling work and entail some 6 to 24 repeated experiments (usually a minimum of six repeats of the same study but, depending on the purpose of the study- research, regulatory product submission, up to twenty-four repeats may be required and so on). Therefore, the cost of experimentally generating only one estimation is as much as £25,000 - £30,000. Hence the expansion of research in model development is needed.

2.2 Skin histological layers

The largest organ of our body is skin which covers any area of about two square metres. Skin protects the body organs from external harmful molecules. It also controls the temperature of the body with the sweating system. The pain can be transferred from the skin to the nerves via a range of specialist receptors, such as nociceptors. Skin consist of three main layers called the epidermis, the dermis and the subcutis layers from outside to inside. A diagram (Figure 2.1) of the human skin from skin care forum shows these layers with more details (© KLEINHANS RED, Source: www.skin-care-forum.basf.com).

The first, outermost, layer is the epidermis. It consists of 5 layers: the *stratum basale*, the stratum *spinosum* (prickle cell layer), the *stratum granulosum* (granular layer), the *stratum lucidum* and the *stratum corneum* (horny layer) from inside to outside. The thickness of the epidermis varies in different types of skin. it may be from only 0.05 mm thick on the eyelids, to 1.5 mm thick on the palms and the soles of the feet. The *stratum corneum* is the main barrier of the skin for the external chemicals to enter the skin. It includes dehydrated and keratinised multilayer (bricks in vertical columns) in a lipid environment. *Stratum corneum* consists of two amorphous lipophilic (ability of a compound to dissolve in fats) and the hydrophilic (solubility of the chemical in water) layers. The lipophilic layer which is the main part of this layer, contains keratin and skin fat. Hydrophilic layer in contrast, include mostly corneocytes and natural moisturizing agents. The *stratum corneum* is a dense tissue, swelling more than many times its own thickness, in the water. The area between the cells is filled with cohesive laminae. Each cell is contained by a proteinaceous ingredient envelope.

The next layer is the dermis and is located exactly underneath of the epidermis layer. Its thickness is variable. Based on the site of the body it is located, it could vary from 0.1

to 0.4 cm which is 10-20 times thicker than the epidermis layer. This layer's responsibility is to provide flexibility and tensile strength for the skin. It includes collagen embedded in a gel texture area mix of mucopolysaccharides. It provides protection for internal body organs from injury and infection. In addition, the epidermal layer obtains the necessary nutrition from this layer.

Finally, the last inner layer, the subcutaneous layer, is located beneath the epidermis and dermis. Its fatty tissue has the responsibility to protect the skin from heat and shock. This layer carries the nutrition to the upper layers through the blood vessels and it also can transmit the pain through the nerves located in this layer. The thickness of the subcutaneous layer can be different in various body locations. The lower layer of this layer covers the muscles and the periosteum of the bones.

Apart from the skin layers, many appendages such as hair follicles, sebaceous glands, eccrine and apocrine sweat glands are linked to the skin. Normally, human skin consists of 40–70 hair follicles and 200– 250 sweat ducts per cm^2 of skin. The sebaceous glands are responsible for lubricating the skin surface and maintain the pH at around 5. Eccrine or sweat glands play an important role in heat exchange and temperature maintenance and they also respond to emotional stress.

As mentioned earlier, the chemicals applied to the surface of the skin should pass the *stratum corneum* which is the main barrier of the skin. It is likely that lipophilic chemicals tend to remain in the *stratum corneum*; however, the skin appendages can provide a pathway that molecules can enter the lower layers of the skin without having to pass through the *stratum corneum* barrier of the skin. There are a number of pathways for passing the chemicals through the skin which are so called polar and lipophilic pathways (Williams (2003); Moss and Cronin (2002)). The *reservoir effect* is where the drug diffuses into the *stratum corneum*. This is the slowest process in skin absorption, called the rate-limiting step. So, when chemicals pass into the skin the rate of passage across the *stratum corneum* is normally slow and the concentration builds up over time. This is the *reservoir effect*. It also means that, when a product is removed from the skin surface treatment / delivery is not at an end as there will still be drug in the *stratum corneum* which can be delivered deeper into the skin (Moss et al. (2015)).

2.2.1 Skin layers thickness

In general, the *stratum corneum* provides a most important barrier to the passage of applied chemical through the skin. So, it is important to consider the thickness of this layer of the skin at which the experiments have been performed and also to consider which layers of the skin are excised and used in *in vitro* experiments. The stratum corneum' s thickness varies from 10 and 40 μm on different body sites. Figure 2.1 shows the details of each layer of the human skin.

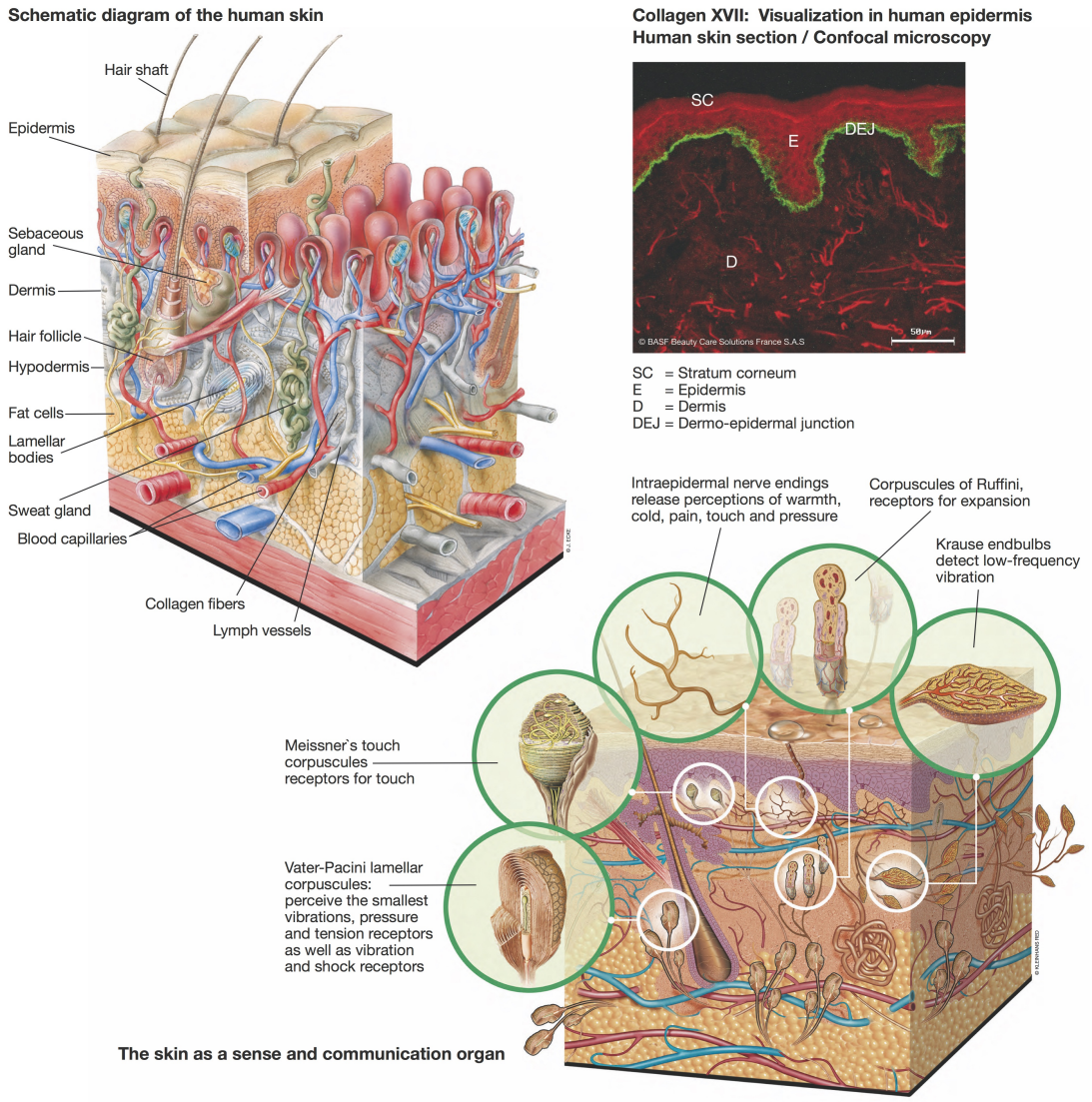


Figure 2.1: Skin layers from skin care forum (© KLEINHANS RED, Source: www.skin-care-forum.basf.com)

2.3 Physicochemical properties of the skin

In this section we define some of the terms related to skin and also the physicochemical properties of the skin we will see later in the QSAR/QSPR equations.

- The first term is **permeation**: it is defined by the amount of movement of drug through the membrane which includes partitioning of the molecules through the different layers of the skin.

- The next term is **penetration**: the amount by which the molecules enter into the tissue which does not necessarily includes passing the molecules out of the tissue.

- **Diffusion** is described by movements of molecules through a domain, for example: from a high concentration liquid to a low concentration liquid.

- The **diffusion coefficient (*D*)** of the molecule through the skin, defines how easily the permeant traverse through the tissue. It is expressed in units of area/time(cm^2/h or cm^2/s).

- **Partitioning** demonstrates the molecular distribution from one domain to another, such as from a lipid aqueous domain to water.

In addition some physicochemical properties such as partition coefficient, membrane thickness and lag time are defined as follows:

- **Partition coefficient (*P*)** illustrates the distribution of the molecules between two phases. To study the percutaneous absorption of exogenous chemicals, the partition coefficient usually employed, is the partition coefficient between octanol and water. Octanol simulates the *stratum corneum* lipids of the outer most layer of the skin to see how the molecules may traverse between *stratum corneum* and water (or the receptor in in vitro studies).

- **Membrane thickness (h)** shows the membrane thickness for various numerical calculations of the chemicals permeation. This measurement is usually difficult due to the thickness of the skin might be variable based on the skin hydration. In addition, it is not clear that the barrier to permeation only exists in *stratum corneum* or it may also occur in the epidermal layer of the skin. Due to the difficulty to measure the thickness of the skin membrane, researchers have estimated more reasonable ways through the *stratum corneum* than a simple assessment of the thickness. These researches revealed that traverse of the chemicals through the skin may occur through

different pathways (such as appendages) which are not simply obtained by measuring the thickness of the skin.

- **Lag time (L)** is the period that lasts during the rate of permeation through the membrane is increasing. It is obtained by measuring the time which the pseudo-steady state portion of a plot of the cumulative amount traverses the skin (Williams (2003)).

2.3.1 Flux (J_{max}) and permeability coefficient (K_p) definitions

J and K_p indicate the rate of absorption through the skin. J_{max} is the maximum dose of solute able to be delivered over a specific period of time and area that it is applied. These measurements predict the drug absorption uptake and toxicity, both of which are related to uptake or transport of cosmetics and other chemicals. The first law defined by Fick (1855) is mainly used to describe the ability of a compound to pass through the unit area of the skin and it is described by the below equation:

$$J = -D\frac{\partial C}{\partial x} \tag{2.1}$$

where J defines the rate of transfer per unit per area of the surface, defined by mol/cm^2/h) (i.e. the flux), C is the concentration of the diffusion substance, x is the spatial co-ordinate measures normal to the section and D is the diffusion coefficient, or diffusivity. Its dimension is area per unit time, which is defined by m^2/h or cm^2/s.

Fick's second law specifies that the change of concentration during the time in a particular region is proportional to the change in the concentration gradient at the same point; it is defined by Martin et al. (1993):

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} \tag{2.2}$$

solving Fick's second law is usually difficult due to its dependency to boundary conditions for an experiment and that is why usually the first law is used. More details can be found in Moss et al. (2015).

The dermal permeability coefficient, K_p, is defined by the equations:

$$J_{ss} = K_p C_v \tag{2.3}$$

or

$$K_p = \frac{J_{ss}}{C_v} \tag{2.4}$$

where C_v illustrates the concentration of the penetrant in the vehicle when sink conditions apply, J_{ss} is the steady-state flux of the solute. K_p expresses the permeability coefficient (cm/s or cm/h) which is the concentration-corrected flux and allows chemicals of different aqueous solubilities to be compared in a single data set. Permeability coefficient is dependent on the vehicle used. If the concentration (C_v) and the solubility (S_v) of the solute in the vehicle are known, J_{max} may be estimated from the experimental steady-state flux (J_{ss}). J_{max} defines the maximum steady-state flux of the solute (J_{ss}) and it is obtained from dilute solutions as follows:

$$J_{max} = \frac{S_v}{C_v} J_{ss} \tag{2.5}$$

K_p can also be obtained by below equation:

$$K_p = \frac{K_m.D}{h} \tag{2.6}$$

where D is the average diffusion coefficient (cm^2/s or m^2/h), K_m represents the partition, or distribution coefficient between the *stratum corneum* and the vehicle and finally h is the thickness of the skin.

From the aspect of the passage of chemicals through the skin, the *stratum corneum* is essentially a lipid layer, which interfaces with an aqueous medium placed underneath it. For penetration of lipophilic chemicals through the *stratum corneum*, they are transferred directly into an aqueous medium, therefore highly lipophilic compounds could not pass from the *stratum corneum* and mainly remain in it. Hence, there should be a number of pathways at which compounds may penetrate through the skin—the so-called polar and lipophilic pathways (Bronaugh and Maibach (1989)). Based on what was mentioned

earlier, it might be inferred that some descriptors such as hydrophobicity (typically quantified by the logarithm of the octanol–water partition coefficient) may have a high influence on the absorption rate of the skin. In addition, physicochemical properties such as molecular weight/size and possibly electronic properties such as hydrogen bonding play important roles in skin permeation. These properties obtained from graph theory and counts of paths and connections between individual atoms. Due to the difficulty in interpreting the topological indices, the use of other descriptors should be considered in a QSAR. To predict the permeability of a chemical, QSAR models should consider the physicochemical and/or structural properties of the compounds (Williams (2003); Moss et al. (2002)). These descriptors should be appropriate for that particular passive diffusion process. A considerable point to be noticed is that it is not only important to have a precise prediction of the chemicals permeation through the skin, but also for designing the drugs and for their effective passage through the skin, it is important to also take the physical mechanism into account. That is to examine which of the chemical features have the most important role in chemicals permeability and consider that when it comes to formulating the drugs (Lam et al. (2010)).

In the next sections, the use of descriptors in QSAR models are explained in more details.

2.4 The importance of experimental conditions

One important consideration is to understand the specific experimental conditions under which an experiment has been conducted, as even small changes could potentially alter the outcome of the experiment substantially. These conditions include temperature of the skin/ environment, the device used to apply the chemical to the skin (the formulation, or the vehicle), the cell type used in the experiment and the body site that chemical applied to the skin surface, *in vitro* or *in vivo* experiments. It is also interesting to know that repeating the same experiment with the same chemical applied to the same skin, may result in different measured permeability values due to the difficulty in keeping the experimental conditions constant at all times (for example temperature). A very important notion which should be born in mind is to develop predictive models for skin penetration based on QSAR, all

the data used for this purpose, should be measured using the same protocol, with the skin from the same animal (source), the same site and under the same circumstances such as the same workers and laboratory. Otherwise, the error of the model will be increased and the validity of the model generated based on this data will be decreased. Therefore, it is likely that QSAR models described in this research are subject to experimental errors if the mentioned laws have not been complied with.

2.4.1 Temperature

The effect of changing temperature on the physiological structure and activity of the skin is highly complex, affecting both local blood flow and metabolism. However, in the vast majority of published studies of skin absorption it has been demonstrated that the absorption rate will increase as the temperature increases and that it decreases in lower temperatures by up to one order of magnitude (Woolfson and McCafferty (1993)).

2.4.2 Static and flow-through diffusion cells

A diffusion cell is a cell contains receiver compartment and it is defined as one of the experimental conditions. There are two diffusion cell types:

- **Static diffusion cell:** an application containing the chemical is applied over the upper excised human/ animal surface of the membrane and the permeant samples are collected from the beneath in the stirred compartment periodically like every 2 or 4 hours.
- **Flow-Through diffusion cell:** the chemical molecules passing through the membrane are carried away by the solvent flowing underneath of the membrane and the samples are also collected periodically.

There are devices to automate the measurement of the chemicals fractions in the solvent to calculate the efflux solvent. In this research, we investigate the effect of using the static or flow-through cells on the prediction performances.

2.4.3 Regional variation (body site)

It is important to know that rate of absorption in the same individual could vary widely across different skin sites. Studies by Scheuplein (1967) and Elias et al. (1981) imply that skin permeability can be ranked in different body sites by the below descending ranking:

posterior auricular skin (behind ear) > scrotum > head and neck > abdomen > forearm >thigh > instep > heel > plantar

2.5 Risk assessments in human/animal data

It is difficult to assess the skin permeability of the chemicals using only *in vivo* experiments, especially when applying new compounds to the skin. General pharmaceutical/ risk assessment *in vitro* models is that it is difficult to obtain biological tissue. In addition, using specific cells from species is laborious. Another important point to be considered is that, testing on live animals and human volunteers need lots of risk assessments. Usually human skin from various sources, such as cosmetic surgery and amputations are employed *in vitro* studies of the skin permeability. Due to the fact that obtaining human skin data is difficult, some other animal's skin, particularly pigs, were found to be most comparable to permeation into and across human skin. Rodent skin, by contrast, has been shown to poorly mimic human skin despite its widespread use in toxicology studies. In this research, we investigate the similarity of using the animal skin data compared to the human skin data to see whether they have the same mechanism in terms of permeability. There is also a debate on reducing animal testing which depends on whether the models show similar or different mechanistic information, to support or refute the use of animals in such experiments, as animal testing has increased in recent years within the EU. The following section represents more details about the human and animal skin variation.

2.5.1 Human and animal skin variation

The study by Moss et al. (2011) examines pig, rat, mouse, human and artificial skin data. It reveals that permeation across rodent (mouse and rat) and pig skin is similar. In addition, the artificial skin can not be a good replacement for animal or human skin. It also appears

that the nature of the data and the size affect the model quality. This suggests that, in order to construct reliable models to predict the permeability of different mammalian membranes, we should consider including as much as commonality as possible in the models.

2.6 QSAR/QSPR models

Over the last 30 years extensive research has been carried out on finding the relation between the skin permeability of the compounds and the physicochemical properties of their molecules. This led to the prior studies of quantitative structure predictive models (QSPRs) which have noted the importance of employing such properties to predict the absorption rate of the chemicals through the skin. In these studies, skin permeability could be defined as either flux or permeability coefficients. To date, various methods have been developed and introduced to measure QSPR values. In many of these studies, the direct linear relationship between hydrophobicity and skin permeability has been reported (Roberts et al. (1977); Scheuplein (2011)). Although QSAR models could be derived from these research, each model can be defined for a particular or specific chemicals. In addition to the mentioned problem, there is a physicochemical variety in some of the descriptors. The results also reveal co-linearity relationship between some descriptors such as molecular weight and hydrophobicity. These findings indicate that it is not possible to discriminate the effect of hydrophobicity from molecular size on permeability of molecules.

Table 2.1 shows a list of the most cited QSAR models for predicting percutaneous absorption from Moss et al. (2012) paper. In this table K_p is the permeability coefficient (as cm/s or cm/h); $\log K_{ow}$, P_{pct} and $\log P$ define the octanol–water partition coefficient; K_{psc} is the permeation coefficient of the lipid fraction of the *stratum corneum*; K_{pol} is the permeation coefficient of the protein fraction of the *stratum corneum*; K_{aq} is the permeation coefficient of the aqueous permeation layer; *MW (*g/mol*)* is the molecular weight. The Dalton, is also sometimes used as a unit of molecular weight (molar mass), especially in biochemistry, with the definition 1 Dalton = 1 g/mol. *MPt* is the melting point (C).

Table 2.1: Most cited QSAR models for estimating the percutaneous absorption (Moss et al. (2012))

Name of the model and the citations*	Algorithm
Potts and Guy (1992)	$\log K_p = 0.71 \log K_{ow} - 0.0061MW - 6.3$
Brown and Rossi (1989)	$K_p = 0.1 \left[\frac{P_{oct}^{0.75}}{120+P_{oct}^{0.75}} \right]$
Cleek and Bunge (1993)	$K_p^{adj} = \frac{K_p}{1+(1400.K_p.\sqrt{MW})}$
Cronin et al. (1999)	$\log K_p = 0.77 \log P$–$0.0103MW$–2.33
Wilschut et al. (1995)	$K_p = \frac{1}{\frac{1}{K_{psc}+K_{pol}}+\frac{1}{K_{aq}}}$
Barratt (1995)	$\log K_p = 0.82 \log P_{oct}$–$0.0093MW$–$0.039MPt$–$2.36$
Moss and Cronin (2002)	$\log K_p(cm/h) = 0.74 \log P$–$0.0091MW$–$2.39$

2.6.1 The Flynn(1990) data-set and the related QSAR analysis

In 1990 a publication by Flynn (1990) on 97 permeability coefficient for 94 compounds, provided the largest and most significant data in skin permeability of *in vitro* experiments on human skin (with the exception of *in vivo* studies for toluene, ethylbenzene and styrene). The data were gathered from 15 different compilations, while they were all obtained from *in vitro* human skin studies. It is therefore, likely that such data has several sources for experimental error, due to the variability in their laboratory experiments. This variability could also be noted more, if the human skin obtained from different sites of the body under different experimental circumstances such as temperature. However, this publication was a significant source of data used in next researches and had a key role in development of QSAR models. Some of the major studies using Flynn (1990) dataset, are going to be reviewed as follows.

To predict the skin permeability from this data-set, Flynn points out that skin permeability was highly affected by the molecular weight and also partition between aqueous and non-aqueous layers (hydrophobicity in terms of the octanol–water partition coefficient). Flynn stated a simple algorithm regarding the skin permeability based on molecular weight and hydrophobicity/hydrophilicity of the chemicals which is illustrated in Table 2.2 on the following page. To have a better visibility on the algorithm, the relationships are also plotted in Figure 2.2. This algorithm indicates that the skin permeability ($\log K_p$) is directly related to size of molecule, in addition to the hydrophilic and hydrophobic properties; which indicates very hydrophilic and hydrophobic compounds had low and high skin permeability respectively, and additionally they are separated based on the high and low molecular weights. In his publication, he did not provide a statistical fit assessment of this model.

The data provided by Flynn has been used and analysed by many researchers afterward. One of the QSAR models, demonstrated by Potts and Guy (1992) using Flynn (1990) dataset as :

$$\log K_p = 0.71 \log K_{ow} - 0.0061MW - 6.3 \tag{2.7}$$

which was for 93 observations.

Table 2.2: Permeability coefficient ($\log K_p$) estimation based on the Flynn algorithm (Flynn (1990))

	Compounds with low *MW* (<150 Dalton)	**Compounds with high *MW* (>150 Dalton)**
$\log K_{ow} < 0.5$	$\log Kp = -3$	$\log K_p = -5$
$0.5 \leq \log K_{ow} \leq 3.0$	$\log Kp = \log K_{ow} - 3.5$	
$0.5 \leq \log K_{ow} \leq 3.5$		$\log Kp = \log K_{ow} - 5.5$
$\log K_{ow} > 3.0$	$\log K_p = -0.5$	
$\log K_{ow} > 3.5$		$\log K_p = -1.5$

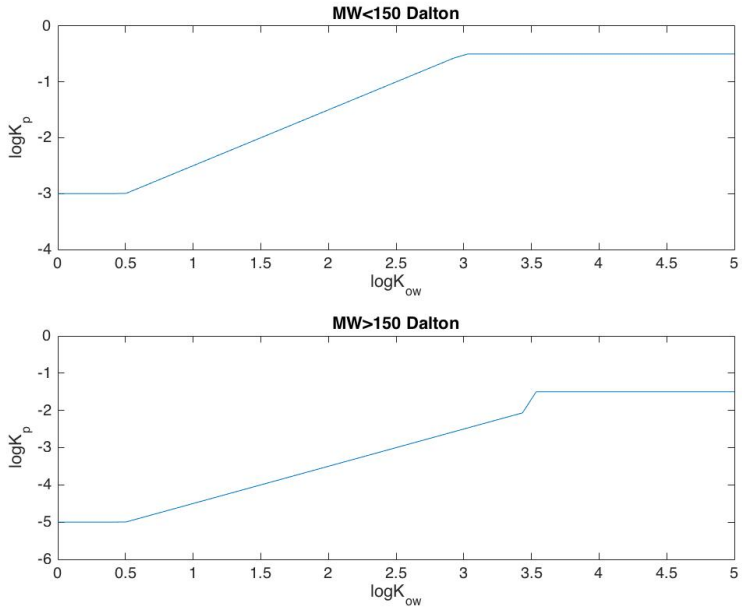


Figure 2.2: Permeability coefficient ($\log Kp$) relationship with *MW* and $\log K_{ow}$, based on the Flynn algorithm (Flynn (1990))

At the same time with Potts and Guy in 1995, Abraham et al. (1995) conducted a research on the Flynn dataset. Their later research in 1997 (Abraham et al. (1997)) and 1999 (Abraham et al. (1999)) on the 46 non-electrocyte compounds notified that some compounds such as steroids, have problems in their permeability coefficients. Additionally, their work showed the importance of hydrogen bonding in QSAR models. The positive coefficient with the parameters of molecular size, shows the correlation of the parameters with hydrophobicity ($\log K_{ow}$). Another important model was proposed by Moss and Cronin (2002), where $\log K_p(cm/h) = 0.74 \log P$–$0.0091MW$–2.39. We should note different methods have been used to calculate $\log P$ values (Moss et al. (2002)). Later in 2003 another model was introduced by Cronin and Schultz (2003), where $\log K_p(cm/h) = 0.77 \log P$–$0.010MW$–2.23 for 107 observations. In addition to the former researches, a recent study has been done by Magnusson et al. (2004) on five datasets. In this study, epidermal permeability coefficients (K_p) are optimally correlated to solute octanol–water partition coefficient (K_{ow}) and molecular weight (MW) was found to be the dominant determinant of J_{max} for on the literature data set: $\log J_{max} = -3.90 - 0.0190MW$ (n=87, r^2 =0.847, p<0.001). Adding the other physicochemical properties by forward stepwise regression could improve the predictions by a few amount. This equation was also validated with the other four datasets and the complete dataset. The datasets include full- and split-thickness skin data and their study revealed that the dermal resistance had only a small effect on the overall J_{max}. These datasets and the results are explained in greater detail in Chapters 3 and 7.

In this thesis, it is shown that trainable non-linear machine learning models outperform the traditional QSAR/QSPR prediction models for skin permeability.

Chapter 3

The Descriptions of Datasets

To do the experiments of this study, 16 datasets are used. The datasets include the physicochemical properties of the chemical compounds. Depending on the data different target is defined in each dataset. The datasets are collated from various sources and they are from the *in vivo* and *in vitro* experiments performed in the labs using the chemicals applied to the human and the the non-human (animals) skins.

I divide the datasets into four major groups. These four groups include human skin datasets, animal skin datasets (Mouse, rat and pig), magnusson datasets (Magnusson et al. (2004)) and the enhanced ratio dataset (Pugh et al. (2005)). The human datasets are collated by Prapopoulou (2012) as one set, which has been divided by her into 16 sub sets and we have used 6 of these subsets in this study. Animal datasets are also gathered by Prapopoulou (2012).

The aim of this chapter is therefore to explain the nature of the data sets, including the sources of their data, the descriptors used to characterise the members of the data set and to discuss the rationale for subdividing the dataset into representative subsets.

3.1 Terminology

All the physicochemical properties of chemicals in the datasets along with their units are illustrated in the following list:

- Molecular weigh(MW; g/mol)

- Molecular refractivity(MR; J mol^{-1} K^{-1})
- Melting point (MPt; *C*)
- Hydrophobicity *(P* or K_{ow} ; shown with log values, no units)
- Solubility parameter (*SP*; $(cal/cm^3)^{1/2}$))
- Counts of the hydrogen bond donors (*HD*; no units) and acceptors (HA; no units)
- Chain lengths of carbon atoms (*CC*; no units)
- H-bonding atoms (*HB*, no units)
- Solubility(S; shown with log values, unit mol/L)

In addition, the common pharmaceutical expressions/definitions used in this chapter are as follows:

- *Stratum corneum:* The outermost layer and the main barrier of the skin.
- J_{max}/flux: The maximum dose of solute able to be delivered over a specific period of time and area that it is applied (mol cm^{-2} min^{-1} or μg cm^{-2} h^{-1}).
- K_p: The permeability coefficient (cm/s or cm/h) which is the concentration-corrected flux(cm/s or cm/h)
- ER: The enhancement ratio of enhancers.

3.2 Human skin datasets

The datasets used in this section are collated from various sources. They are *in vitro* human skin permeability studies whose core is the Flynn (1990) data set as modified by Moss and Cronin (2002) and added new chemicals by Prapopoulou (2012). There are six datasets in this category which range in size from n=9 to n=86. The number of data points in each dataset is obtained after refining the data by removing the missing data and repetitions. In some of the datasets, there are chemicals with variable target values (data inconsistency).

The variability in the data depends on the biological variation and variation in methods used and in the quality of the work carried out (Chilcott et al. (2005)). Therefore, it results in needing multiple repeats of an experiment. It should be noted that this variability can not be entirely removed from a biological system. In order to deal with this issue in computational methods, the mean of target values are used for the same chemicals with different permeability values (inconsistent data). In Chapter 6, section 6.6 the Monte Carlo method is employed as another technique to solve the data variability problem.

Therefore the data refinement can be summarised in the below steps:

1. If a chemical has any missing value in the vector of its features, that chemical is deleted from the data.

2. The chemicals that have the same molecular features and target values are omitted.

3. For the same chemicals (with same molecular features) and different target values, the mean of those target values are used and assigned to the chemical and the repetitions are removed. So, at the end each chemical has only one target value assigned to it.

In addition, in gathering all the human data, a complete human dataset with 145 chemicals (after data refinement) is obtained which is used in the experiments in the following chapters. Table 3.1 shows the original and refined number of data points in each dataset.

There is also a degree of overlap between the datasets. For instance, dataset E includes the majority of chemicals in datasets A and B. Similarly, dataset F includes most chemicals in dataset C. All the common data among the original datasets are shown in Table 3.2. The reason for this is that the datasets are divided based on their experimental conditions and they may share some experimental features. This can be useful to see the same chemical prediction performance in different datasets. It can be especially used in pharmaceutical filed to examine which chemicals can be considered together to provide a model that can be used for the prediction purpose.

Table 3.1: Number of data-points in human datasets

Datasets sizes	Human A	Human B	Human C	Human D	Human E	Human F	Complete Human
# Original datasets	11	42	38	99	92	148	642
# After refining datasets	9	25	21	57	51	86	145

Table 3.2: Summary of the common data among the subsets

Similarities	Human A	Human B	Human C	Human D	Human E	Human F
Human A	11	9	0	3	10	0
Human B	9	42	0	20	41	0
Human C	0	0	38	32	0	36
Human D	3	20	32	99	32	59
Human E	10	41	0	32	92	0
Human F	0	0	36	59	0	148

3.2.1 Chemical features

There are seven measured molecular physicochemical properties for each member of the dataset. They are molecular weight (*MW*) which ranges from 18.02 to 454.45 g/mol, molecular refractivity (*MR*; 0 to 116.22 J mol^{-1} K^{-1}), melting point (*MPt*; -142 to 866*C*) , *P* (hydrophobicity; -4.47 to 8.39, log values, no units), solubility parameter (*SP*; 7.51 to 32.83 $(cal/cm^3)^{1/2}$), counts of the hydrogen bond donors (*HD*; 0 to 6, no units) and acceptors (HA; 0 to 10, no units) on each molecule. The absorption rate is measured by the permeability coefficient, K_p (as either cm/h or cm/s) and adjusted to cm/h in these studies; values range from -6.32 to 0.34 (log values). Usually the log of *P* is reported because the values of *P* have a wide range between 10^{-7} to 10^{7}. For the same reason, the log is normally reported for percutaneous absorption K_p (Williams (2003)). It is important to bear in mind that there are a number of missing or zero values assigned to *MR* and *MPt* features, which might be problematic.

It should be noted that these values are not the same scale. For example MW of 1 to 10 is not the same as MR or log P 1 to 10. Therefore, before performing the experiments, these values are normalised as *Z-score*. In this method, the average values of the features (vector of average values) are first subtracted from the data and the results then are divided by the d-dimensional standard deviation of the features.

3.2.2 Experimental condition features

The experimental conditions for each dataset are also various . For example in some cases the temperature at which absorption measured was skin temperature but for some of them the temperature was set to a specific amount. The experimental conditions include temperature (numerical feature) and diffusion cell types (explained in Chapter 6), the body site and the skin layer at which the chemical is applied are nominal features. These features are important to be considered as they can have a large effect on the permeability of the chemical through the skin. As an example, one obtains different permeability coefficient values when the chemical is applied to the skin of upper arm and forearm of the same person.

The skin thickness are 0.2-0.5 mm (dermatome) for dataset A and can be varied from full thickness, dermatome (with any length), epidermal and *stratum corneum* for datasets B-F. Sites are abdominal for dataset A and all types for datasets B-F. The cell type in which the chemical applies to the skin is flow-through for datasets A, B and E, it is static in datasets C and F and mix of both static and flow-through cells for dataset D. The nominal features that are used for further experiments in Chapter 6 are the skin thickness, body site and cell type.

The temperature is constant (37°C) in human sets A to D, while it varies in the human E from 31° to 37°C, and in the human F dataset from 22°C to 45°C. The temperatures 32°C at the surface of the skin are considered as 37°C (human sets A to D). The other experimental conditions are the 'time' which is the duration that the experiment has been completed, vehicle and receptor fluid. For more information, the experimental features of the datasets are summarised in Table 3.3. The datasets, their chemical features and experimental conditions are reported in Appendix B, section B.1.

3.2.3 Human sets numerical features analysis

To have a better visualisation on the features and their ranges in each dataset, the six datasets with the same features in addition to their logK_p ranges are plotted in box-plot separately in Figure 3.1. From this figure, the differences among the features and permeability ranges of the 6 datasets can be seen and the features' statistics can be used later for the purpose of data and performance analysis. The figure shows that, human datasets D and F

Table 3.3: Summary of the experimental conditions for the human datasets

Conditions/Datasets	Human A	Human B	Human C	Human D	Human E	Human F
Skin thickness	dermatome 0.2-0.5 mm	**All types	**All types	**All types	**All types	**All types
Site	Abdominal Cell	All body sites	All body sites	All body sites	All body sites	All body sites
Cell type	Flow-through	Flow-through	Static	Flow-through/Static	Flow-through	Static
Temperature	37°C	37°C	37°C	37°C	from 31° to 37°C	from 22° to 45°C
Time	24 – 72 h	24 – 72 h	24 – 72 h	Not specified	Not specified	Not specified
Vehicle	*Specific vehicles	*Specific vehicles	*Specific vehicles	Any	Any	Any
Receptor fluid	Normal saline	Normal saline	Normal saline	Any	Any	Any

*Specific vehicles: Water, physiological buffer, propylene glycol, ethanol, methanol
** All types: Full thickness, dermatome, epidermal, stratum corneum

seem to cover the same range of chemical features but not temperature (which is constant, 37°C in dataset D) and dataset F covers a slightly larger target range than dataset D. Dataset A seems to cover a smaller range on *MW, MPt, MR, SP, HA* and target values compared to the other datasets which is probably due to the small size of this dataset.

3.3 Animal skin dataset and chemical features

The animal datasets used for this research are mouse, rat and pig datasets as they are commonly used as reasonable replacements for human data. However, thickness of the *stratum corneum,* the number of appendages per unit area and the amount of lipids in the skin of human are different from many of animals (Moss et al. (2011)). However, the experiments since 1992 on animal skin are usually validated with human skin data and in this study it is also investigated (see Chapter 2, section 2.5).

There are five measured molecular physicochemical properties in each dataset. They include *MW* which ranges from 18.02 to 959.17*, SP* (8.14 to 32.83)*,* $\log P$ (-4.27 to 8.10)*, HA* (0 to 11) and *HD* (0 to 8) and the permeability coefficient, K_p(cm/h) ranging from -6.29 to 0.10 (log values). As there were a lot of missing values for MR and MPt, they have not been used for the experiments. The datasets are collated from the same sources as the human datasets (Flynn (1990), Moss and Cronin (2002) and Prapopoulou (2012)). The sizes of these datasets after refining and removing the repetitions can be found in Table 3.4.

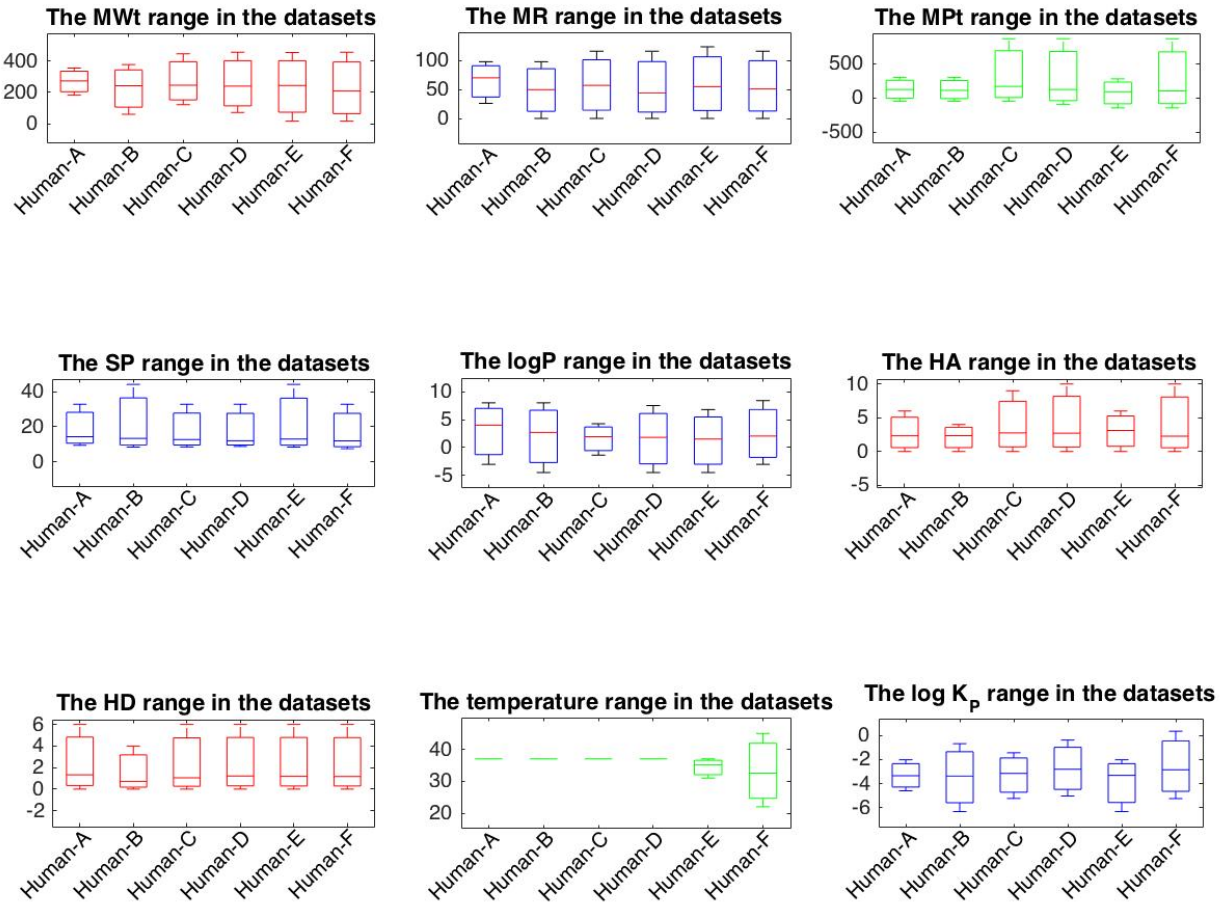


Figure 3.1: Ranges of features and targets in human datasets

Table 3.4: Number of data-points in animal datasets

Datasets sizes	Mouse	Rat	Pig
# After refining datasets	46	26	14

Table 3.5: Summary of the experimental conditions for the animal datasets

Conditions/Datasets	Mouse	Rat	Pig
Skin thickness	Full thickness	**All types	Full thickness and epidermal
Site	All body sites	All body sites	Outer ear
Cell type	Flow-through/Static	Flow-through/Static	Flow-through/Static
Temperature	25°C and 37°C	30°C and 37°C	from 30°C to 37°C
Time	Not specific	Not specific	4, 24, 48h
Vehicle	Any	Any	*Specific vehicles
Receptor fluid	Mainly Saline	Any	phosphate buffered saline PH 7.4

*Specific vehicles: Water, physiological buffer, propylene glycol, ethanol, aqueous saturated solution, isotonic phosphate buffered saline (pH7.4) with ethanol, propylene glycol, Azone
** All types: Full thickness, dermatome, epidermal, stratum corneum

3.3.1 Animal sets experimental conditions

Table 3.5 illustrates the experimental conditions for mouse, rat and pig datasets. The mouse, rat and pig datasets features and full details can be found in Appendix B, Tables B.2.1.2, B.2.2.2 and B.2.3.2, respectively.

3.3.2 Animal sets numerical features analysis

Figure 3.2 shows the features' ranges and the permeability coefficient ranges related to each compound. It can be seen in the plots that mouse dataset covers a larger range in all features. Although pig dataset has the smallest number of chemical compounds, it covers the largest range of permeability coefficient values ($\log K_p$).

3.4 Magnusson datasets

Datasets in this group are obtained from Magnusson et al. (2004). Authors either acquired J_{max} values from aqueous solution across human skin or estimated them from experimental data and correlated them with solute physicochemical properties. J_{max} (mol per cm^2 per h) is the maximum dose of solute which can be delivered over a given period of time and within a defined area from a given vehicle. The database consist of five separate sets: (1) Mag-set A, include a training set of 87 records (85 in our dataset after refining); (2) Mag-set B, full and split-thickness skin (from 0.01 to 2 mm) set of 56 records (50 after refining); (3)

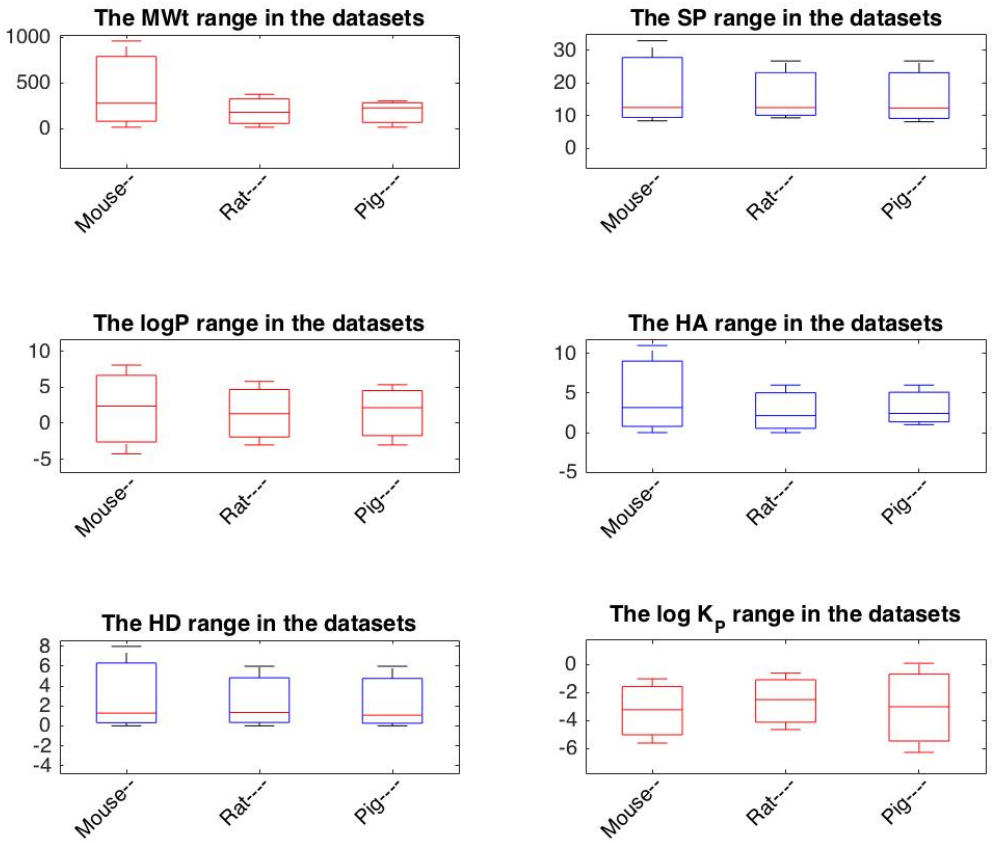


Figure 3.2: Ranges of features and targets in animal datasets

Mag-set C, including a pure liquid vehicle set of 34 records (27 after refining); (4) Mag-set D, including an ionised solutes set of 54 records (45 records after refining); (5) Mag-set E, PG vehicle set of 36 records (Kasting et al, 1987). Appendix B, section B.4 shows datasets records (from Table S1 in Magnusson et al. (2004)).

3.4.1 Magnusson datasets features analysis

The datasets used for this research are obtained under experimental temperature ranging from 22°C to 39°C. Ranges of molecular properties are: *MW* from 18 to 764.90, $\log K_{ow}$/$\log P$ from -5.60 to 8.70, *Mpt* from 147 to 573, *HA* from 0 to 13 and *HD* from 0 to 8. The other experimental conditions are not specified. All data features, their ranges along with the $\log J_{max}$ values are demonstrated in Figure 3.3. The plots show that in general, sets A and B cover larger ranges of numerical features and log J_{max} values.

3.5 The enhanced ratio (ER) dataset

Various methods have been used to enhance percutaneous absorption. To do so, the enhancers can be mixed with the chemical formulation. One way is to to add the enhancers such as ethanol or propylene glycol to the formulation vehicle. The enhancement ratio (ER) dataset is obtained from Pugh et al. (2005). It includes 73 enhancers of hydrocortisone permeation from propylene glycol across hairless mouse skin (we used 71 data records after removing the repetitions and averaging the target for the same compounds). The molecular properties are chain lengths (*CC*) from 0 to 16 carbon atoms, H-bonding atoms (*HB*) from 1 to 8, *MW* from 60 to 450, $\log P$ (calculated) 1.7 to 9.7 an*d* $\log S$ (calculated) 7.8 to 0.7. These predictive properties were chosen because of their ready availability. ER is defined as hydrocortisone transferred after 24 hours relative to control. Values of ER are ranged from 0.2 to 25.3. The $\log ER$ values are usually considered, because ER have a wide range between 10^{-7} to 10^{7}. Using principal components analysis (PCA) (see Chapter 5 for PCA), Pugh's study showed that that good enhancers could be identified by a combination of relevant descriptors: CC, HB and molecular weight. The ER dataset can be found in Appendix B, sectionB.3.

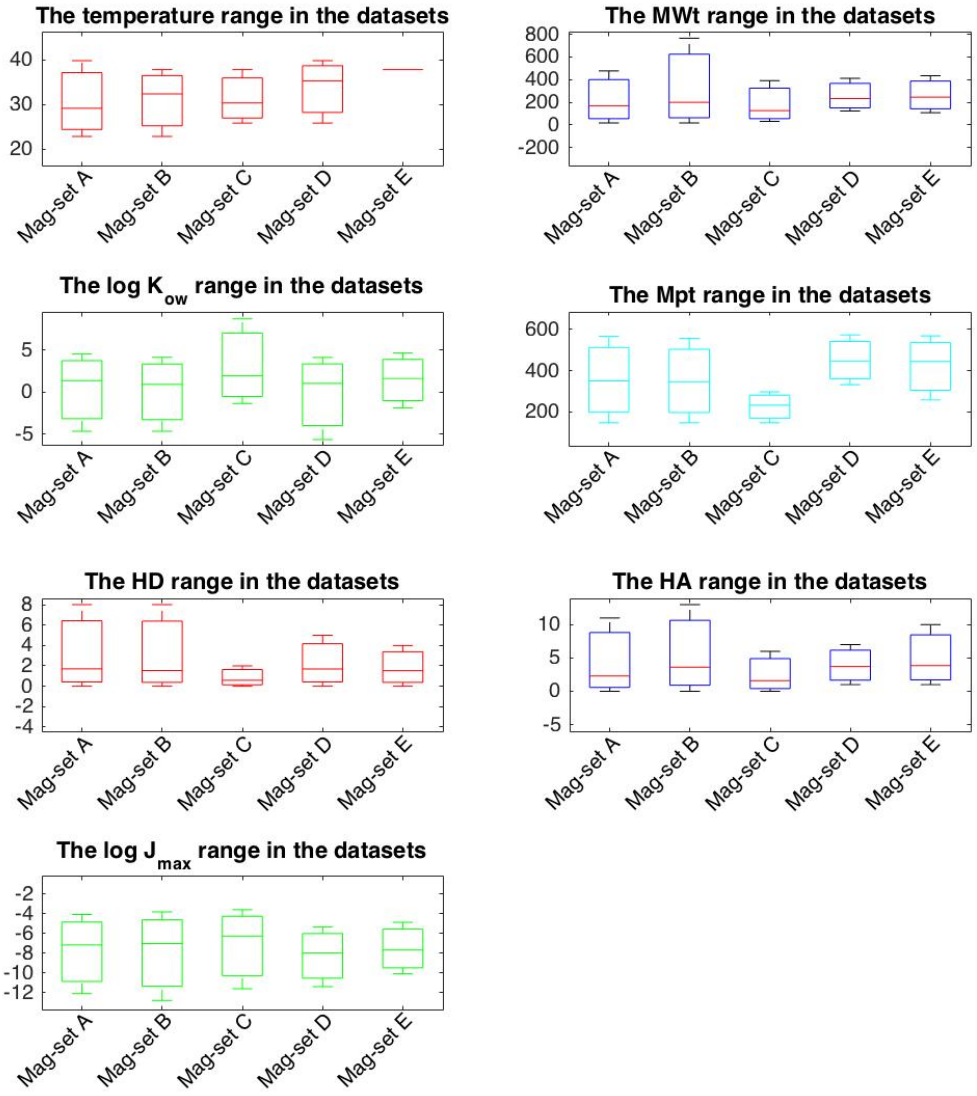


Figure 3.3: Ranges of features and targets in magnusson datasets

3.6 Analysis of the same numerical features among all the datasets

In this section, common features in all datasets are compared together and the ranges of features can be seen in Figure 3.4. In this figure, the empty columns above some datasets in each plot show that feature does not exist in the dataset. It can be seen that in general, the range of each feature varies among datasets. However, the mean values over all datasets are similar, except for *MPt*, where mean values of human sets are lower than those in the Magnusson sets (Animal data do not have reported *MPt*). *MW* range is largest in the Mouse dataset and Mag-set B, however, the mean values are almost the same overall the datasets. The *MPt* ranges are larger for datasets Human C, D and F and it is the smallest in Mag set-C. The mean values of SP is almost the same for all datasets and the ranges are larger in Human B and Human E datasets. However, both mean and ranges of log *P*/log K_{ow} vary over all the dataset. *HA* and *HD* have the similar average values in all the datasets and their ranges are various among the datasets.

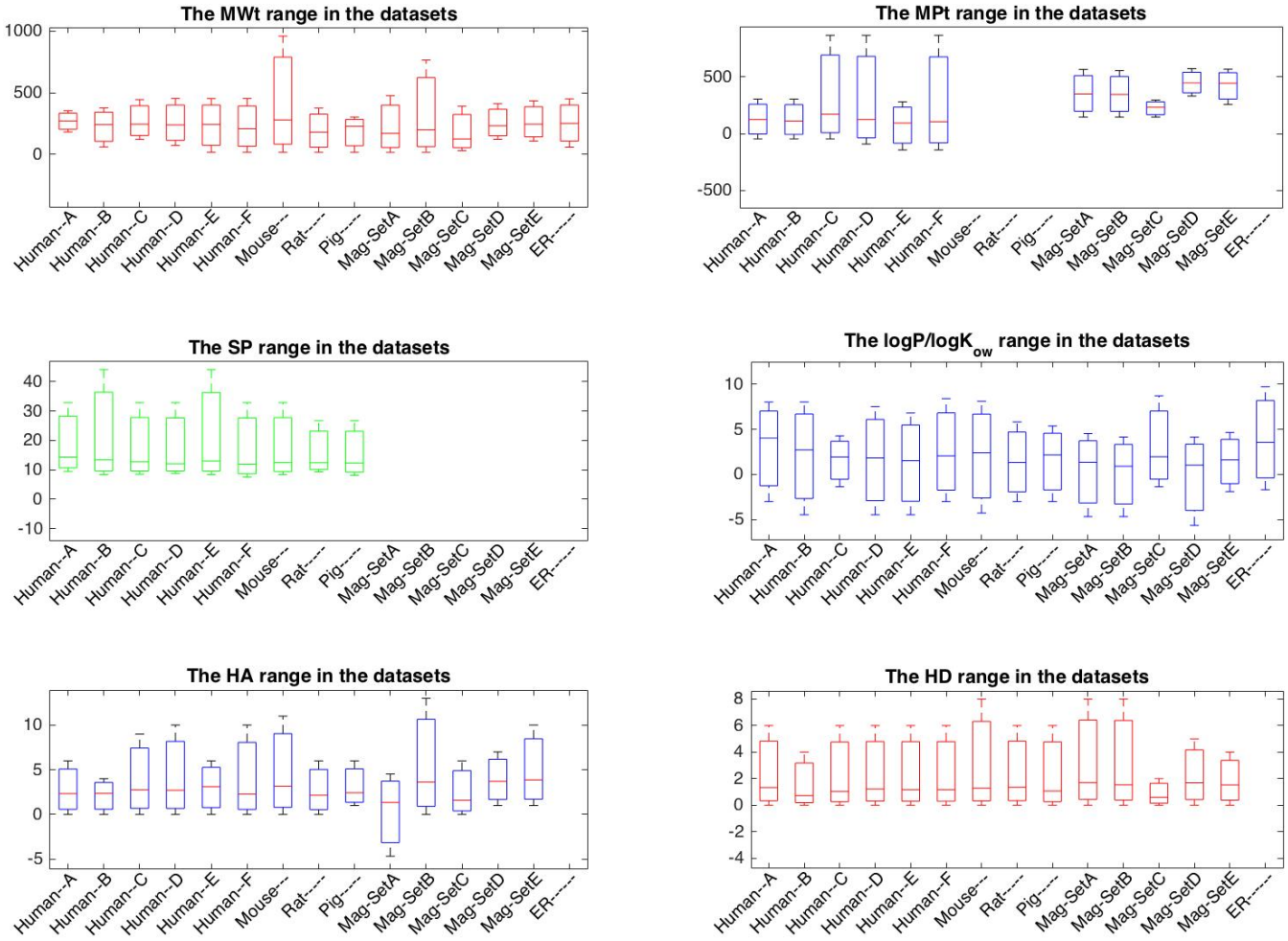


Figure 3.4: Comparison of ranges of features in all datasets

Chapter 4

Machine Learning Techniques

Machine Learning (ML) is a type of artificial intelligence method (AI); ML techniques explore the data and develop algorithms that can learn from data. It enables the algorithms to make a prediction or decision on new data. ML is generally divided into two main types: supervised learning and unsupervised learning, and which also considers reinforcement learning which is a blend of these two types. Supervised learning models the relationship from inputs to targets (or outputs). A large number of pharmaceutical and clinical data analyses are performed using this method. Supervised learning is subdivided into two primary tasks: classification and regression. In classification the outputs are discrete labels, whereas in regression the outputs are continuous variables. In this thesis the focus is predominately on regression, which is prediction of a continuous quantity and is dependent on a set of continuous and sometimes nominal (categorical) inputs, from noisy measurements. In other words, regression is the process of abstracting from functional data. In a linear case, fitting a hyperplane to the data, or in the case of non-linear regression, fitting an appropriate function to the data. An appropriate function is the one which with enable predictions to be made for novel data effectively.

In this Chapter, all the ML and computational/statistics methods used in this thesis are explained with details. The Gaussian Processes regression is the major method used in this thesis. A detailed description of GP, kernels and optimisation methods can be used for selecting hyper-parameters are given in Sections 4.2 and 4.3. To compare the linear QSAR model that is used in this thesis, a trainable linear regression model (a Single Layer Network

method) is also described in section 4.4. Sections 4.5 and 4.6 discuss two more widely used non-linear regression methods KNN and SVM, respectively. Moreover, a clustering algorithm, Growing Neural Gas, is shown in Section 4.7. The advantage of this algorithm is that the user does not need to initialise the number of clusters. In addition, the Monte Carlo method is mentioned in section 4.8 which has been used to deal with data inconsistency. Finally, section 4.9 discusses performance measurements applied in this thesis..

It should be noted that the visualisation methods used to discover the patterns and relationships in the datasets are discussed in Chapter 5.

4.1 The Prediction Problem

To do predictions, let us consider a dataset ($\mathscr{D}$) with N number of data points ($\mathbf{x}_1$,...,$\mathbf{x}_\mathbf{N}$). Each data point has D number of descriptors and their corresponding continuous target values $y_n(n = 1,...,N)$. Our assumption is that the outputs are noisily observed from an underlying functional mapping $f(\mathbf{x})$. Our aim is to find a probability distribution over likelihood functions from the data $\mathscr{D}$. There are several methods for performing regression. Gaussian Process is the primary method I investigate with details in this thesis. GP defines a probability distribution over functions *p(f)*. This can be used as a Bayesian prior for the regression. To make predictions from data then, Bayesian inference can be used as:

$$p(f|\mathscr{D}) = \frac{p(\mathscr{D}|f)p(f)}{p(\mathscr{D})} \tag{4.1}$$

This brief overview shows how GP solves the regression problem, as it provides probabilistic predictions of possible interpolating functions f (Snelson (2007)). The complete Gaussian Process usage for regression is explained with more details in the next sections. It is notable that all the notations used in this and the next section, are from Rasmussen (2006a) and Rasmussen (2004).

4.2 Gaussian Process introduction

The Gaussian process (GP) is a simple and general class of probability distributions on functions. GP is a non-parametric regression method, which means it does not assume a particular functional form, but allows the form of the relationship between inputs and targets to be determined entirely by the data which may include infinitive number of functions. It is assumed that the underlying function that produces the data will remain unknown but that the predictions are generated from a set of functions with a Gaussian distribution in the function space. Gaussian Process (GP) is an increasingly important area in machine learning to find a nonlinear regression such as a function estimation from the training data. It has been successfully used in various applications such as predicting skin permeability of the chemicals (Sun et al. (2008); Moss et al. (2009); Sun et al. (2011)), transmission spectroscopy by Gibson et al. (2012) and prediction of ozone concentration in the air Petelin et al. (2013). To be able to understand the posterior Gaussian Process to achieve the predictions for the unseen test set based on the training set, knowing the expressions, below, is necessary:

Definition 1: Likelihood functions A likelihood function is a function of the parameters of a statistical model such as GP. If the model involves a set of parameters θ(for GP, they are all hyper-parameters), then the likelihood of θ, given outcomes $\mathbf{y}$, is equal to the probability of the data given the hyper-parameters, $\mathscr{L}(\theta|y) = p(y|x,\theta)$.

Definition 2: Inference methods Inference methods compute the posterior and the negative log marginal likelihood and its partial derivations with respect to the hyper-parameters. Depend on its usage for regression or classification, we may use different inference methods such as Exact or Expectation Propagation (EP) inference method. The Exact method approximates the function with Gaussian likelihood and it can be only used for Gaussian likelihood. The EP is a method by Minka and Picard (1999), which is an iterative method that approximates each data point likelihood term by a scaled Gaussian. Therefor, it gives an overall Gaussian approximation the non-GP likelihood function. In this thesis the Exact inference method is employed as the Gaussian likelihood function is also used.

4.3 Gaussian Process for regression

A simple regression method finds the weighted average of target of two points nearest to the new point. The actual Gaussian Process uses a Gaussian function of distance as the weight. To give a simple example, suppose we have two values of x, for which we know the values of the dependent variable y, as $(x_1, y_1) = (1,2), (x_2, y_2) = (3,4)$. We want to predict the value of y for a new value of x, shown as x_* . This is achieved in the GP modelling by using a weighted average of known values of y, with the weighting determined by the closeness of x_* to each value for x in the data. For example, if $x_* = 2$, it is equally close to the original x values (1 and 3). So each is given equal weight and the prediction is the average of y, 3. If x_* = 2.5, which is 3 times as close to 3 as to 1, then a weighted average of the known values of y would be [(0.75 × 4) + (0.25 × 2)], or 3.5. It should also be noted that using the reciprocal of the distance between points is limited as (1 / distance) approaches zero, yielding inconsistent results. In a GP model, the actual weighting is not exactly proportional to the separation from the known values; rather, it is a Gaussian function of that distance. To have a more insight into the basics of the GP, and see more examples, Ashrafi et al. (2015) publication in the international journal of SAR and QSAR in environmental research can be found in Appendix C.

A GP is a collection of random variables $f(\mathbf{x})$ which have a joint Gaussian distribution for any set of inputs. As mentioned previously, it is a method to deal with nonlinear regression problems. If we choose a particular finite subset of these random function variables such as **f**= $\{f_1, f_2, ..., f_N\}$, with corresponding inputs **X** = $\{x_1, x_2, ..., x_N\}$, then any set of random function variables is distributed multivariate Gaussian:

$$f \sim \mathscr{GP}(m,k), \tag{4.2}$$

where $\mathscr{GP}(m,k)$ illustrates that the function f is distributed as a GP with mean function m and covariance function k. The simplest mean function is usually considered to be zero. Therefore, targets are predicted from a joint Gaussian distribution with zero mean and a covariance matrix. The covariance function or kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ is the expected correlation between values of $f(\mathbf{x})$ at the two points $\mathbf{x}_i$, $\mathbf{x}_j$. In other words, the kernel defines the nearness or similarity between data points, and allows specifying a-priori information

from training data for solving the regression problem. The kernel methods are standard machine learning algorithms; they use kernel functions to embed the data into a high or infinite dimension feature space. A kernel function computes the inner product of the embedding of two data points under a certain mapping in the feature space. The feature space that the data is embedded to, is expected to capture and enhance the patterns and regularities in the data. Then using standard algorithms of classification or regression the regularities of the data are investigated in the feature space.

For every input $\mathbf{x}$ there is an associated random variable $f(\mathbf{x})$, which is the value of the (stochastic) function f at the location. The Gaussian Process is over functions. However, to be able to generalise the Gaussian Process we may move from the Process to the distribution using a finite vector as m and a covariance matrix as $k(\mathbf{x},\mathbf{x}')$ to make the Gaussian Process feasible.

4.3.1 Covariance functions for numerical data

A variety of kernels or covariance functions can be used in a GP model. In our initial research, the Matérn, Polynomial and Gaussian covariance functions are applied to the data.

The Matérn covariance functions has a positive parameter, ν. This function becomes especially simple when ν is half-integer: $\nu = p + 1/2$, where p is a non-negative integer. The Matérn covariance function can be defined as a product of an exponential and a polynomial of order p. The most interesting cases for machine learning are $\nu = 3/2$ and $\nu = 5/2$ and they are defined as (Rasmussen (2006b)):

$$\mathrm{k}_{\nu=3/2}(r) = \left(1+\frac{\sqrt{3}r}{l}\right) exp\left(-\frac{\sqrt{3}r}{l}\right) \tag{4.3}$$

$$\mathrm{k}_{\nu=5/2}(r) = \left(1+\frac{\sqrt{5}r}{l}+\frac{5r^2}{3l^2}\right) exp\left(-\frac{\sqrt{5}r}{l}\right) \tag{4.4}$$

where r $=|\mathbf{x}-\mathbf{x}^{'}|$ and l (length scale) is a positive hyper-parameter. l defines the distance can be moved in input space, before the function value is changed significantly. They are called hyper-parameters since they are not set directly from the training set and studies have

proved they can affect the model predictions sometimes with a great amount (Snoek et al. (2012); Bergstra et al. (2011)). Note that as GP is a non-parametric method, there are not any ordinary parameters set from the training set. The covariance function, should be then multiplied by signal variance, σ_f^2 (noise in the data). More information can be found in Abramowitz and Stegun (1964).

Polynomial covariance functions are defined as :

$$k_{Poly}(\mathbf{x}, \mathbf{x}^{'}) = (\sigma_f^2 + \mathbf{x} \cdot \mathbf{x}')^p, \tag{4.5}$$

where p is a positive integer and σ_f^2 is the signal variance. Dot product covariance functions are invariant to a rotation of the coordinates about the origin, but not transitions. A simple example is the covariance function $k(x,x') = \sigma_f^2 + x \cdot x'$ which can be obtained from linear regression by putting $\mathcal{N}(0, 1)$ priors on the coefficients of $x_d (d = 1, ..., D)$. D is determined by the number of chemical compounds' properties. Polynomial covariance functions are usually not positive definite for all input dimensions but their validity is restricted up to some maximum dimension D. Therefore, we should not chose the D values larger than data dimension. Being positive definite is necessary for covariance functions in the GP (otherwise, calculations generate error). A symmetric matrix like covariance function, is positive definite if all its eigenvalues are non-negative.

In addition, the Squared Exponential covariance function is defined by the equation:

$$k_{SE}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 exp(\frac{-r^2}{2l^2}), \tag{4.6}$$

where l is the length scale, $r = |\mathbf{x} - \mathbf{x}'|$, and σ_f^2 is signal variance. The squared exponential covariance function corresponds to a Bayesian linear regression model with an infinite number of basis functions (Rasmussen (2006b)).

As we shall see later (Chapter 6), the best prediction performances are obtained using the Matérn covariance functions. To see the shape and smoothness of the Matérn covariance function, it is plotted by changing the ranges of r and l values. Figure 4.1 represents the 3D plot of the Matérn covariance function with $\nu = 3/2$, changing the l values in a range from 0.1 to 10 and r values from 0.1 to 10.

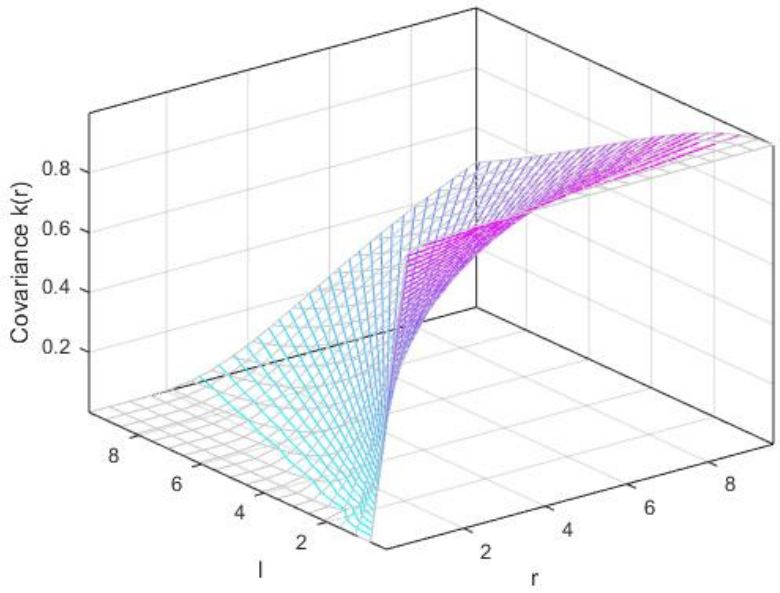


Figure 4.1: Matérn covariance function $\nu = 3/2$ and changing the r and l values

4.3.2 Covariance function for categorical data

Some of the pharmaceutical data used for the purpose of this research involve the experimental conditions and nominal data, such as body site, cell type and membrane (skin layer) in which the chemical compounds applied through the skin. To investigate the probable effect of this information, we need to calculate a categorical covariance matrix to be combined with the numerical covariance matrix (such as Matérn). Having this data, a kernel called categorical kernel by Couto (2005) is employed. It is then added to the Matérn kernel used to get the GP predictions for numerical data (molecular features).

4.3.2.1 Hamming Distance Kernel Function

The kernel function employed for categorical data is based on Hamming distance, which simply defines the number of positions at which the corresponding symbols are different between two strings of equal length. This functions embeds the categorical data into an inner product feature space. This kernel does not rely on a generative model or *a priori*

information about the nature of the data. The kernel function $K_H(s,t)$ between two categorical inputs s and t is defined as (Couto (2005)):

$$K_H(s,t) = \sum_{i=1}^{n} \phi_u(s)\phi_u(t) = \sum_{u \in D_c} \prod_{i=1}^{c} \lambda^{\delta(u_i,s_i)} \lambda^{\delta(u_i,t_i)}, \tag{4.7}$$

where $\phi_u(s)$ and $\phi_u(t)$ are the mapping of categorical objects s and t into the feature space which is defined by the u coordinate. $\lambda \in (0,1)$, $\delta(x,y)$ is 0 when $x = y$ and 1 otherwise. c is the number of categorical features and D_c is the number of categories in each of the categorical features. The kernel can be calculated recursively as follows:

$$K^0(s,t) = 1$$

$$K^j(s,t) = ((\lambda^2(|D_j| - 1 - \delta(s_j,t_j)) + (2\lambda - 1)\delta(s_j,t_j) + 1)))K^{j-1}(s,t),\ 1 \leq j \leq c$$

$$K_H(s,t) = K^n(s,t). \tag{4.8}$$

$|D_j|$ is the number of categories in each of the categorical features of the data. Therefore, it can be different from feature to feature in each dataset.

The normalised kernel $\tilde{K}$ of a kernel function K is computed as follows:

$$\tilde{K}(x,y) = \frac{K(x,y)}{\sqrt{K(x,x)K(y,y)}} \tag{4.9}$$

This kernel then can be combined (weighted multiplied/added) with the numerical kernel to be used for the prediction of the unseen points target values. More information about this kernel could be found in Couto (2005).

Test for Toy data

To understand the functionality of this method, a toy data with three nominal features is introduced. Three features are gender, nationality and native English speaker (Y/N).

Table 4.1: Toy data with nominal features

No	Gender	Nationality	Native English Speaker
1	Female	American	Yes
2	Female	Australian	Yes
3	Male	Egyptian	No

The data is defined in Table 4.1. Here the values of D_j are 2, 3 and 2 for the mentioned dimensions, respectively. Various values of λ between (0,1) can be considered. In this example it is defined as λ=0.3. To obtain the δ value between the features of two objects, we compare the features one by one and according to the explained rules it is either 0 or 1 for each feature of the two objects. As an example, the δ between the features of the first and second object is (0, 1, 0). This means they have two similar features and their second feature is different. The initial value of K is 1 and then it is calculated recursively for all the features of these two objects using Equation 4.8. Using this equation, the $K(1,2) = 0.58$ between the first and second objects. Similarly the other K values between the objects can be obtained. The values in the main diagonal of the K matrix is 1 as it defines the similarity of each object with itself which is 1. The final K matrix is:

$$K = \begin{bmatrix} 1 & 0.58 & 0.09 \\ 0.58 & 1 & 0.09 \\ 0.09 & 0.09 & 1 \end{bmatrix}$$

This matrix shows that the similarity between the first and second data is more than the similarity between the first and third or the second and third data (with no similarities!).

4.3.3 Posterior Gaussian process

The most important aim of finding the posterior is to make predictions for the test data target values. To make a prediction y_* at a new input $\mathbf{x}_*$, the conditional distribution $p(y_*|y_1,...,y_N)$ on the observed data $[y_1,...,y_N]$ should be assessed. Since our model is a Gaussian, this distribution is also a Gaussian and is completely determined by its mean (prediction of new inputs) and variance (predictive variance), which can be calculated using standard linear algebra:

$$E[y_*] = \mathbf{k}_*^T.(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} \tag{4.10}$$

$$var[y_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_*, \tag{4.11}$$

where k_* is the covariances between the test point and training data. $\mathbf{K}$ denotes the covariance matrix of the training data, $\mathbf{I}$ denotes identity matrix with 1's on the diagonal and zeros elsewhere, σ_n^2 denotes the variance of an independent identically distributed Gaussian noise which means observations are noisy, and y denotes the vector of training targets.

In Equation (4.11), $k(\mathbf{x}_*, \mathbf{x}_*)$ is the variance of $\mathbf{x}_*$. The predictions of a GP, are based on the weighted average of the known values of $\mathbf{y}$, with the weighting given by the proximity of $\mathbf{x}_*$ to each $\mathbf{x}$ in the training data. The matrix $(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}$ is completely independent on a new point and is a linear transformation that forces the Gaussian weights to perform interpolation, which results in zero value for all but one of the known data points. The transformed weights are multiplied by $\mathbf{y}$ values and the prediction achieved. GP regression is also able to provide the variance of the predictions, which reports the error of the predictions. If the weights are high it means that the new data point should be near to some of the original data and our prediction is precise. Based on the Equation (4.11), if the weights are large then the variance should be low and vice-versa (Rasmussen (2006b)).

Considering the examples in Rasmussen (2004) and to make the process much clearer, using 20 observed data, a sample from posterior process is drawn in Figure 4.3. In this example, the mean and covariance functions are $m(\mathbf{x}) = 0$, and $k(\mathbf{x}, \mathbf{x}') = exp(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^2)$. The prior GP is explained in the Appendix A.1. Comparing this figure with Figure 4.2 (obtained from prior GP, see A.1) one can see the uncertainty decreases close to the training data (20 observed data). It could be inferred that the posterior variance is always smaller than the prior variance, since the data has given us some additional information.

So far, we have seen how we can update the prior using the training data to obtain the posterior. However, in this case we should have a prior information about a dataset to specify its mean and covariance functions. Unfortunately, in ML problems we do not usually have this detailed prior information about the data. In order to obtain the prior

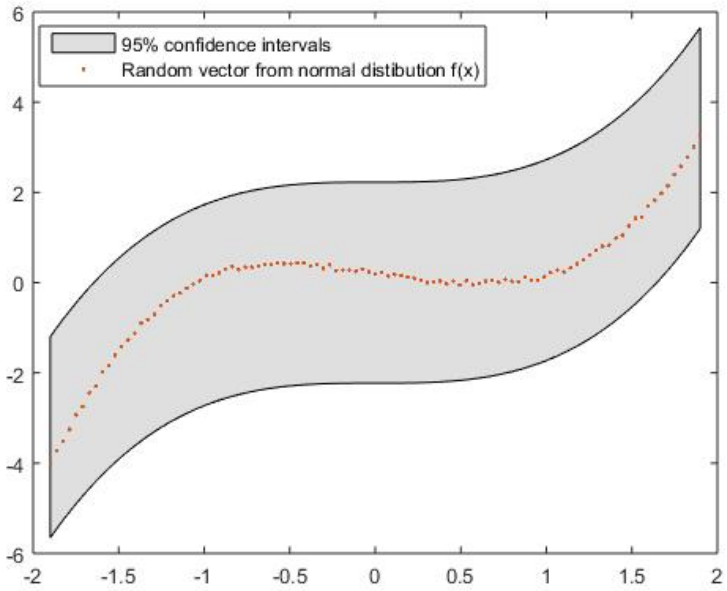


Figure 4.2: A random functions from GP. The shaded grey area shows the 95% confidence intervals. The dots are the values generated from Equation (A.4).

information, we should use mechanisms to choose the mean and covariance functions or in other words, we should *train* the GP model. this process is explained in the next section.

4.3.4 Functions and hyper-parameters selection

As mentioned earlier, various mean functions and covariance functions such as Matérn, Polynomial and squared exponential can be used in GP regression. The best one for the used data can be obtained by model selection methods such as bootstrap criteria, cross-validation criteria and Bayesian methods. After choosing the proper mean value and covariance function, we may choose the appropriate parameters of these functions. hyper-parameters control the mean and covariance functions and pre-setting the hyper-parameters is an important task. In the following example Rasmussen (2004), the assumption is that the mean function is considered to be zero and the covariance function is squared exponential. The aim is to find the best hyper-parameters of this model. A generalisation form of this example is as follows:

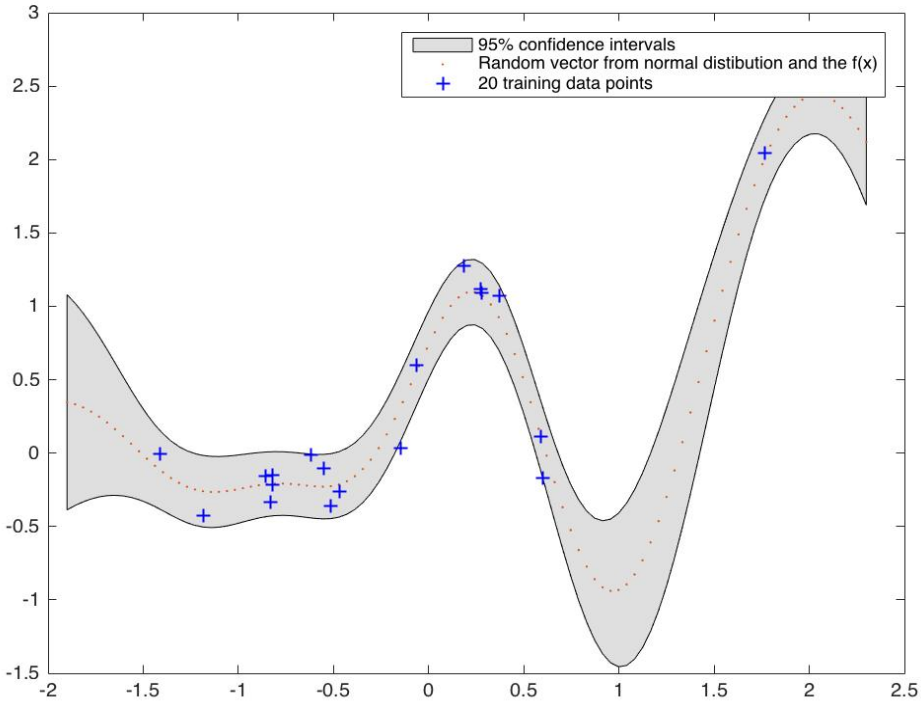


Figure 4.3: A random function from the posterior, given 20 training data points and a noise level of σ_n= log (0.1). Comparing it with Figure 4.2 shows that the uncertainty decreases close to the observations.

$$f \sim \mathcal{GP}(m,k), \text{ where } m(x) = 0 \tag{4.12}$$

and

$$k(\mathbf{x},\mathbf{x}') = \sigma_f^2 exp(-\frac{(\mathbf{x}-\mathbf{x}')^2}{2l^2}) + \sigma_n^2 \delta_{ii'}$$

The hyper-parameters are $\theta = \{\sigma_f^2\ ,\ \sigma_n^2, l^2\}$. Here, l is the characteristic length-scale, σ_f^2 is the signal variance, σ_n^2 is the noise variance; and δ_{ii} is the Kronecker delta. Our aim is to make inferences about the hyper-parameters considering the data. In order to do this, one way is to compute the probability of the data given the hyper-parameters. Fortunately, considering the Gaussian distribution for the data, this calculation would not be difficult:

$$L = \log p(\mathbf{y}|\mathbf{x},\theta) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu) - \frac{n}{2}\log(2\pi). \tag{4.13}$$

Quantity L is called the *log marginal likelihood*. The best hyper-parameter values are obtained by maximising the marginal likelihood. For this purpose, we should set the initial parameters and then the best hyper-parameters suitable for the data are observed. In Equation (4.13), the first term, $-\frac{1}{2}\log|\Sigma|$ is a complexity penalty term, which measures and penalises the complexity of the model. The second term is a negative quadratic, and plays the role of a data-fit measure (it is the only term that depends on the training set output values y). The third term is a log normalisation term, which is independent of the data (Rasmussen (2004)).

Figure 4.4 shows the predictions obtained of a model trained by maximising the marginal likelihood. The best hyper-parameters obtained are l^2=0.06 , σ_f^2 =0.25 , σ_n^2 =0.01. In this particular example, the approach worked pretty well, even before optimising the hyper-parameters of the model (Figure 4.3), but this is not true for all the applications.

4.3.4.1 hyper-parameter optimisation

The previous example is an example of how the hyper-parameters may be chosen. In practice, there are different ways to choose the hyper-parameters. The cross validation

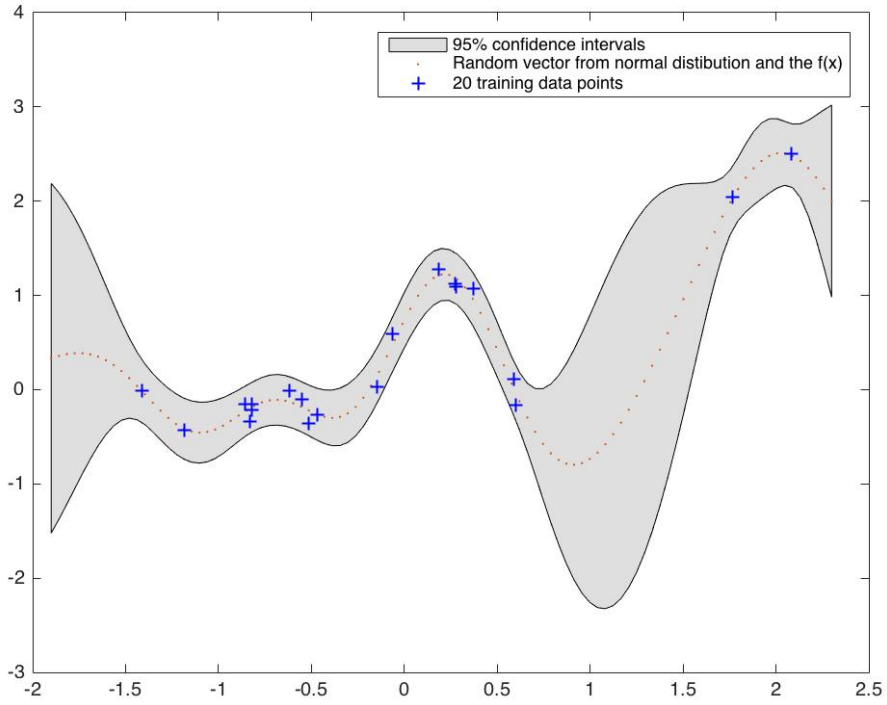


Figure 4.4: Mean and 95% posterior confidence region with parameters learned by maximising marginal likelihood, for the same data as in Figure 4.3.

method is explained in the 'model selection' section in work by Ashrafi et al. (2015) (see Appendix C). Since GP is a probabilistic based model, the optimisation can be performed by *maximising marginal liklehood (MML)* as in the previous example. In this study, various methods are used considering *MML* and they are demonstrated in the following.

- **Conjugate gradient**

Essentially we are trying to find the minimum of a cost function (the cost function is negative log-likelihood) and the first order method is simply to follow the maximum gradient downwards. Usually, however, second order methods are used such as *Conjugate Gradient* (CG) which is an iterative method to solve the linear equation Shewchuk (1994)):

$$A\mathbf{x} = b, \tag{4.14}$$

where $\mathbf{x}$ is an unknown vector, b is a known vector, and $A_{n\times n}$ is a known, square, symmetric ($A = A^T$) and positive-definite ($\mathbf{x}^\top A \int > 0$ for every non-zero $\boldsymbol{x}$) matrix. The unique solution of this equation is x_*. We can show that the solution x_* is also the unique minimiser of the following quadratic function:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - b^\top \mathbf{x}, \tag{4.15}$$

A is a positive definite matrix, therefore, $f(\mathbf{x})$ has a paraboloid bowl shape. At the bottom of the paraboloid bowl the gradient is zero. To minimise the $f(\mathbf{x})$, we can take the derivative (gradient) of $f(\mathbf{x})$, $f'(\mathbf{x})$, as the following:

$$f'(\mathbf{x}) = A\mathbf{x} - b, \tag{4.16}$$

A is positive-definite and symmetric. By setting the gradient of Equation 4.16 to zero, one can obtain a mathematical expression of $\mathbf{x}$ in Equation (4.14).

in practice, we start at an initial random point x_0 and slide down to the bottom of the paraboloid. By taking some steps which at most is equal to n (the size of the matrix A) we are getting closer to the solution until we are satisfied that the error is small enough. The direction of a step in each iteration, should be chosen so that f decreases most quickly, which is the direction opposite $f'(\mathbf{x})$. According to Equation (4.16), this direction

is $-f'(\mathbf{x}) = b - A\mathbf{x}$. For each iteration i, the residual $r_i = b - A\mathbf{x}_i$ indicates how far we are from the correct value of b or on the other hand, the error transformed by A into the same space as b. More details on Conjugate definition can be seen in Appendix A.2.

- **Grid search**

To do this requires a grid search over all the the parameters (hyper-parameters in this study) within specific steps (ideally equal sized steps). To keep this computationally tractable, one can search through limited number of steps. For example if 20 equal steps are chosen in each parameter range and there are three parameters in the model, the total 8000 ($20\times20\times20$) different parameters combination sets are obtained . The best parameters are from a parameters-set that resulted in the best prediction performance for the validation set. The same parameters are then used to predict the test set targets.

- **Random search**

The idea of the random search is taken from the method defined in work by Bergstra and Bengio (2012). They proved using random search over the same domain as used in the grid search can find models that are as good or better as grid search within a small fraction of the computation time. Granting random search the same computational budget, random search finds better models by effectively searching a larger but less promising configuration space.

- **Hyper-prior**

Hierarchical model specification is commonly used to gain a joint regularisation for individual models. The first level are parameters, which could be the parameters in linear or non-linear models. At the second level hyper-parameters θ, control the distribution of the parameters of the first level. Finally, at the top level we may have a (discrete) set of possible model structures which are called hyper-priors ($\mathscr{H}$) and are the prior distributions of the hyper-parameters. The prior over models $\mathscr{H}$ is often taken to be flat, so that we do not favour one model over another (Rasmussen (2006b)). The prior models that we used in this research are Gaussian, Laplacian and non-linear *Smoothbox* prior methods. Univariate smoothed box prior distributions defined with quadratic decay in the log domain and it supports the whole real axis infinitely. It is built by cutting a Gaussian into two parts

and inserting a uniform distribution from a (lower bound parameter) to b (upper bound parameter), the parameters of the *Smoothbox prior*. There is also a parameter η, which balances the probability mass between the constituents so that $\eta/(\eta + 1)$ is used for the box and $1/(\eta + 1)$ for the Gaussian sides. In this research, a and b are considered as the minimum and maximum values of the hyper-parameters ranges. Larger values of η make the distribution more box-like. Prior Smooth Box distribution is given as :

$$\mathscr{H}(\theta) = \frac{1}{w \cdot (\frac{1}{\eta+1})} \cdot \begin{cases} N(\theta \mid a, s^2), & t \leq a \\ 1 & t \in [a,b] \\ N(\theta \mid b, s^2) & b \leq t \end{cases} \tag{4.17}$$

$$w = \mid b - a \mid,\ s = \frac{w}{\eta\sqrt{2\pi}}, \tag{4.18}$$

where a is the lower bound parameter, b is the upper bound parameter, η>0 is the slope parameter, $\theta_{(1*N)}$ contains query hyper-parameters for prior evaluation, and $t \in \theta$ is first initialised and optimised in each step. For sample *a, b* and η values, the hyperprior smooth-box function is plotted, and shown in Figure 4.5 . To generate this plot a=2, b=10, η(slope)=2, and t is considered in a range between 0.0001 and 12.

More information about the priors can be found in Rasmussen and Nickisch (2015).

The mean and variance parameters of the Gaussian and Laplacian priors should be initialised based on the data. These values can be obtained using cross validation in each of the datasets.

- **Evolutionary Algorithm**

These methods involve a subset of evolutionary computation, a genetic population-based meta heuristic which tries to optimise the results in each generation. To evolute the populations, one of the methods of reproduction, mutation, recombination and selection or a mix of them can be used. A fitness function is defined to determine the quality of the solution each time. For instance, the fitness function in this research is the *Negative Log Likelihood function (NLL)* (see section 4.9), in which we want to minimise using different combinations of hyper-parameters. A notable point is that this fitness function results are static and

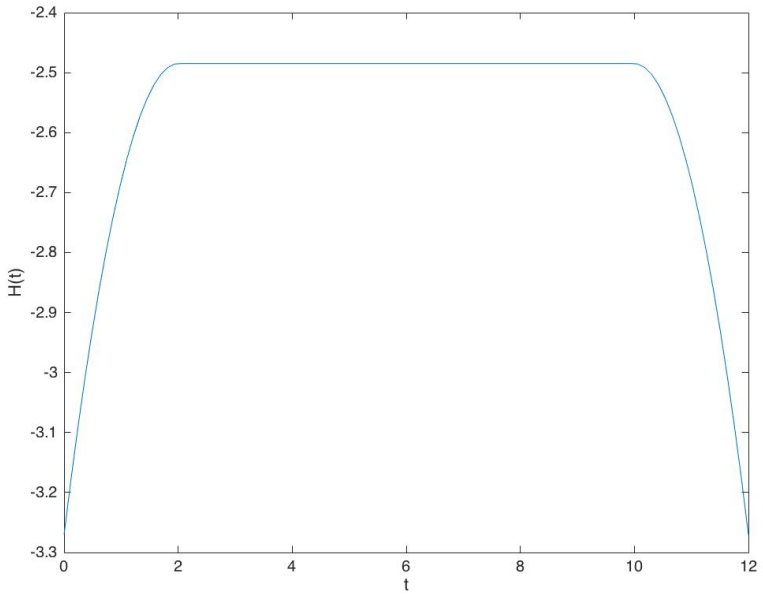


Figure 4.5: Hyperprior smooth-box kernel shape, a (lower bound)=2, b (upper bound)=10, η(slope)=2, $t \in [0.0001, 12]$

only depend on the current information. In order to improve the fitness function to vary dynamically based on current and previous result states, a heuristic crossover function is used. In this method, the population in the first generation is initialised randomly. Next generation's population can be obtained as follows:

- To make the reproduction possible, two parents from the previous generation resulted in the best performance (according to the fitness function) are selected. Children of parents can be obtained using crossover function (recombination function). We can use various crossover functions; the one used for this study is heuristic function. If parent1 has a better fitness value than parent2, the function returns the child as follows (Houck et al. (1995)):

$$child = parent2 + ratio * (parent1 - parent2) \tag{4.19}$$

  The obtained child is closer to the parent with better fitness value and has more distant from the other parent with worse fitness value. To specify the amount by which the child is far from each parent, a ratio parameter is defined. The value of ratio can be set between 0 to 1.2 and it should be chosen based on the data to accelerate the convergence. This process continues till the required number of population in the current generation is achieved.

- In each generation a number of children are obtained using mutation. In this case, the variables (hyper-parameters) are chosen randomly from the variable ranges (minimum ta maximum defined ranges for hyper-parameters) from a gaussian or uniform distribution. Performing mutation, population of a generation are not biased only on the parents and can be chosen from the entire variables space.

- To keep the best results of the last generations, we can add a fixed number of best children from last generation to the next generation. Adding *Elite children,* it is guaranteed that the performance of the model is not fallen over generations.

As an example, if the Population size is 20, the Elite count is 2, and the Crossover fraction is 0.8, the numbers of each type of children in the next generation are as follows:

There are two elite children which are brought directly from the previous generation. From the 18 individuals other than elite children, there are 14 crossover children (which is obtained by 0.8*18 = 14.4 , the algorithms then rounds it to 14) . The remaining four individuals, other than elite children, are mutation children which are chosen randomly from the entire variables space (Mitchell (1998); Winter et al. (1996)). Genetic Algorithm (GA) Matlab optimisation toolbox by Houck et al. (1995) is used to do the Evolutionary Algorithm hyper-parameter optimisation.

A simple example of evolutionary algorithm

To understand the process, a simple example is given with the fitness function: $f(x) = 100 \times (x_1^2 - x_2)^2 + (1 - x_3)^2$ that we aim to minimise. We consider 4 initial populations for the first generation that initiated randomly for three parameters of the model.

Initial random population=

	$x_{1,2,3}$ (Parameters values)			Fitness values
Populations	1.86	4.47	4.97	117.85
	3.40	2.79	4.14	7701.15
	1.87	4.49	4.03	107.81
	2.67	3.57	3.11	1271.03

We choose two best scores (lowest) as parents for the second generation. Here, the first and third populations have the lowest scores. The first child from these two parents is obtained using Equation 4.19 and if we consider ratio to be 0.7 to be more similar to the better parent (third one with lower score), the first child of the next generation is obtained as Ch1=[1.86 4.48 4.31] with fitness value of 110.63. The obtained score is not better than one of the parents. This child is stored and using the parents and the obtained child, we will continue the same process to obtain the other children of the second generation. As mentioned previously, to make sure we do not lose the best scores, in each generation we may choose to keep the desired number of best populations from the last generation (elite children). In addition, a number of children may be chosen by mutation (randomly, from a gaussian or uniform distribution), to prevent the population of a generation only biased on the parents. The process ends when we get to the termination condition, which in the experiment in this thesis, is the number of generations.

4.4 Single Layer Network

A Single Layer Network (SLN) is used to consider whether the relationship between the data features and the target values are linear. An SLN can be considered as a simple generalised linear model. SLNs are known as a statistical technique for linear regression. These models have a linear combination of the input features, the coefficients are the parameters of the model. They also include an activation function tailored to the data being modelled.

If we denote the input values to the network by x_i where $i = 1, ..., d$ then the network will consist of c linear combinations of these inputs (c is the number of outputs). This gives us a set of intermediate variables a_j:

$$a_j = \sum_{i=1}^{d} w_{ji} x_i + b_j \quad j = 1, ..., c. \tag{4.20}$$

Each output unit has one variable a_j. w_{ji} defines the weight matrix and b_j are the bias parameters. To obtain the output values y_j, the a_j variables should be transformed by the activation functions of the output layer. For the regression case we use the linear function of the form:

$$y_j = a_j \tag{4.21}$$

We can also obtain the error function for the output values (Nabney (2002)).

4.5 K-nearest-neighbour

K-nearest-neighbours (KNN) is a simple algorithm that stores all the training and test features and their corresponding targets, then predicts the numerical target based on a similarity measure between the training and the test cases. It is a non-parametric method. The simplest KNN regression algorithm calculates the average of the numerical target of the K nearest neighbours. Another approach uses an inverse distance weighted average of the K nearest neighbours.

In case of numerical variables distance algorithms such as Euclidean, Manhattan and Minkowski can be used. But when it comes to the categorical variables hamming distance

could be used, which calculates the number of instances in which corresponding symbols are different in two strings of equal length.

Choosing the best K depends on the nature of the data. In general, we can trust more on a large K value predictions as it reduces the overall noise; however, it may cause to blurred distinct boundaries within the feature space. Another way to do this, is to use cross-validation which determines a good way using an independent dataset to validate the K value. The default and optimal K value is usually chosen from odd numbers (Hastie et al. (2001)). As an example, the simplest Euclidean distance between one dimension p and q is shown by the below equation:

$$d(p,q) = ||p-q|| \tag{4.22}$$

4.6 Support Vector Machine Regression

The $\varepsilon-$SVM method has first introduced by Vapnik (1998). This method uses the maximum margin algorithm which is a non-linear function, learned by linear learning machine mapping into high dimensional kernel induced feature space. The parameters involved in the SVM do not depend on the dimensionality of feature space. Similar to the same as classification, in regression problems, the generalisation bounds should also be optimise. To do this, the loss function should be defined which is also called 'epsilon intensive' loss function. Our goal is to find a function $f(x)$ that has at most ε deviation from the actually obtained target points for all the training data, and at the same time is as flat as possible (derivatives should vanish at those points). Here, we do not care about the errors as long as they are less than ε, but any deviation larger than this is not acceptable (Smola and Schölkopf (2004)). In this method, the input x is first mapped onto a m-dimensional feature space using some fixed (nonlinear) mapping, and then a linear model is constructed in this feature space. The linear $f(x)$ function is given by the below mathematical notation:

$$f(\mathbf{x},\mathbf{w}) = \sum_{j=1}^{m} w_j g_j(\mathbf{x}) + b, \tag{4.23}$$

where $g_j(\mathbf{x})$, $j = 1,..,m$ defines a set of nonlinear transformations, and b is the 'bias' term. As we usually consider the data to have zero mean, the bias term can be dropped.

The quality of this estimation is measured by the loss function $L(y, f(x,w))$ which is called ε-insensitive loss function proposed by Vapnik (1998):

$$L_\varepsilon(y, f(x,w)) = \begin{cases} 0 & if\, y - f(\mathbf{x},\mathbf{w}) \leq \varepsilon \\ |y - f(\mathbf{x},\mathbf{w})| - \varepsilon & Otherwise \end{cases} \tag{4.24}$$

In order to reduce the complexity of the model, or keeping the function as flat as possible we can minimise the $||w||^2$. To do this, one should introduce the non-negative slack variables ζ_i, ζ_i^*, $i = 1,..n$ to measure the deviation of the training samples outside ε-insensitive zone. Therefore, the SVM regression is defined to minimise the following function:

$$Minimise\ \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*) \tag{4.25}$$

$$Subject\,to \begin{cases} y_i - f(x_i,w) \leq \varepsilon + \zeta_i^* \\ f(x_i,w) - y_i \leq \varepsilon + \zeta_i \\ \zeta_i, \zeta_i^* \geq 0,\ i = 1,..n \end{cases}$$

the constant $C>0$ defines the trade off between the flatness of f (ideally to minimise the w) and the maximum amount that the deviation larger than ε can be tolerated. Figure 4.6 shows the situation graphically. Here the derivations are penalised in a linear fashion and that is why only the points outside the shaded region causes the cost.

To solve this optimisation problem easier, we can transform it into the Lagrange dual problem as the below:

$$f(x) = \sum_{j=1}^{n_{sv}}(\alpha_i - \alpha_i^*)K(x_i,x), 0 \leq \alpha_i \leq C, 0 \leq \alpha_i^* \leq C, \tag{4.26}$$

where n_{sv} is the number of support vectors (SVs) and the kernel function:

$$K(x,x_i) = \sum_{j=1}^{m} g_j(x)g_j(x_i) \tag{4.27}$$

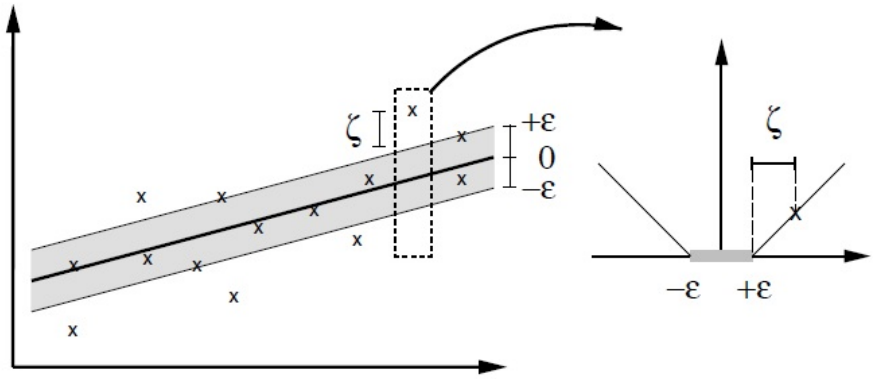


Figure 4.6: The soft margin loss setting for a linear SVM (Smola and Schölkopf (2004))

The SVM prediction performance (estimation accuracy) depends on the parameters C, ε and the kernel function (with its parameters). Selecting the best model depends on the distribution of the input values of the training data and the application the SVM regression is used. Both of parameters C and ε affect the model complexity in different ways. Parameter C controls the model flatness and the maximum threshold to which the derivations can be tolerated. Parameter ε determines the width of the ε-intensive zone that used to fit the training data. The number of support vectors are affected by the ε value. The bigger ε values cause more flat estimates. The kernel functions for SVM regression are varied from linear to non-linear for algorithms which can be expressed in terms of dot products. Following kernel functions can be used for the SVM regression:

$$K(X_i,X_j) = \begin{cases} X_i \cdot X_j & linear \\ (\gamma X_i \cdot X_j + C)^d & Polynomial \\ \exp(-\gamma |X_i - X_j|^2) & RBF \\ tanh(\gamma X_i \cdot X_j + C) & Sigmoid \end{cases},$$

where $K(X_i,X_j)$ is the kernel function and it is a dot product of input data points mapped into the higher dimensional feature space by transformation. Gamma (γ) is an adjustable parameter of certain kernel functions (Cherkassky and Mulier (2007); Cherkassky and Ma (2002); Chapelle and Vapnik (1999); Smola and Schölkopf (2004)).

By far, the RBF kernel has been proved to be the most popular choice of kernel types used in SVM . This is due to the fact that unlike the other three kernel functions, the variable of RBF kernel function can be considered as the Euclidean distance between two points $||X_i - X_j||$. However linear, polynomial and sigmoid kernel functions are based on the inner products of the vectors $((X_i \cdot X_j)$) (Pedrycz and Chen (2013)).

4.7 Growing Neural Gas (GNG) for clustering

It is important from pharmaceutical point of view to see if the data can form clusters that are classified based on specific physiochemical or experimental conditions. If we train the model based on the data obtained in the new clusters and perform the predictions for the unseen data, we may obtain better results. To do so, the Growing Neural Gas (GNG) by Fritzke et al. (1995) is applied to cluster the complete dataset. GNG is an unsupervised clustering algorithm which makes use of unlabelled data. It does not need to have any *a priori* knowledge about the number of clusters. Clustering is performed based on the similarity of the compounds' features, which means similar compounds are clustered together. This algorithm creates a graph structure of the data. It then generates sub-graphs in which the number of sub-graphs shows the number of clusters. Graph structure and sub-graphs are generated based on the following steps Fritzke et al. (1995) :

1. Two nodes a and b should be placed randomly in multi-dimensional space.
2. An input signal ξ from a probability density function $(P(\xi))$ should be generated.
3. The nearest node should be recognised and, w is assigned to this node as the winner and the second nearest node is identified as s, the second winner.
4. All the edges emanating from w should be incremented and be kept as age.
5. The distance between the input signal ξ and w should be added to a variable, *error*.
6. The node w and its immediate neighbours should be moved towards ξ by multiplying the error to the constant values α and β, respectively.

7. If w and s are connected by an edge, the age should be reset to zero. If they are not connected by an edge, an edge should be created and initialised to zero.

8. All edges with greater age than a_{max} should be removed. All nodes with no emanating edges should be removed.

9. If the number of generated input signals is an integer multiple of the parameter λ, then a new node should be inserted based on the following rules:

    - The node q with the greatest error value should be selected. A new node r should be inserted between the nodes q and its neighbour f which have the largest error value.
    - The original edge that connecting q and f should be removed and edges should be inserted between the new node r and the nodes q and f.
    - The error variables of q and f should be decreased by a quantity called δ. The error variable of r should be initialised by the new error value which was obtained for the node q.

10. All error variables should be decreased by constant d.

11. Steps 1-10 should be repeated until the maximum numbers of nodes are inserted.

4.8 Monte Carlo method

The Monte Carlo method is an application of the probability and statistics to the natural sciences. It is mainly used in statistical analysis. In this algorithm, various distributions of random numbers are generated. Each distribution shows a particular process of the total processes that approximates the real data values (Anderson (1986)). Monte Carlo techniques can be used for optimisation, sampling and estimation purposes. In general this method involves random sampling from certain probability distributions (Kroese et al. (2014)) . The following steps should be taken to generate the random distributions:

- A domain of possible input data should be defined.

- Inputs are randomly generated from a probability distribution over the domain.
- Computational algorithms should be performed on the generated inputs.
- Based on the application the Monte Carlo algorithm is used, the results should be aggregated.

4.9 Performance measures

The *correlation coefficient (corrcoef), Negative Log Likelihood* (NLL) *and Improvement Over Naïve (ION*) are used to measure the model performance (Sun et al. (2012)). Probably the simplest prediction (*Naïve*) always predicts the mean of the target value in the training set independently of the input.

If we produce a predictive distribution at each test input x_*, in the dataset $\mathscr{D}$, the negative log probability of the target under the model can be evaluated. If we consider μ as the mean prediction, as GPR produces a Gaussian predictive density, one obtains (Rasmussen (2006b)):

$$NLL = -\log p(y_*|\mathscr{D}, x_*) = \frac{1}{2} log(2\pi\sigma_*^2) + \frac{(y_* - \mu)^2}{2\sigma_*^2}, \tag{4.28}$$

where the predictive variance σ_*^2 for GPR is computed as $\sigma_*^2 = \mathrm{V}(f_*) + \sigma_n^2$. $\mathrm{V}(f_*)$ is the variance of the prediction. As we are predicting the noisy target y_*, we must add the noise variance σ_n^2 . To standardise this loss, we may subtract the loss from the obtained NLL using the mentioned equation considering mean and variance of the training data. We denote this the Standardised Log Loss (SLL). The mean SLL is denoted MSLL (Rasmussen (2006b)). The MSLL will be approximately zero for simple methods and negative for better methods.

In contrast, *ION* measures how much better a predictor is than the *Naïve* predictor. *ION* is given by the below equation:

$$ION = \frac{(MSE_{naïve} - MSE_{GP})}{MSE_{naïve}}, \tag{4.29}$$

where *MSE* denotes the *Mean Squared Error* and may varies between 0 and $+\infty$. *ION* values may range from $-\infty$ to 1. Large positive *ION* values represent better performance, while smaller positive MSE values represent better model with low error.

Correlation coefficient (*CorrCoef*) finds the statistical relationships between the targets of the test set and predicted values for them, and it may vary from -1 to 1. In this study, larger positive correlations defines good prediction performance (Sun et al. (2012)).

Chapter 5

Data Visualisation

Visualising the datasets can reveal more details about the nature of the data and the relationship between features and the data distribution. The datasets employed in this study contain more than two numerical features. We need to use techniques to reduce the dimensionality of the data. One way to do this is to apply Principal Component Analysis (PCA).

5.1 Principal Component Analysis (PCA)

5.1.1 PCA for numerical data

PCA is a well known technique for analysing numerical data. The Principal Components (PC) of our data manifold, are a set of orthogonal vectors that progressively account for the variance in the data. Figure 5.1 shows PC1 against PC2 for a sample data. Note that if the data manifold is n-dimensional, then there will be exactly n principal components.

We can use as many principal components as we need to have an insight into the data. We can consider the PCA as an n-dimensional ellipsoid that we want to fit to the data, each of whose axis defines a principal component. A smaller unit value of the axis, represents the small variance along that axis, and if we remove that axis and its corresponding principal component from the data set presentation, we only miss a small amount of information. The axes of the ellipse are obtained using the following process (Jolliffe (2002); Shlens (2014)):

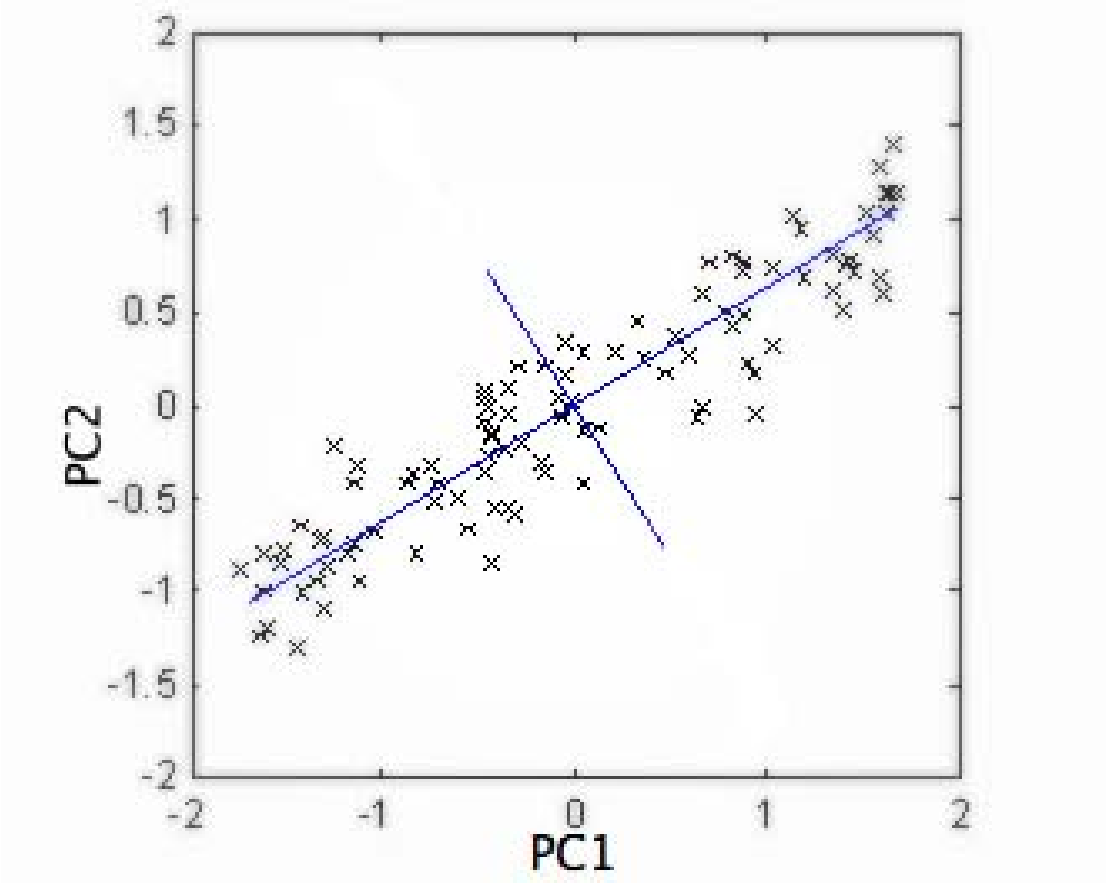


Figure 5.1: PC1-PC2 for a sample dataset

- First, for the d-dimensional data, we should calculate the *d-dimensional* mean vector of the data features. Then to normalise the data, the mean data should be subtracted from the dataset. We can also use *Z-score* method to normalise the data.

- Then, we should compute the covariance matrix of the data (or correlation matrix if *Z-score* is applied first), and calculate the eigenvectors $(\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_d})$ and corresponding eigenvalues $(\lambda_1, \lambda_2, ..., \lambda_d)$.

- The eigenvectors should be sorted based on the decreasing eigenvalues rank. Then the k eigenvectors with the largest eigenvalues can be chosen and $d \times k$ dimensional matrix $\mathbf{W}$ *(each column of* $\mathbf{W}$ defines an eigenvector).

- Each of the eigenvectors can be considered as an axis of the ellipsoid that fitted to the data.

- The amount at which each eigenvector variance represents, can be calculated by dividing its corresponding eigenvalue by sum of all the eigenvalues.

The next step is to project the principal components to the real data, so that we can visualise the projected data and be able to analyse the data. To do this, we can multiply the d-dimensional data to the eigenvectors $\mathbf{W}$, that we selected as the most important ones (eigenvectors with the largest eigenvalues) and obtain the projected data $\mathbf{P}$, considering only the important features. This projected data can be plotted and analysed. One of the main uses of PCA is the dimensionality reduction. For example, if the first two principal components stand for 90% of the variance of the data, the plotting of the data manifold in the PC1-PC2 space will give a very precise two dimensional representation of the data manifold.

Figure 5.2 represents eigenvectors corresponding variance of the data. From this figure we can see that the fifth and sixth PCs variance is close to zero. So, if we remove them, we do not miss much information.

To see the relationship between the most important principal components of dataset human D (see Chapter 3 for dataset details), PC1-PC2 is plotted and can be seen in Figure 5.3. In this figure, the first and second principal components account for 50% and 26%

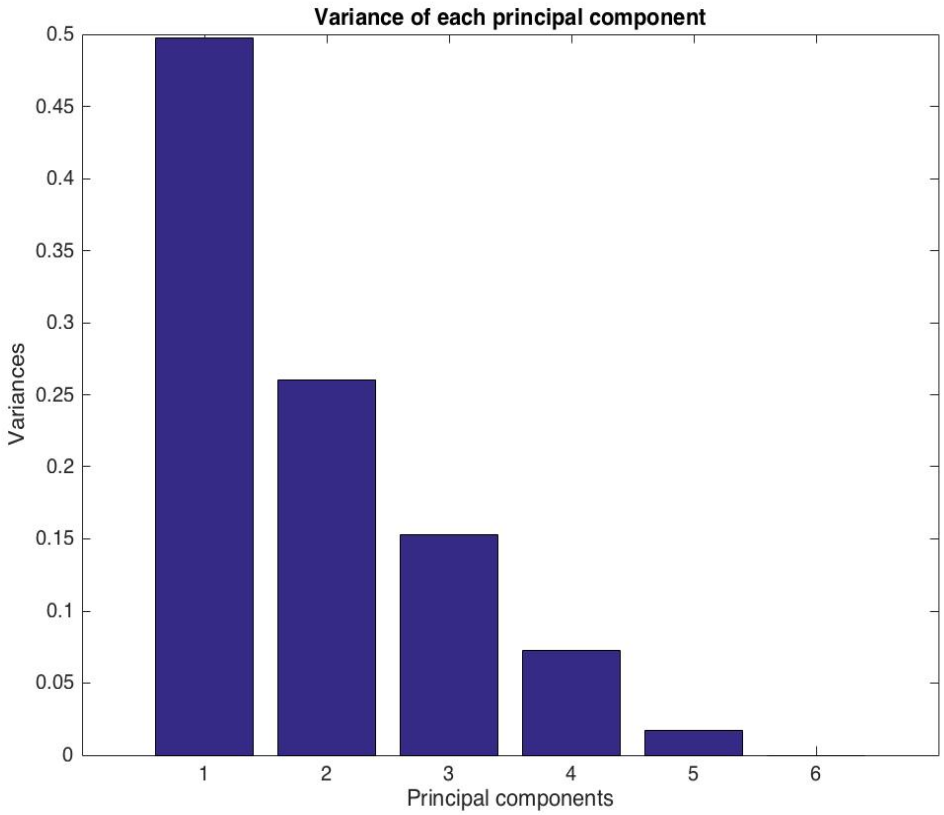


Figure 5.2: Variance of the Principal components

variances of the data, respectively. From this figure, we can see the area that data points are gathered together and a few outliers can also be seen. These outliers are explained later.

5.1.2 PCA for nominal data

In addition to the numerical data that is mentioned in the previous section, we also need to visualise the nominal data which are items that differentiated by a naming system. Numbers may be assigned to nominal data, to simplify capturing and referencing, but it does not mean they can be considered as numerical or ordinal data. An example of a nominal data can be the countries people were born or different colours of a particular object. Nominal items are usually categorical. To visualise the nominal data, initially we need to find a

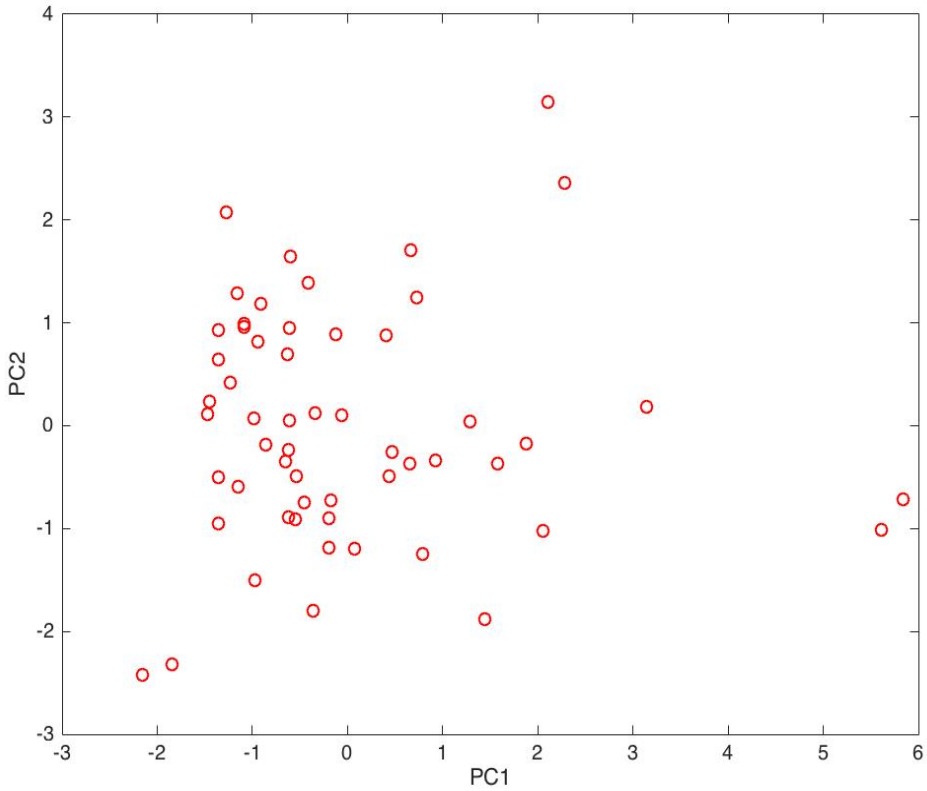


Figure 5.3: The PCA plot of human dataset D (numerical data)

way to calculate the covariance/correlation matrix of the features. Three various methods are used to plot the PCA for dataset human D nominal data and they are illustrated in the following sections.

5.1.2.1 Using correlation matrix between the nominal features

The first method byWakelam et al. (2016), finds the correlation between the nominal features. As an instance, if we have three objects with two dimensions or features, as:

	data features/dimensions (d)	
Data objects (D)		
A=	a	b
	a	b
	c	b

,

the correlation matrix $\mathbf{K}_{(2\times2)}$ is symmetric with 1 in the main diagonal (correlation between each feature by itself). The other values of the matrix can be calculated based on the following instruction:

- First, we should consider all the ways data objects can be compared together. In our example we can compare D1 with D2, D2 with D3 and D1 with D3. Number of comparisons is stored in *comp* variable, which is 3 in this example, and the comparison pairs P= (D1 D2,D1 D3,D2 D3).
- To find the values of the matrix **K**, we will compare the features of all pair of objects in P. The '*count*' always starts at 0.
  - If values of the first feature d1, are the same, then we look at the second feature d2. If both of their d2 are also the same we will add 1, if not we deduct 1 from *count*.
  - If values of the first feature d1, are not the same, we look at the second feature d2. If both of their d2 are also not the same we will add 1, if they are equal, we will deduct 1 from count.
  - This process is continued till we add/deduct count for all the P pairs objects.

- The obtained values from the previous steps then are divided by *comp* and we locate them in the covariance matrix based on their relation.

For example, the matrix **K** for the data A is **K**=$\begin{pmatrix} 1 & -0.33 \\ -0.33 & 1 \end{pmatrix}$ which in this case, -0.33 shows there is not a strong correlation between features 1 and 2 of the data. The reason is that although we have 2 data points with the same features, the last data point has the same second feature but not the same first feature as the other two, and it therefore made the correlation as weak as -0.33. This highlights the issue of dealing with variable data of a biological origin.

Having the covariance/correlation matrix for nominal data can perform the same process as numerical data to obtain the principal components of the human dataset D. To see the PCA plot for the same dataset in the previous section, we plotted the nominal data (we have only 3 features in this dataset) PC1 against PC2 in Figure 5.4. PC1 and PC2 represent %58 and 42% of the data features, respectively which cover all the data variance. The figure shows a linear relationship between the PC1 and PC2 with some outlier points.

The method described above on computing correlation on nominal features has advantage to be learned and used easily, as it does not involve complicated mathematical calculations. In addition, it does not take a long time to be executed. On the other hand, its weak point is that it is not appropriate to be used for ordinal and interval variables.

Relationship between the numerical and nominal data: To investigate the probable relationship between the chemical features (numerical data) and of experimental conditions (nominal features) on the predictions, their PCA plots are compared together. The first thing to be noticed from Figure 5.3 is the four outliers in the left and right side of the figure. These points are shown in green in Figure 5.5(a). The same data points' projections in categorical data are also shown in plot (b) of the same figure with red stars. One can see that 4 outliers of PC1 and PC2 projected numerical data are 3 points in the categorical PC data projection (there is one repetition). At least 2 of these points seem to be outliers in Figure 5.5 (b) as well.

To do more analysis, the linear relationship formed in the nominal data PC plot is also investigated to see whether the same data points in the numerical data PC plot represent

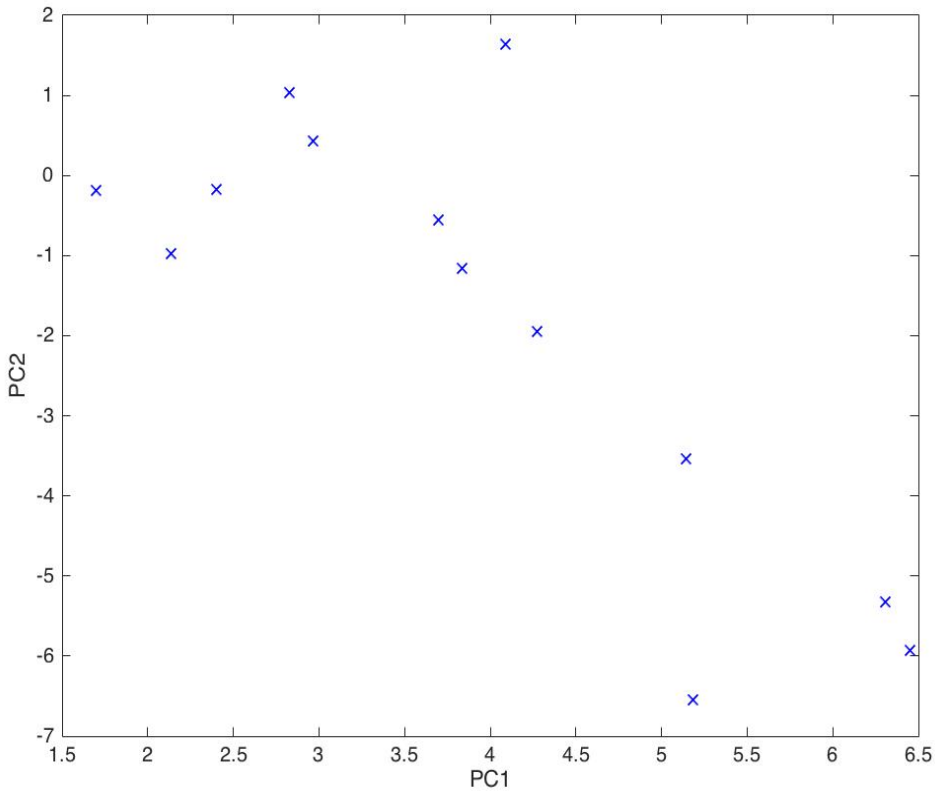
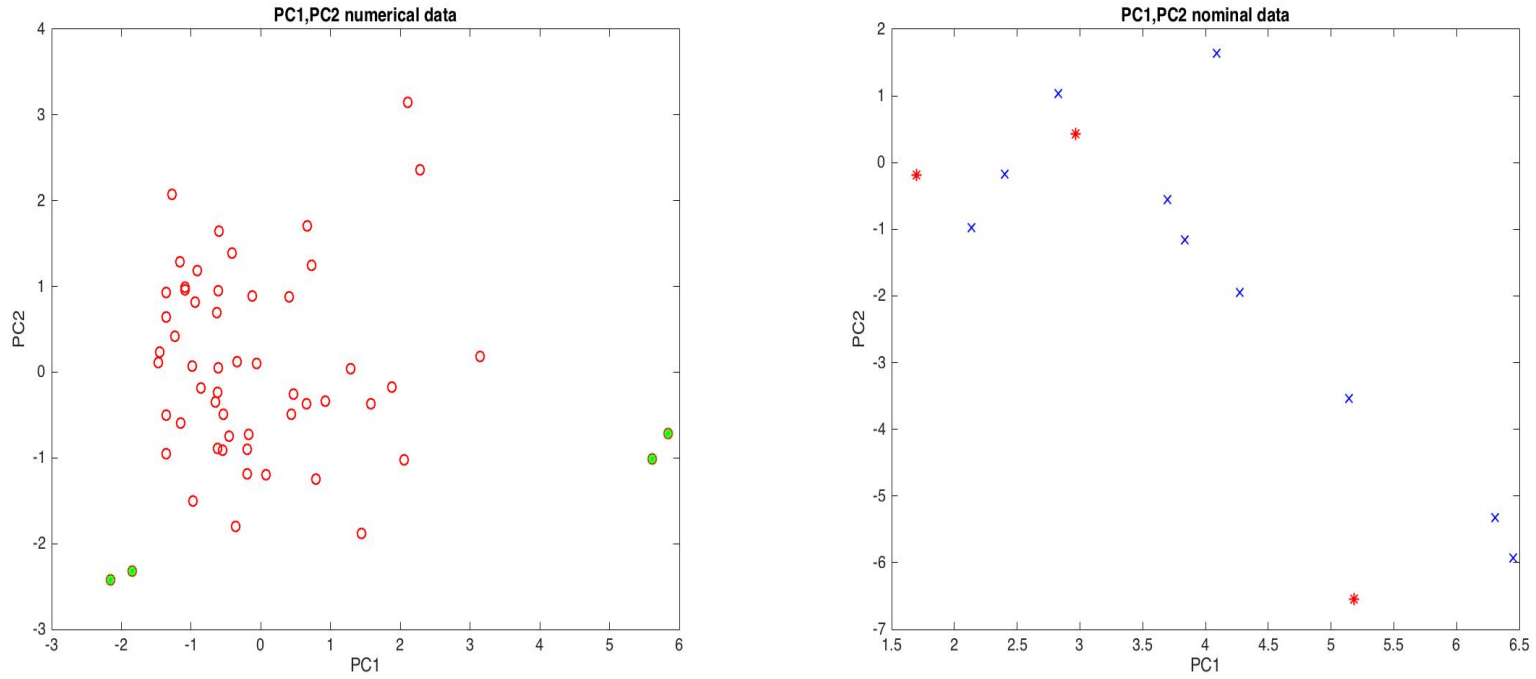


Figure 5.4: The PCA plot of human dataset D (nominal data), using the method byWakelam et al. (2016)



(a) (b)

Figure 5.5: The PCA plot of human dataset D (a) numerical data, outliers are shown in green (b) nominal data, outliers are shown in red stars

any specific relationship between these two. Figure 5.6 (a) and (b) shows this scenario. The red stars in plot (a) represent the liner relationship in the nominal data and the same points are shown in plot (b) by green. As it is expected, we could not find any specific relationship between the numerical and nominal data. To examine the other methods of PCA visualisation for nominal data, two other methods are discussed in the following sections.

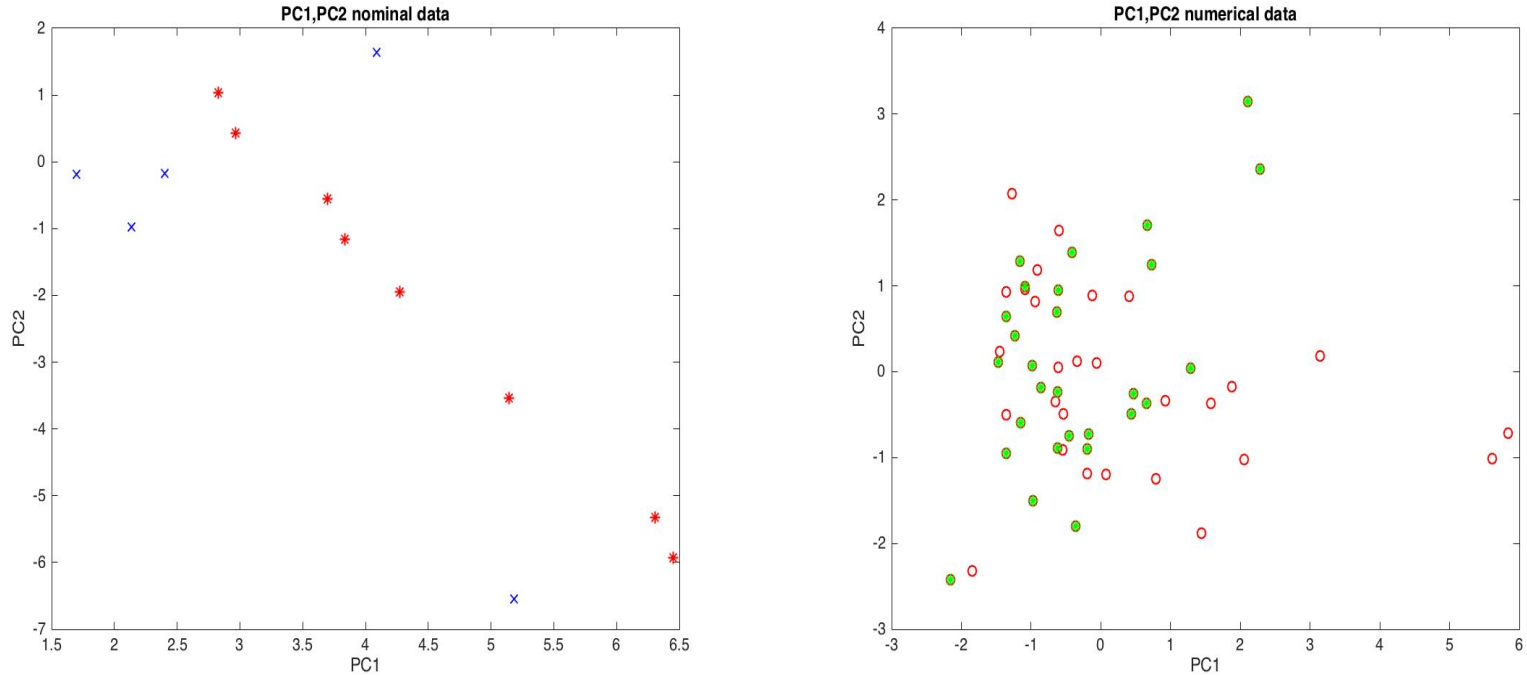
5.1.2.2 Using Hamming distance covariance function

In chapter 4 of this thesis, a kernel function is introduced by Couto (2005) which is calculated based on the Hamming distance. In this section, using this technique, the covariance matrix between the nominal features are calculated and plotted in Figure 5.7. PC1 and PC2 represent %65 and 20% of the data features, respectively which covers 85% the data variance. This figure is different from Figure 5.4, but it also shows a linear relationship between the PC1 and PC2 with some outlier points. To investigate the linear relationship in this figure and their corresponding data in the numerical PCA plot, Figure 5.8 is generated. Similar to the result in the previous section and as expected, the green points in plot (a) of this figure do not yield a specific relationship with the green points (the corresponding points in numerical data) in the same figure plot (b).

To visualise the nominal data, another method based on multiple correspondence analysis (MCA) is used (Linting and van der Kooij (2012)), and the PCA plot is shown in Figure A.1. The details of applying this method can be found in section A.3. Similar to the previous techniques used for nominal data visualisation, no correlation can be found between the numerical and nominal data.

5.2 Conclusion

In this chapter, I have shown how PCA can be used to analyse both numerical and nominal data. As expected, we could not find any correlation between numerical and nominal data. Although the number of nominal features is small in the studied datasets, the possibility of using different methods to set up the covariance/correlation matrix for nominal data, was



(a) (b)

Figure 5.6: The PCA plot of human dataset D (a) categorical data, linear data shown in red (b) Numerical data, the green points represent the same points

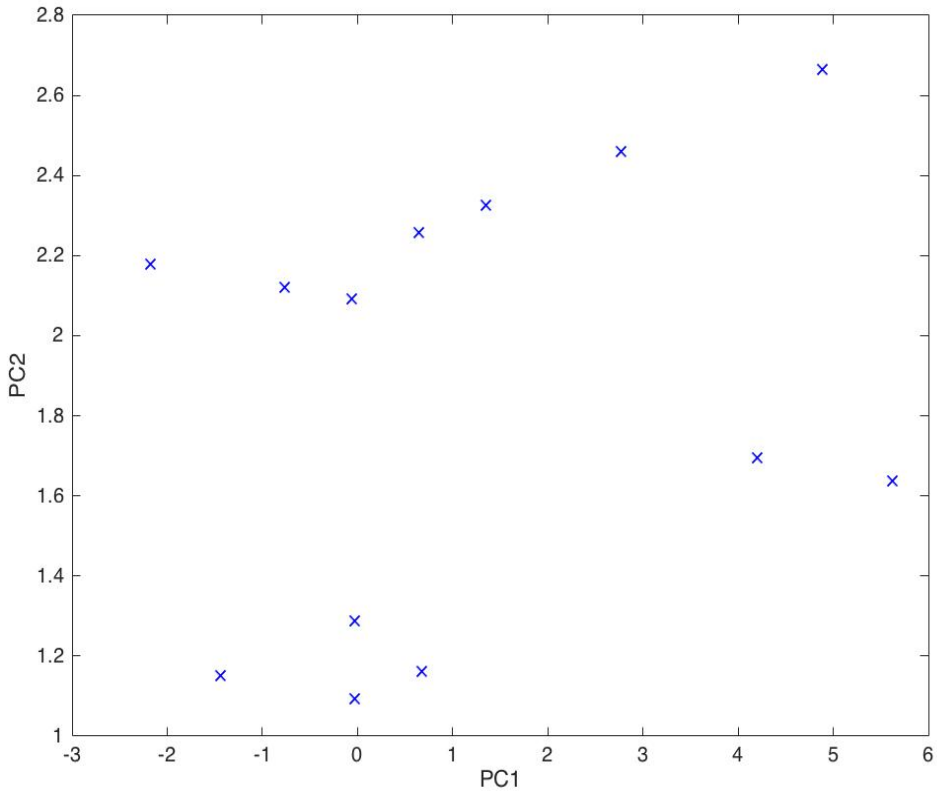
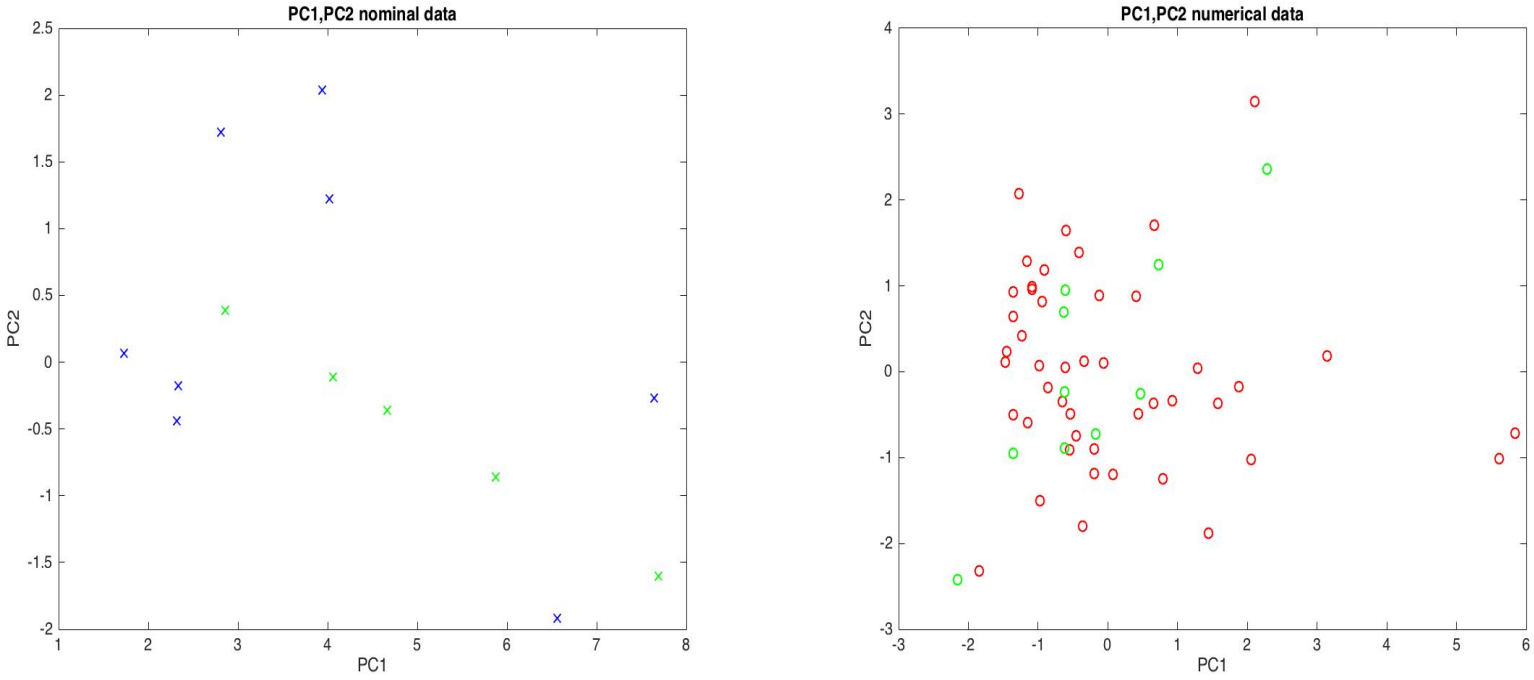


Figure 5.7: The PCA plot of human dataset D (nominal data), using Hamming distance method Couto (2005)



(a) (b)

Figure 5.8: The PCA plot of human dataset D (a) categorical data, linear data shown in green (b) Numerical data, the green points represent the same points. Hamming distance method is used to obtain the feature-feature covariance matrix.

investigated. This can be useful when dealing with a dataset involving a large number of nominal features.

Chapter 6

Experimental Results

In this Chapter, the major experiments that have been performed throughout this research are defined and the results are discussed. Various Machine Learning (ML) methods are applied to the human and animal datasets described in Chapter 3. It should be noted that in all of experiments, the training data is normalised as *Z-score*. The test set is also normalised with the training set mean and standard deviation.

These experiments include:

- Experiment 1: Various regression methods, kernels and hyper-parameters settings are applied to human datasets with 7 and 5 numerical features (Section 6.1).
- Experiment 2: The Gaussian Process and SVM methods are used to examine the effect of using human and non-human models to predict the performance of each other (Section 6.2).
- Experiment 3:The efficiency of considering experimental conditions on permeability predictions is investigated (Section 6.3).
- Experiment 4: A kernel containing a mix of of numerical and nominal data is used in the GP regression model (Section 6.4).
- Experiment 5: A clustering method, Growing Neural Gas, is applied to the complete human dataset. The prediction performances within their original sets and the new GNG clusters are then compared (Section 6.5).

- Experiment 6: The last experiment is using Monte Carlo method to deal with inconsistent data (Section 6.6).

6.1 Experiment 1: on Human Data (Applying GP, SLN, QSAR, KNN and SVM methods)

The main aim of this section is to investigate the chemical descriptors, different regression methods, their performances and the suitable parameters (and hyper-parameters) to be used for the defined datasets in this study.

First of all, I examine the effect of using all the descriptors compared with using some of the descriptors on the performance. GP is applied (see section 4.3.1 of Chapter 4) to the six human datasets considering 7 and 5 features, separately. The 5 features used in the experiments are widely used descriptors in skin permeability laboratory experiments ($MW, SP, logP, HA$ and HD). As mentioned in Section 3.2.1, the reason these particular features are selected is missing or zero values that are assigned to a number of chemicals for their *MR* and *MPt* features. This might cause issues with predictions. All covariance functions's hyper-parameters are optimised using conjugate gradient method.

Furthermore, different covariance functions have been applied in order to find the best performance of the GP. Polynomial, Squared Exponential and *Matérn* covariance functions using *leave-one-out* method and different parameters setting are employed and their performance are compared together.

Linear methods including the traditional QSAR technique that is commonly used in physical and biological sciences areas, including drug delivery and environmental sciences, and the SLN algorithm (see Chapter 4 section 4.4) are used to see whether linear equations can be fitted well to the data to estimate the permeability of the unseen data. The datasets are quite small and it may raise the over-fitting and under-fitting problems. Therefore, I use KNN (see section 4.5 of Chapter 4) that is known to cope with small datasets. Due to the problems of data similarity and inconsistency with GP (see 6.1.1), the SVM regression method (see Chapter 4, section 4.6) has also been applied to the datasets. The *Leave-one-out* technique is used in all the experiments of this section.

6.1.1 Gaussian Process

The experiments in this section are performed using GP as described in section 4.3 (Rasmussen and Nickisch (2015)). The nature of the data in datasets illustrates that there are many similar or the same feature vectors in each of the datasets with different target values corresponding to them that cause problem in GP predictions. To perform the regression in the GP, the *Cholesky* Decomposition[1] of the covariance matrix of training vectors (**K**) should be calculated and and it should be positive definite, otherwise it generates an error. Similar or same vectors tend to result in not positive definite covariance matrix. In order to omit this effect on the predictions, the same feature vectors (of chemical descriptors) are removed and only one of them is kept. The target value is obtained by averaging values of the target (e.g. $\log K_p$) for those with same features.

In addition, using a threshold, the very similar vectors are also removed and one of them is kept using the average value of their targets. The threshold of 0.01 is considered between the features vectors in training set. If the distance between the points become less than the threshold, then GP does not work well on the data. To solve this issue, the similar points should be replaced with only one point with the same features and target that is the average of different target values. The similarity is defined as the following:

$$Distance = \frac{d_{i,j}}{max(|f_i|,|f_j|)} \leq 0.01, \tag{6.1}$$

where $d_{i,j}$ is the Euclidean distance between two vectors *i* and *j*, and $|f_i|, |f_j|$ are the magnitudes of vectors *i* and *j* respectively. As an example if we have two vectors with three features each, as *i*=[1.5 5 3.5] and *j*=[2 5.5 3], the similarity between them is calculated as the following:

$d_{i,j} = 0.87$, $|f_i| = 6.28, |f_j| = 6.58$

$Distance = \frac{0.87}{6.58} = 0.13$

In this example the distance is more than the threshold and we can keep both points without causing any problem. As another example, if *i* is the same vector and *j*=[1.5 5 3.4], then the distance between *i* and *j* is calculated as following:

[1] In linear algebra the *Cholesky* Decomposition is the product of a lower triangular matrix and its conjugate transpose.

$d_{i,j} = 0.10$, $|f_i| = 6.28, |f_j| = 6.23$

$Distance = \frac{0.10}{6.28} = 0.01$

In this case, as the distance is 0.01, we should remove one of the two vectors randomly. As in the skin datasets, there is a target value (e.g. permeability) assigned to each features vector. The average value of the targets might be calculated and assigned to the vector that is remained in the dataset.

6.1.1.1 Considering 7 chemical compound descriptors

In this experiment, *Matérn* covariance function with $\nu = 3/2$ is employed. 6 datasets with 7 molecular features of $MW, MR, MPt, SP, logP, HA$ and HD are examined. Details on the datasets can be found in Chapter 3. The aim is to predict the permeability coefficient. The mean values of *ION* and *MSE (*the mean value is obtained from all the *ION* and *MSE* values for all the predicted points*)* and *Correlation coefficient (CorrCoef)* performance measures can be seen in the first result row of Tables 6.1, 6.2 , and 6.3. Result for each performance measure method (*e.g. ION*) are collated and represented in a single table to make the comparisons convenient. Each row in these tables represents the performance of a regression method (with various numerical features and parameter settings) for the datasets used in this study. To have a better visualisation on the overall *ION* and *MSE* performance of various methods, the best methods' performances for each dataset (in each column) are coloured by blue and the second best values are coloured by green in Tables 6.1 and6.2. Correlation coefficient results are important in pharmaceutical field; therefore, the *Corrcoef* results are added for the main experiments. Other kernel methods with their details are defined in the next sections.

6.1.1.2 Considering 5 chemical compound descriptors

In this experiment, 5 features, which are $MW, SP, logP, HA$ and HD, are used and the model performance is shown in the second result row of Tables 6.1, 6.2 , and 6.3. As explained,

Table 6.1: *ION* performance of the GP, larger positive ION demonstrate better results.

Experiment/Data	**Dataset A**	**Dataset B**	**Dataset C**	**Dataset D**	**Dataset E**	**Dataset F**
GP, 7 features, *Matérn* cov, $\nu = 3/2$	-0.03	0.00	0.12	0.31	-0.10	0.52
GP , 5 features, *Matérn* cov, $\nu = 3/2$	0.19	-0.02	0.35	0.50	0.01	0.37
GP, 5 features, Polynomial Cov, D=5	0.02	0.00	0.00	-0.04	-0.01	-5.85
GP, 5 features, SE cov	0.18	-0.02	0.35	0.55	0.01	0.28
GP, 5 features, *Matérn* cov, $\nu = 1/2$	0.17	-0.01	0.30	0.45	0.00	0.42
GP, 5 features, *Matérn* cov, $\nu = 5/2$	0.19	-0.03	0.35	0.52	0.01	0.35
QSAR (Potts and Guy)	-7.31	-3.49	-0.06	-0.30	-2.69	-0.62
SLN , 2 features (MW and log$P$)	-1.28	-0.1	-0.14	0.11	-0.03	0.12
SLN, 5 features	-6.86	-0.46	0.2	0.01	0.04	0.06
KNN, 5 features	-0.22	-0.48	0.45	-0.09	-0.4	-0.02

Table 6.2: *MSE* performance of the GP, smaller MSE demonstrate better results with less error between predictions and real target values.

Experiment/Data	**Dataset A**	**Dataset B**	**Dataset C**	**Dataset D**	**Dataset E**	**Dataset F**
GP, 7 features, *Matérn* cov, $\nu = 3/2$	1.15	1.64	1.62	1.17	1.23	0.97
GP , 5 features, *Matérn* cov, $\nu = 3/2$	0.81	1.68	1.19	0.84	1.11	1.27
GP, 5 features, Polynomial Cov, D=5	0.97	1.64	1.84	1.75	1.14	13.76
GP, 5 features, SE cov	0.82	1.68	1.14	0.76	1.11	1.44
GP, 5 features, *Matérn* cov, $\nu = 1/2$	0.81	1.66	1.28	0.92	1.12	1.16
GP, 5 features, *Matérn* cov, $\nu = 5/2$	0.81	1.69	1.19	0.81	1.11	1.30
QSAR (Potts and Guy)	9.37	7.38	1.95	2.20	4.15	3.26
KNN, 5 features	1.21	2.04	0.78	1.45	1.3	1.68

Table 6.3: *Corrcoef* performance of the GP, positive *Corrcoef* values closer to 1, demonstrate higher correlation between predictions and real target values.

Experiment/Data	**Dataset A**	**Dataset B**	**Dataset C**	**Dataset D**	**Dataset E**	**Dataset F**
GP, 7 features, *Matérn* cov, $\nu = 3/2$	-0.94	-0.21	0.40	0.54	-0.01	0.72
GP , 5 features, *Matérn* cov, $\nu = 3/2$	0.19	-0.23	0.52	0.69	0.11	0.62
QSAR (Potts and Guy)	-0.96	-0.99	0.04	0.20	-0.81	0.18

MR and *MPt* features are removed, since there they contain a number of zero or missing values. Comparing these results with those obtained in the previous experiment (first row in all tables) shows that except from dataset B with a slightly decrease in performance and dataset F that the performance is also dropped, in general using 5 features, more precise estimation of the target values are obtained. Considering these results, the 5 mentioned physicochemical features are used for the next experiments.

6.1.1.3 Various covariance functions

In this section, the results are obtained using different covariance functions. Polynomial covariance function with parameter D (order)= 5, is employed and the *ION* and *MSE* results are demonstrated in row three of Tables 6.1 and 6.2, respectively.

It is clear that Polynomial Covariance function does not demonstrate a good performance compared to the *Matérn* function in all the datasets. Further investigation shows that $\mathbf{K}$ and $\mathbf{K}^*$ which are related to the weighting factor of the predictions, are very small (almost zero) . Therefore, this covariance function is not suitable to be used on our datasets.

The Squared Exponential (*SE*) covariance function is also applied to the datasets. The results are also shown in Tables 6.1 and 6.2. Comparing the performance of using this method with the others, *SE* gives better performance than Polynomial covariance function (in 5 out of 6 datasets in both *ION* and *MSE*). *SE* also shows better *ION* and *MSE* performance than *Matérn* Covariance function in only one dataset (dataset D), and *Matérn* performs better or the same in the other datasets. Additional investigation about this method shows small weighting factors and the large distance values between test point and training points (r values) are obtained which might make the model unreliable

In addition, *Matérn* function's performance is also examined with parameters $\nu = 1/2$ and $\nu = 5/2$, and the *ION* and *MSE* results are shown in Tables 6.1 and 6.2, respectively. As one can see from these tables, regardless of the performance measure method, the results are quite similar for all three values of ν. As the average performance of the *Matérn* kernel with $\nu = 3/2$ is slightly better for the datasets used in this study, *Matérn* covariance function with $\nu = 3/2$ is chosen as the bench mark method in GP to apply regression in the next experiments.

6.1.2 Linear methods

In order to investigate the performance of linear methods in estimating the permeability values, traditional QSAR and SLN methods are applied to the datasets.

6.1.2.1 Traditional QSAR method

Potts and Guy (1995) QSAR method (see Equation 2.7) is used to estimate the permeability of the chemicals. The *ION*, *MSE* and *corrcoef* are shown in Tables 6.1, 6.2, and 6.3, respectively. The main outcome from these results considering all performance measurement approaches, show that in all 6 datasets, GP works extremely better than linear *QSAR* method. This has been also previously shown in studies by Moss et al. (2011; 2002; 2009), and Sun et al. (2010, 2012) in this field. To make this visually clear, an example of comparison between real target values (permeability coefficients), with the predicted values using GP (*Matérn* covariance function,$\nu = 3/2$), and linear QSAR methods, for all the chemicals in one of this study's datasets (human C) is shown in Figure 6.1. In this figure, it is demonstrated that except from three chemicals (Chemicals number 1, 3 and 11), all the other GP predicted values are closer to the target values than QSAR estimated permeability values.

6.1.2.2 SLN

In addition to the previous method, using the linear *SLN* method (Nabney (2002)), model is trained and tested. The results are presented in Tables 6.1 and 6.2. The results are presented for using two features and five chemical features. Due to the reason that QSAR uses two features (log$P$ and MW) to estimate the skin permeability values, the same molecular features are also used in linear model. The first thing to be noticed is that in both datasets A and B, the linear model works worse than the *Naïve* model and in particular on dataset A with 5 features the *SLN* works very poorly. This is almost certainly due to the fact that these datasets are very small with few vectors. It is also apparent that using *SLN* on the other datasets brings little if any benefit.

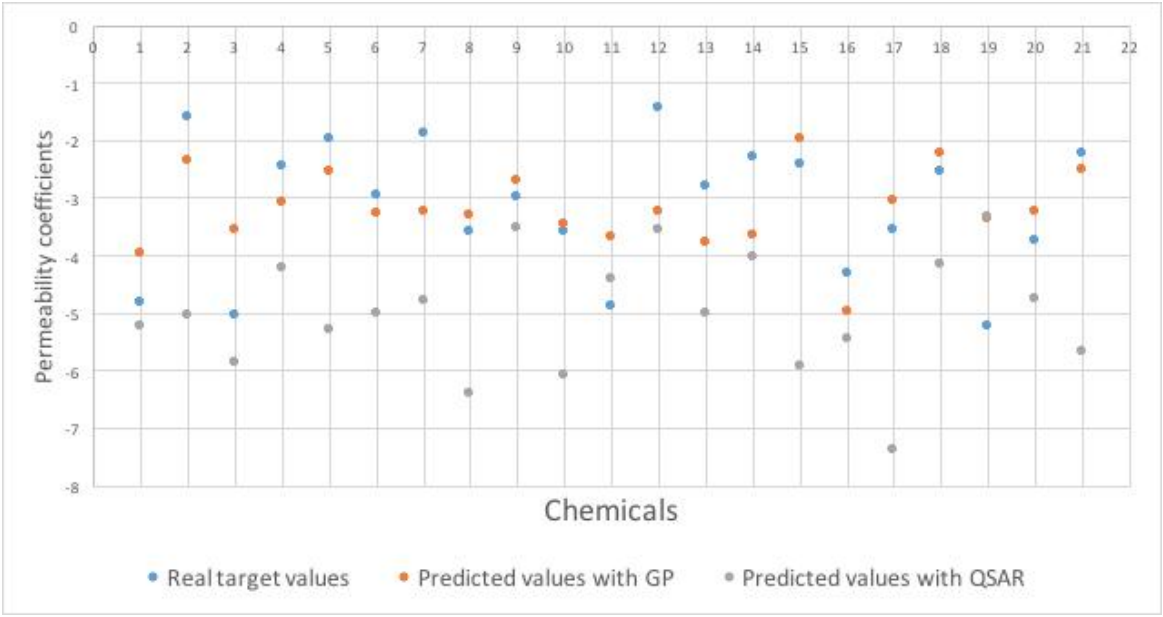


Figure 6.1: Comparison between real target values (permeability coefficients), with the ones predicted with GP (*Matérn* covariance function,$\nu = 3/2$ and *leave-one-out* method), and linear QSAR methods, for all the chemicals in the dataset human C.

6.1.3 KNN application

The performances of applying KNN are shown in Tables 6.1 and 6.2. From both *ION* and *MSE* results, it is clear that KNN does not work properly on most of the datasets (worse than *Naïve* predictor in 5 out of 6 datasets). KNN performance illustrates that the model tends to have over-fitting due to the complexity of the data. The simple Euclidean Distance is implemented to measure the distances in KNN technique.

6.1.4 SVM application

As mentioned earlier, similar feature vectors in training set are problematic in the GP, while they are not an issue in a Support Vector Machine (SVM) regressor. In this method, similarity between the same chemicals with different permeability do not cause any mathematical problems in calculations. Therefore, all the data in the datasets are used to do the experiment. Using a few SVR types with different parameter settings, the one with the best performance for the skin datasets is *epsilon*-SVR (ε-SVR) with RBF (Radial Basis Function) kernel (Chang and Lin (2011)). The optimised values of parameters γ and C are obtained by grid search. The *ION* results are demonstrated in Table 6.4, respectively .

It should be noted that these results are not comparable with the ones in Table 6.1(the test sets of the models are not the same as all the chemicals's features including inconsistent data are used in this dataset). The SVM performance confirms that inconsistent data are not mathematically problematic in this technique, and it can provide a comparable prediction performance with the ones obtained from GP. Studies by Obrezanova and Segall (2010) and Cortes-Ciriano et al. (2015) show that both GP and SVM with relevant parameter settings, result in better performance (classification/regression) than the other methods. In addition, a research by Shah et al. (2016) illustrates that a global approach to modelling a biological process may not necessarily be the best method. The authors suggest that a'mixed-methods' approach (using GP and SVM) to different parts of the chemical space can improve the model performance.

Table 6.4: SVM applying to the data-Considering 5 and 7 data features. The results can be compared to the GP performance in the same table.

Performance/Data	Dataset A	Dataset B	Dataset C	Dataset D	Dataset E	Dataset F
***ION* SVM-5 features (mean)**	-0.07	-0.15	0.28	0.35	0	0.42
***ION* SVM-7 features (mean)**	-0.3	-0.07	0.18	0.37	-0.04	0.52

6.1.5 Conclusion

The aim of the experiments in this section was to evaluate the performance of different regression methods, with proper parameter settings on the skin permeability estimation problem. Blue and green coloured values in Tables 6.1 and6.2 show the best and second best performance obtained for each dataset, using various methods and parameters. Therefore, methods (rows) with more blue and green values demonstrate better models. The first conclusion from these tables is that QSAR method works worse than all the other methods in all the datasets. Additionally, it can be seen that GP with *Matérn* covariance function (with all three values ν= 1/2, 3/2, 5/2) perform similar, using 5 chemical features; however, the average of performance of this function with $\nu = 3/2$ in all datasets, outperforms the GP with *Matérn* covariance function with $\nu = 1/2, 5/2$. Using GP method with *SE* kernel shows comparable performance with *Matérn* kernel, however, as discussed, small weighting factors and large distance values between test point and training data might make this model unreliable.

In addition, the result represent that SVM performs reasonably good and comparable to GP with the *Matérn* covariance function ($\nu = 3/2$) to predict the permeability values. A mix of these methods may be used for best quality predictions. Since this study's main focus is on GP, this approach is employed (using 5 features) as the fundamental technique to perform further experiments in this thesis.

6.2 Experiment 2: Comparing human, mouse, rat and pig models

In this section, four datasets (complete human, mouse, rat and pig) are used to perform the experiments. The GP is applied to one of the datasets each time to obtain a GP model. The model of that dataset then is used to predict the permeability of the other datasets. As an example, the model is first trained, using the human data and then predict the mouse data permeability using human GP model. This is especially an important question in pharmaceutical field that which of the mammalian skins may perform as a better model to be trained to result in better permeability estimation for the other types. Animal membranes can be used as they are economical in both time and financially, and for safety reasons, so humans do not get exposed to problems with chemicals. However, the relevance of rodents in particular has been questioned with most researchers now agreeing that rodent skin has little correlation to human skin in terms of permeability.

From pharmaceutical point of view, this experiment is important from two aspects. First, if we can make models work better (than animals), we do not need to use animal models as it might be unethical. Second, we need to consider whether the outputs (from animal data) match human in both predictions and mechanistic understanding, and this would allow us to conclude as to the good or bad use of animal tissues to replace human studies.

6.2.1 Datasets

The details about the nature and the size of the four datasets employed in this study are explained in Chapter 3. Each dataset includes measurements of both the chemical properties (5 descriptors are used) and the absorption rate of the variety of compounds. As mentioned previously, the average value of the permeability for inconsistent chemicals is used to deal with the inconsistency problem. There is also a degree of overlap between datasets after removing the inconsistency which are also shown in Table 6.5.

Table 6.5: Overlap among the complete human, rat and mouse datasets

Datasets	Complete human	Rat	Mouse	Pig
Complete human	*(121)	17	23	4
Rat	17	(26)	5	4
Mouse	23	5	(46)	2
Pig	4	4	2	(14)

* As only five chemical descriptors are used in this study (temperature is not considered), the size is less than the the complete human dataset in Table 3.1(with temperature the size=145)

6.2.2 Experiments and results

Two major experiments have been performed in this section.

1. The first one is to measure the performance of the model for each dataset separately.
2. The second experiment investigates the performance of employing one dataset model to predict the skin permeability of the other types of the mammalian skins.
3. Finally, the last experiment is performed the same as the second experiment but SVM is used to train the model.

6.2.2.1 The performance of each model using GP model

The *Leave-One-Out* technique is used for each of the datasets, and the *ION* values are shown in Table 6.6. The results show that the best prediction performance is obtained for pig and that the prediction performance of rat and human data are better than mouse data. Comparing the ranges of chemical features and permeability values of pig and other datasets (can be seen in Figures 3.2 and 3.4) illustrates that the comprehensive performance of Pig dataset (with only 14 data) may be caused by having the largest range of permeability values (targets) among the others.

Table 6.6: GP prediction performances using *leave-one-out* in each dataset

Dataset	***ION (mean)***	***CorrCoef***
Complete human	0.40	0.63
Mouse	0.32	0.55
Rat	0.42	0.65
Pig	0.77	0.88

6.2.2.2 Effect of using one mammalian model to predict the skin permeability of the others

Having repetitions in training and test datasets: In this experiment, I examine how using one of the human, mouse and rat datasets to train the model hypothetically may help to predict the other groups permeability coefficients. In this method, one dataset is used to train the model and the other dataset is used as the test set at which the permeability values are predicted. The same chemicals exist in both test and training sets are kept. The following stages are taken to do this experiment:

1. The training and test datasets are first normalised as discussed earlier. The GP model is trained with human dataset and then each time one of the mouse, rat and pig datasets is used as the test set to estimate their targets based on the model which has been had trained. The prediction performances (*ION*) are then calculated.

2. This process is repeated for mouse, rat and pig datasets as well and then the model is used to get the performance of the other datasets and the prediction performances are reported as *ION* in Table 6.7. The datasets shown in red are the training sets and the blue datasets are the test sets and the prediction performances are reported for them.

The best *ION* (0.57) is achieved for the Rat dataset where the model is trained with Mouse data. However, the best *Correlation coefficient* (0.80) is related to the Rat dataset as well, but the model is trained with Human data. Comparing these results with those in Table 6.6 shows that rat dataset permeability values can be predicted better when the model is trained with either human or mouse dataset, However, the vice versa is not correct as the datasets sizes are not the same. The important finding is that the best permeability predictions for human data are obtained when the model is trained by mouse data. Training the model

with Pig dataset for predicting the other datasets target values is not beneficial on any of the datasets.

Table 6.7: GP prediction performances training the models and test on the other datasets. Comparing these results with the ones in Table 6.6 shows rat dataset permeability values can be predicted better when the model is trained with either human or mouse dataset.

To be tested on	***ION***				***CorrCoef***			
	The training set				**The training set**			
	Complete human	**Mouse**	**Rat**	**Pig**	**Complete human**	**Mouse**	**Rat**	**Pig**
Complete human	——	0.36	0.21	0.12	——	0.64	0.43	0.36
Mouse	0.22	——	0.10	0.11	0.64	——	0.26	0.34
Rat	0.52	0.57	——	0.07	0.80	0.65	——	0.15
Pig	0.19	0.10	0.03	——	0.34	0.30	0.10	——

Removing repetitions from training or test datasets: The second part of this analysis considers removing the repetitions between training and test data and investigates its effect on predictions performances. The steps have been taken to conduct the experiment are as follows:

1. The repeated vectors of chemical features are removed from the training set where the number of data is more than the number of data in the test set, otherwise, the repeated data is removed from the test set.

2. All the process in the previous experiment is then followed for each of the datasets and the results are reported in Table 6.8. The same as the previous experiment, the datasets with red colour are the training sets and the blue datasets performance are reported in the results table.

As expected, due to removing the repetitions between training and test sets, there is a decrease in the prediction performances. However, interestingly the best performances are obtained for the same training and test sets of the previous experiment (0.50 and 0.64 for *ION* and *Correlation coefficient*, respectively). The results illustrate that human and mouse data result in better performances when they are used to train the model and then the model is used for the rat data permeability estimation, compared to the model which is trained with

rat data (Table 6.6). Similar poor performance is obtained when the models are trained with Pig data.

Table 6.8: GP prediction performances training the models and test on the other datasets, after removing the repetitions Comparing these results with the ones in Table 6.6 shows rat dataset permeability values can be predicted better when the model is trained with mouse dataset.

To be tested on	***ION***				***CorrCoef***			
	The training set				**The training set**			
	Complete human	**Mouse**	**Rat**	**Pig**	**Complete human**	**Mouse**	**Rat**	**Pig**
Complete human	——	0.32	0.11	0.09	——	0.56	0.31	0.31
Mouse	0.18	——	0.11	0.07	0.51	——	0.32	0.28
Rat	0.45	0.50	——	0.06	0.64	0.61	——	0.12
Pig	0.15	0.7	0.02	——	0.26	0.24	0.06	——

6.2.2.3 Effect of using one mammalian model to predict the skin permeability of the other groups (using SVM)

Considering all the experiments in this section, I also apply SVM to the datasets (included all the inconsistent data) to investigate whether better performance can be obtained. The outcome shows that SVM did not change the performance by a great amount. Therefore, I do not report the results in this section.

6.2.3 Conclusion

To summarise this section, the best results were obtained using the mouse and human data to train the model and then using the obtained models to predict the permeability of the rat data. The pig dataset shows to have a thorough performance (using *leave-one-out*) comparing to the other ones (*ION*=0.77 and *CorrCoef*=0.88); however, training the model with Pig dataset to predict the other datasets' permeability values result in poor performance.

The other interesting finding is that training the model with mouse data to obtain the human data permeability values, could result in better performance than training the model with rat data. It should be noted that, all models work better than the *Naive* models.

6.3 Experiment 3 : The effects of experimental conditions (environment temperature and diffusion cells type) on permeability predictions

A closer look at the datasets and their prediction performances raised an interesting question about the effect of experimental conditions on the permeability predictions. The question is, whether the experimental conditions such as temperature and various diffusion cells can affect the performance of the model. In this section the effect of temperature change and using the static or flow-through cells on the prediction performances are investigated.

6.3.1 Temperature effect on the model performance

To examine the effect of temperature on chemical's permeability across the skin, this numerical feature is added to the training data. Temperature values that used for this study includes 37^{o}C (all 32^{o}C skin surface temperature are considered to be 37^{o}C as it shows the temperature of the diffusion cell during the experiment), in addition to 30^{o}C, 27^{o}C, 26^{o}C, 25^{o}C, 23^{o}C. For the *'not given'* values the corresponding chemical features are removed. then omitted. The temperature ranges in the datasets are demonstrated in Chapter 3, Figure 3.1. Temperature values are added with other numerical molecular features and the *ION* performance of the experiments are shown in Table 6.9. To make the comparison easier, the *ION* performance of the datasets before adding the temperature is also shown in the table. Since among datasets A-D, the temperature is constant, the experiment is performed only on the datasets E and F. The results yield much better performance in the dataset F (about 73% increase). This is due to the fact that dataset F has the largest range of temperature values.

6.3.2 Conclusion

This shows that adding temperature as a numerical feature to the data can be helpful to increase the prediction performance, and it is especially effective when the temperature range in the dataset is large.

More exploration is also performed employing 7 features together with the temperature values (using GP) results do not change by great amount.

Table 6.9: *ION* performances with and without temperature added to the 5 features

Using 5 features	**Dataset E**	**Dataset F**
***ION*-without temp**	0.00	0.37
***ION*-adding temp**	0.01	0.64

6.3.3 Using only flow-through or static diffusion cells

In this experiment, the static and flow-through cell data are separated and the performance of each of these groups is calculated separately. There are in total 93 static and 53 flow-through data in the complete human dataset. In the first experiment, the model is trained based on the flow-through data and the predictions are obtained for flow-through data only (*leave-one-out* is used). The results of the experiment have been shown in Table 6.10. Then similarly, the model is trained only based on the static data and the predictions are achieved only for static data and the results are shown in Table 6.11. The results show that using flow-through data to train the model results in poor prediction performance for the flow-through cell data(almost the same as *Naive* predictor); However, training the model with static data to predict the static data permeability values, yield so much better performance.

Table 6.10: GP prediction performances considering only flow-through cell data to train the model and predict the flow-through cells data permeabilities (*leave-one-out*). The results compared to the ones in Table 6.11 show that static data results in much better prediction performance than the flow-through data.

Flow-through only for Training	**Mean**
MSE_GP_flow-through	0.84
***ION*_GP_flow-through**	0.04
MSE_Naive_GP_flow-through	0.87
***CorrCoef*_GP_flow-through**	0.20

Table 6.11: GP prediction performances considering only static cell data to train the model and predict the static cells data permeabilities (*leave-one-out*). The results compared to the ones in Table 6.10 show that static data results in much better prediction performance than the flow-through data.

Static only for training	**Mean**
MSE_GP_static	0.98
***ION*_GP_static**	0.42
MSE_Naive_GP_static	1.68
***CorrCoef*_GP_static**	0.66

Since, static data resulted in better prediction performances, in the following experiment I train the model based on the static data and estimate the flow-through data permeability values. In addition, I investigate on training the model with flow-through data and obtaining predictions for static data. The results of these experiments are shown in Tables 6.12 and 6.13. One can see that this approach does not work well on any of the static or flow-through data and the prediction performance for both data types are dropped.

Table 6.12: GP prediction performances Considering only static cell data to train the model and predict the flow-through cells data permeabilities. The results compared to the ones in Table 6.10 show that this training model does not results in better performance for flow-through data.

Static only for Training	**Mean**
MSE_GP_flow-through	1.17
***ION*_GP_flow-through**	-0.35
MSE_Naive_flow-through	0.86
***CorrCoef*_GP_flow-through**	0.07

Table 6.13: GP prediction performances Considering only flow-through cell data to train the model predict the static cells data permeabilities. The results compared to the ones in Table 6.11 show that this training model does not results in better performance for static data.

Flow-through only for Training	**Mean**
MSE_GP_static	1.64
***ION*_GP_static**	0.05
MSE_Naive_static	1.73
***CorrCoef*_GP_static**	0.19

6.3.4 Mixing static and flow-through data

In this experiment, the static and flow-through cell data are collated together (in total 143 data). Subsequently, 10 different training and test sets from the mixed static and flow-through cell data are randomly selected based the following process:

1. As mentioned, the number of flow-through and static cell data are different in the complete data (size=93 for static data and size=53 for flow-through data in the complete data). Therefore, the same number of static and flow-through data are selected and included in each training set (36 data are selected randomly from each of static and flow-through data). The rest of the data that remained in the complete datasets is used as the test set. The test set contains unequal mixed static and flow-through cell data. This process is done for 10 times to obtain 10 random training and test sets.

2. The 10 sets are trained separately and the predictions obtained for their corresponding test sets.

3. As the experiment have been performed 10 times to obtain the predictions on 10 test sets, the mean and standard deviation (*STD*) of the 10 experiments performances are reported in Table 6.14.

Although the results are not comparable with the ones in Tables 6.12 and 6.13 (as their test sets are different), it can bee seen from Table 6.14, that using both datasets to train the model does not bring much benefit and the prediction performance for flow-through data is still very poor.

Table 6.14: GP prediction performances mixing static and flow-through cell data. The results show that mixing the data do not bring much benefit to predict the static and especially flow-through data (with very low performance) permeability values.

Mixed_static_flow-through	**Mean (over 10 experiments)**	**STD (over 10 experiments)**
MSE_GP_flow-through	0.93	0.23
MSE_GP_static	0.96	0.09
***ION*_GP_flow-through**	-0.07	0.09
***ION*_GP_static**	0.43	0.05
***CorrCoef*_GP_flow-through**	0.19	0.16
***CorrCoef*_GP_static**	0.67	0.05
MSE_GP_Naive_flow-through	0.86	0.14
MSE_GP_Naive_static	1.70	0.09

6.3.5 Conclusion

To conclude, the experimental condition features such as temperature and cell type can affect the model prediction. As for temperature, with large range of temperature values in the data, the performance of the model can be improved by a great amount. Furthermore, separating the data based on the different cell types show that the best predictive models are always obtained when static diffusion cell data permeability is predicted compared to models constructed from flow-through cell experiments. These results are obtained regardless of whether data from static or flow-through cells, or mixtures of both, are used to train models. Further, training models based on flow-through cell data only, and predicting the permeability of 'unseen' test data resulted in poor models.

It is apparent that the quality of the model is directly affected by the nature of the input data, and that the inclusion of data from flow-through experiments may reduce overall model quality and predictive power, while models based solely on such data offer poor predictions of skin permeability. This is in significant contrast to models developed from static diffusion cell experiments, which resulted in highly predictive models. It may therefore be suggested that, in order to optimise the model quality, data from only static, franz-type, experiments should be used to construct computation models. It is also interesting to investigate using different chemical features when flow-through cell type is used.

6.4 Experiment 4: Mixing numerical and nominal Data

The aim of this experiment is to consider the nominal experimental features and investigate their probable effect on the predictions. To do this, 6 numerical data features including the temperature and three nominal features as the experimental conditions are considered. The numerical and nominal features of the datasets are illustrated in Chapter 3 (see 3.2.2). To obtain the covariance function of the numerical data, *Matérn* function ($\nu = 3/2$) is used and a categorical kernel function based on hamming distance by Couto (2005) (see Chapter 4, section 4.3.2.1) is employed to calculate the categorical kernel function.

Finally, the covariance matrices are added together considering different weighting factors multiplied by them. The weights are chosen to be μ and 1-μ for categorical and *Matérn* covariance/correlation matrices, respectively; these weighted kernels are then added together. The μ values are considered to be 0.4 and 0.8. So, the weight of categorical and *Matérn* covariance/correlation matrices are first set to 0.4 and 0.6 and the experiment is performed. For the second experiment, the weights are considered to be 0.8 and 0.2, respectively.

The results are reported for different values of $\lambda = 0.01$, 0.5, 0.8 and 0.99 in Tables 6.16 and 6.17. For conveniency, the *ION* results for numerical data including temperature, are added in Table 6.15. Comparing the obtained *ION* values using both numerical and categorical data to the ones using only numerical data, it can be seen performance of the model is so poor when (λ=0.01). Considering the other parameter values for λ and μ , except from dataset E, all other datasets can have a slightly improvement by varying these parameters.

Table 6.15: *ION* performances only numerical data with temperature. This is the benchmark results as the best prediction performances are obtained using GP (with 6 numerical data, *Matérn* function ($\nu = 3/2$)). These should be compared to the ones in Table 6.16.

Using 5 features	**Dataset A**	**Dataset B**	**Dataset C**	**Dataset D**	**Dataset E**	**Dataset F**
***ION*-adding temp (mean)**	0.19	-0.03	0.38	0.33	0.01	0.64

Table 6.16: Adding categorical features to the 6 numerical data features (Higher *ION* better) $\mu = 0.4$. These results should be compared to the ones in Tables 6.15 and 6.17to examine which λ parameter setting performs better among all.

Performance/dataset	Dataset A	Dataset B	Dataset C	Dataset D	Dataset E	Dataset F
ION-mean (λ=0.01)	-0.5	-19.09	-1.85	-70.84	-13.98	-48.45
ION-mean(λ=0.5)	0.21	0.02	0.35	0.3	-0.01	0.66
ION-mean(λ=0.8)	0.21	-0.05	0.38	0.33	-0.01	0.66
ION-mean(λ=0.99)	0.21	-0.05	0.38	0.33	-0.01	0.66

Table 6.17: Adding categorical features to the 5 numerical data features (Higher *ION* better) $\mu = 0.8$. These results should be compared to the ones in Tables6.15 and 6.16 to examine which λ parameter setting performs better among all.

Performance/dataset	Dataset A	Dataset B	Dataset C	Dataset D	Dataset E	Dataset F
ION-mean(λ=0.01)	-0.7	-28.59	-4.42	-159.25	-27.44	-80.53
ION-mean(λ=0.5)	0.2	0.05	0.31	0.29	-0.02	0.65
ION-mean(λ=0.8)	0.21	-0.05	0.37	0.32	-0.02	0.66
ION-mean(λ=0.99)	0.21	-0.08	0.38	0.32	-0.01	0.66

6.4.1 Conclusion

The small improvement in this experiment's performance compared to the ones in Table 6.15 may be caused by involving cell type in the features, as can be seen in 6.3, it can decrease the performance with flow-through cell data. More investigation needs to be done in future to consider this case or to combine the categorical kernel with the numerical kernel using the other weighting factors and combining methods (e.g. by multiplying them together with different weights).

Table 6.18: Number of points in each cluster obtained from applying GNG

Clusters	**DS 1**	**DS 2**	**DS 3**	**DS 4**
Number of points	41	79	13	12

6.5 Experiment 5: Data clustering (using Growing Neural Gas algorithm)

This experiment has been performed on the complete human dataset by gathering all the human datasets and removing the repetitions. The average value of the permeabilities considered for the inconsistent data. The GNG by Fritzke et al. (1995) (see section 4.7 on Chapter 4) is applied to the complete dataset to investigate the nature of the generated clusters. The maximum nodes is set to 100 and the default values are selected for the other parameters of the GNG (Loos and Fritzke (1998)). Employing GNG on the complete dataset with 145 data points with 6 chemical features including temperature, 14 natural clusters are obtained. Some of the obtained clusters include only 2 or 3 data which cause problem in GP (the training set can not have only 1 or 2 data). To solve this issue, the centre of the clusters are calculated and the closer clusters are combined together. As a result, 4 clusters are obtained from GNG. PCA is used to show the clusters in two dimensions. Figure 6.2 shows the PC1 against PC2 of the data points and the obtained clusters from GNG, are shown in different colours. The number of data points in each of the clusters are also shown in Table 6.18.

To predict the permeability of the chemicals in the new clusters, and compare them with the the previous permeability estimations in their original clusters (datasets), the following steps have been taken:

- In each cluster, for the same chemicals with different permeability values (targets), one of them (with its all chemical features) is kept and the average value of the various permeability values is considered as that chemical's target value.

- GP is applied separately to each of the GNG clusters and using *leave-one-out* technique, the predictions for each of the chemicals in their new clusters are obtained.

Table 6.19: Comparing the MSE of predictions in the original datasets and the new GNG clusters and the overall MSE of the predictions

Original human datasets	**Dataset A**	**Dataset B**	**Dataset C**	**Dataset D**	**Dataset E**	**Dataset F**	**Overall datasets**
MSE	0.96	1.68	1.19	0.84	1.11	1.27	1.62
GNG clusters	**DS 1**	**DS 2**	**DS 3**	**DS4**	**NA**	**NA**	**Overall clusters**
MSE	1.16	0.80	0.90	0.59	—	——	0.89

- GP is applied to the original datasets (6 human datasets) and the predictions also obtained for each of the chemicals in the their original datasets (using *leave-one-out*).
- The obtained permeability values for the chemicals in each of the new clusters (from GNG) then are compared to the predictions achieved for the chemicals in their original datasets.
- Finally, the MSE of the predictions are calculated for all the chemical in both their previous datasets and new GNG clusters. It is interesting that for 107 chemicals out of 145, the error of the predictions in the GNG clusters is less than the average value of the errors in the previous datasets.

Table 6.19 shows the MSE in each of the predictions in the original datasets and the GNG clusters. The overall MSE of all compounds are also shown in the same table. From these results, one can see that GNG clusters decrease the prediction errors by 45% and clustering the large data using GNG is suggested to be used in future experiments for target (permeability) estimation.

6.5.1 Conclusion

In summary, GNG clustering works better in predicting 107 chemical's permeability (out of 145) which are the majority of the compounds. In addition, the absolute error box-plot (Figure 6.3) shows that GNG clustering result in better predictions overall the compounds. The MSE over all the compounds improved 45% (from 1.62 in the previous datasets to the

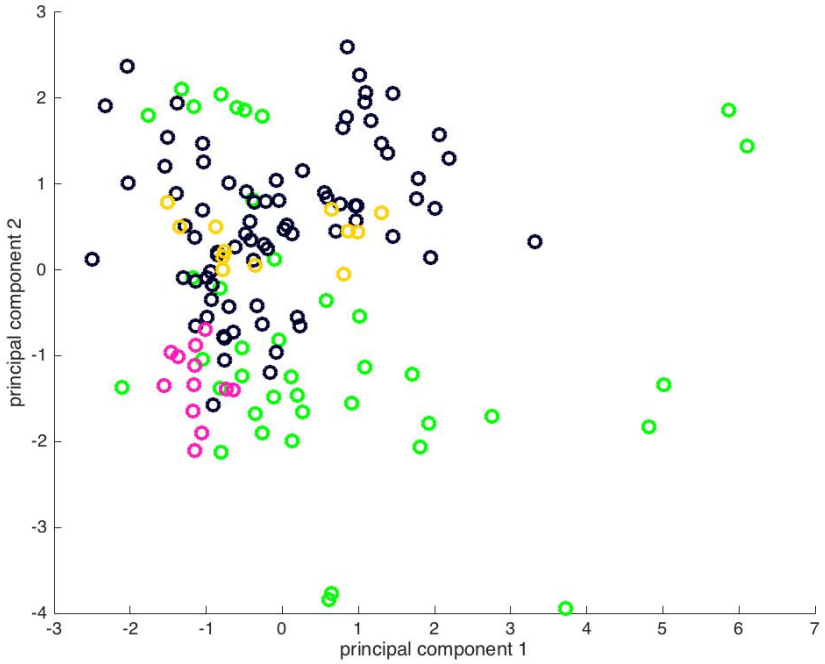


Figure 6.2: PC1-PC2 application of GNG and the natural clusters

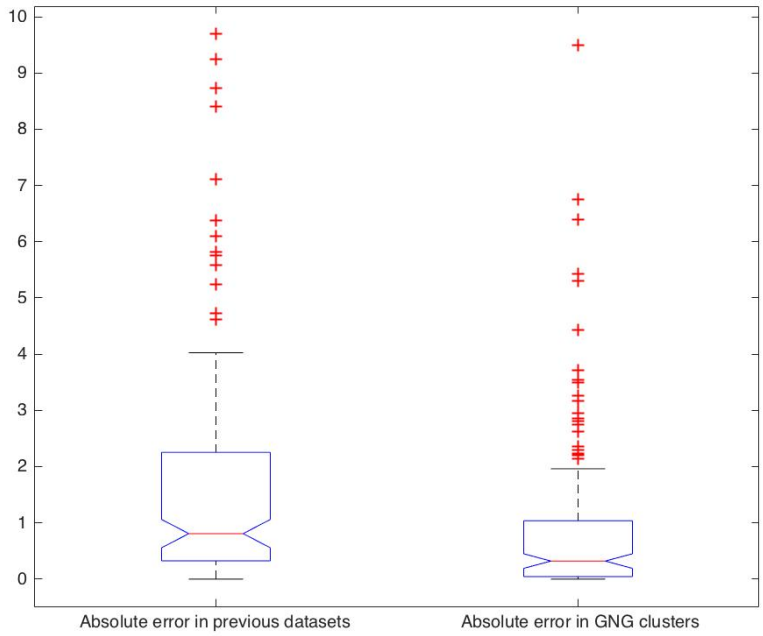


Figure 6.3: Comparing Absolute error of predictions between the chemicals in their previous datasets and new GNG clusters

0.89 in the GNG clusters). However, predictions in the original datasets still work better for 42 chemicals. In the next section, I examine applying the Monte Carlo method to deal with inconsistency in the data.

6.6 Experiment 6: Using Monte Carlo method to deal with inconsistent data

As mentioned earlier, the datasets used for this study may include the same chemicals with various target (e.g. permeability) values. This *'inconsistent data'* may cause problem for predictions. To deal with this issue, the mean of all target values corresponding to a single chemical in each of the datasets is considered in all the previous experiments. In this section, application of *Monte Carlo* method (see 4.8) is investigated to see whether a better solution for the average of target values can be found. To do so, the following steps should be taken:

- The human datasets are gathered in a set 'complete human dataset' (Table 3.1)
- 2/3 of the data is considered as training set and 1/3 is considered as test set.
- For the compounds with more than one target value (e.g. permeability rate), one of the values is randomly chosen and placed in the training set.
- Model is trained with GP and predictions are obtained for the test sets using 3-fold cross validation. The mean *ION* (for three test sets) and correlation coefficients are obtained.
- In order to establish which values worked best, this whole process is repeated with 10000 different training sets and10000 different random choices.
- The best *ION (mean)* and *CorrCoef* among all 10000 repetitions of the experiment is then found and the best selection for each compound target values can be obtained. This dataset can be used as a proper training set for further experiments and target estimations.

Table 6.20: Complete human dataset, the performance of using *Monte Carlo* method is compared with the ones in which average of targets are used for inconsistent data.

Performance measures	
***ION* (with mean of targets)**	0.34
***ION* (with Monte Carlo the best)**	0.41
***CorrCoef* (with mean of targets)**	0.55
***CorrCoef* (with Monte Carlo the best)**	0.64

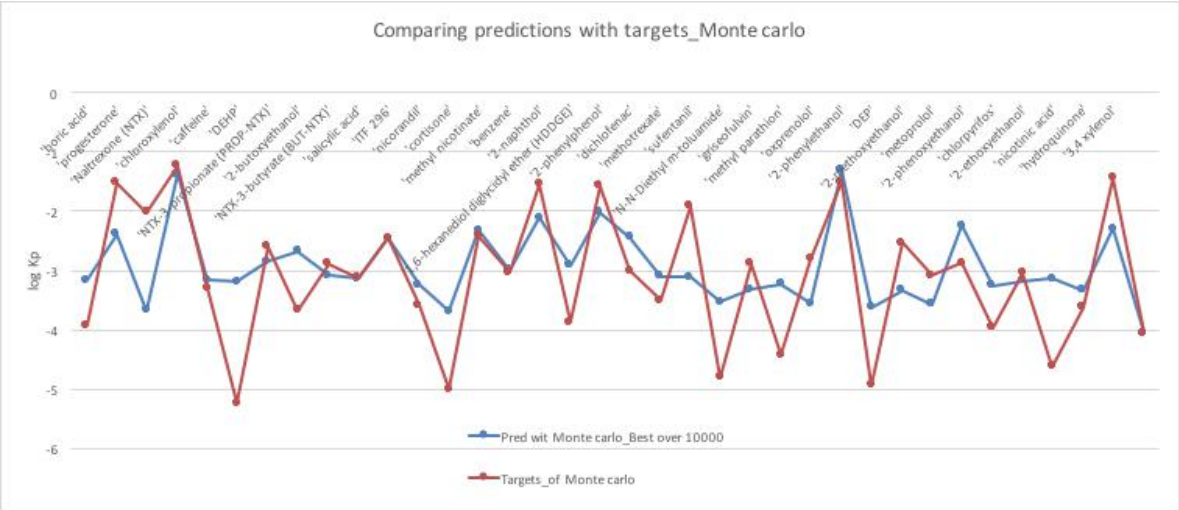


Figure 6.4: Comparison of estimates with targets (*Monte Carlo* method).

The performance of this method are compared with the original dataset that mean values are used for each inconsistent compound. The same test sets with 3-fold cross validation are used to do the experiment on the original datasets (mean of target values for inconsistent data). The comparison of results between these two methods performances is shown in Table 6.20. Figures 6.4 and 6.5 show a summary of the total results – they highlight the estimates for 35 chemicals taken from the "best" Monte Carlo data set. From these results it is apparent that, in most cases, the application of the Monte Carlo method works better than using a data set containing mean values of chemicals in that it produces better estimates. This analysis also yields an 'optimal' data set; starting with the complete data set repetitive data points are removed and the best target values of the statistical metrics are produced.

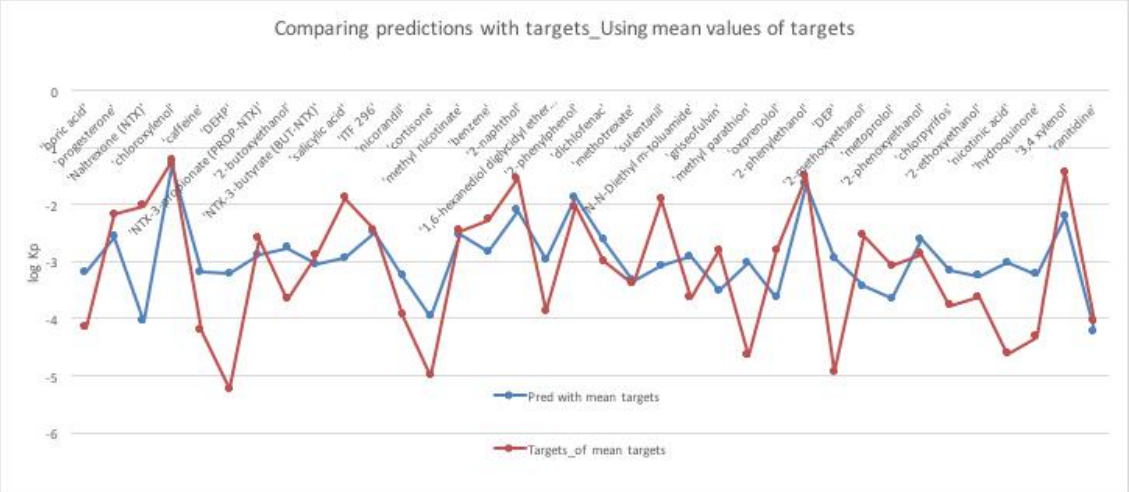


Figure 6.5: Comparison of estimates with targets (conventional GP methods).

6.6.1 Conclusion

The *ION* and *correlation coefficient* for the non-optimised (using the mean value for each inconsistent compound) model are 0.34 and 0.55, respectively; these increased in the optimised model to 0.41 and 0.64, respectively. This experiment shows that following the application of the random sampling approach of the Monte Carlo simulation, results in a measurably improved model. Therefore the Monte Carlo method provides a clear opportunity to optimise a dataset with inconsistent data points.

Chapter 7

Hyper-parameter Optimisation Methods

In the previous chapters, it is shown that machine learning techniques, specifically Gaussian Process (GP) methods have better prediction performance compared to the QSAR traditional models (Moss et al. (2009); Obrezanova et al. (2007); Burden (2001); Cortes-Ciriano et al. (2015)). In this chapter, the problem of finding optimised hyper-parameters involved in GP kernel functions is addressed. In particular, finding suitable hyper-parameters with the small datasets used for this study is addressed. Ii is shown that the choice of hyper-parameters can play an important role in obtaining good model. *Matérn* kernel is used as the covariance function of the model as it shows promising results in my previous work. It is also demonstrated that a §particular method for finding good hyper-parameters namely smooth-box hyper-prior method finds good parameters with all the datasets. Therefore, this method is recommended to be used with small datasets. This is one of the major findings of my thesis.

7.1 Introduction

Obrezanova et al. (2007) show that the Gaussian Processes method is comparable to and sometimes exceeds artificial neural networks in performance. There are hyper-parameters such as length scale (l), signal variance σ_f^2 and noise variance σ_n^2 involved in the GP regression.

An important feature of the data used in this study is that the datasets are relatively small (the sizes of the datasets vary between n=9 to n=85). The study performed by Steyerberg et al. (2000), investigates the logistic regression method with small datasets (61 small containing on average 336 objects). To obtain the small subsets, the data (size=20512) was split into a training and test part. The regression estimation methods in their study included standard maximum likelihood, the use of linear linkage factor, penalised maximum likelihood and the Lasso on univariable regression coefficient. The research indicates that using shrinkage methods in full models including predefined predictors and external information resulted in the best prediction performance in the small datasets. In addition, the work done by Dreiseitl and Ohno-Machado (2002) investigates the logistic regression and artificial neural network models as predictive models. They show that the problem of lacking data can be addressed by applying cross-validation or bootstrapping methods to make the best possible use of the limited amount of data.

Several studies have investigated hyper-parameter optimisation, such as the work by Bergstra et al. (2011) and Bergstra and Bengio (2012) on random search through hyper-parameters space which is also used in my study. Bardenet and Kégl (2010) showed that using Evolutionary Algorithms (EA) to optimise the hyper-parameters can outperform the other methods. Similarly, MacKay (1997) demonstrated that the hyper-parameter optimisation landscape is a multi- modal. This suggests that a slower but more robust global optimiser, such as an EA, may yield better results. Using a variety of EA methods including the advanced ones such as CMA (Covariance Matrix Adaptation) and CMA-ES (Covariance Matrix Adaptation Evolution Strategy) by Hansen and Ostermeier (2001) and accelerating EA as in Büche et al. (2005) showed the usefulness of evolutionary methods in optimisation specifically for GP hyper-parameters. In the hyper-prior technique, a prior distribution is assigned to the hyper-parameters and the distribution is then modified to improve the fit to the actual data.

In this chapter, the performances of different hyper-parameters optimisation strategies in the GP are investigated. In addition to EA and the hyper-prior optimisation method, the results of applying the *Conjugate gradient (marginal likelihood maximisation* (MLM) techniques*),* manual grid search and random search (Bergstra and Bengio (2012)) through the hyper-parameters space are discussed.

7.2 Datasets

In this work, 11 small datasets are investigated. Thay are collated from different sources. The datasets and their features specifications are shown in Chapter 3 with more details.

The number of data records in each dataset after refining are shown in Table 7.1. The small size is due to the fact that gathering pharmaceutical data is difficult, time consuming, and expensive. In these circumstances obtaining good predictions from small data sets is very important.

Table 7.1: Number of data-points in each dataset. The first 5 datasets are related to the same group of the data with the same number of features (5 chemical features which are $MW, SP, logP, HA$ and HD). ER is the enhanced ratio data with 6 features and the last 5 datasets are from Magnusson et al. (2004) with 6 features. All the datasets and their features are introduced in Chapter 3 of this thesis.

Dataset	**Human C**	**Human F**	**Rat**	**Mouse**	**Pig**	**ER**
# Data points	21	84	26	46	14	71
# Used molecular features	5	5	5	5	5	5
Dataset	**Mag-set A**	**Mag-set B**	**Mag-set C**	**Mag set-D**	**Mag-set E**	-
# Data points	85	50	27	45	36	-
# Used molecular features	6	6	6	6	5	-

7.2.1 Experimental set up

7.2.1.1 Software

The software used for this study is Gaussian Process modelling for non-linear regression (Rasmussen (2006b,c)) and the latest version of toolbox defined in Rasmussen and Nickisch (2015), is used for the hyper-prior optimisation method. The Genetic Algorithm (GA) Matlab optimisation toolbox is used for Evolutionary Algorithm hyper-parameter optimisation.

7.2.1.2 Cross validation

A 5 fold cross-validation has been applied in this chapter's experiments. Each time one of the folds is considered as the test set and the rest are considered as the training set

plus the validation set. The mentioned hyper-parameter optimisation methods are applied to the training set and the prediction performances are gained for the validation set. The best hyper-parameters are the ones minimise the average values of negative log likelihood (NLL) over validation sets. Furthermore, the best set of hyper-parameters is used to predict the permeability values of the test sets. Therefore, the results are the mean of 5 prediction performance values together with their standard deviations.

7.2.1.3 Experimental Initialisations

In this section, all initialisations set up for the experiments are explained as the following:

- **Grid search**: To do the manual search through the hyper-parameters space, the hyper-parameters ranges are considered in [10^{-3}, 10^{3}] with 20 equal distance steps. Using 5-fold cross validation, the model is trained with all the 8000 ($20\times20\times20$) different sets of the hyper-parameters and obtained the predictions for the test sets. The average values and their standard deviation among 5 folds are then reported. Looking at the prediction performances, a finer search for better values of the hyper-parameters is performed and the searching range is limited to [0.01 10] with 20 steps, as no better results are obtained using the hyper-parameters out of this range. The model is trained with the new hyper-parameters and the best performance are reported.

- **Random search**: For random search, 20 values for each hyper-parameters are obtained randomly between the same range [0.01 10] which is also considered in the grid search. Using a 5-fold cross validation, the model is then trained and the predictions are obtained. Since, in each run of this experiment, the hyper-parameters are selected randomly, this experiment is run for 5 times and the results are the average of 5 mean and standard deviation values obtained from each 5-fold cross validation.

- **Conjugate Gradient:** The hyper-parameters were initialised to log(0.5). The number of function evaluations is set to 100.

- **Hyper-prior methods**:

- **Gaussian prior and Laplacian prior:** The mean and variance parameters of the Gaussian and Laplacian priors are set to constant values of 0.1 and 0.01 respectively. The mentioned values are obtained as the best prediction performances using cross validation in each of the datasets.
- **Smooth-box prior:** Based on the Equation 4.17 in Chapter 4, in these experiments the a, b and η are set to 10^{-3}, 10 and 2, respectively. Different values of η are tried and 2 is the best value obtained for my datasets.

- **Evolutionary Algorithm**: In this research, heuristic fitness function is employed with ratio=0.7 to accelerate convergence. Various ratio values (in a range from 0.1 to 1.2) are tried and it turns out, ratio=0.7 works better for the datasets used in this study. In each of the 50 generations, there are 50 populations and the optimised hyper-parameters are obtained in the last generation. Elite children is set to 4 and the mutation function is uniform, means the children are randomly selected from a uniform distribution within the range of hyper-parameters. Crossover fraction is 0.8, meaning the rest of children in a population, are 4 Elite children and also reproduced from mutation. The first generation's population is initialised randomly, therefore, similar to the random search, this experiment is also repeated for 5 times. Genetic Algorithm toolbox in Matlab used for this experiment Houck et al. (1995).

7.3 Results and discussion

7.3.1 Results analysis

To see the effect of size of the dataset on the prediction performance, the datasets in the Tables 7.2 and 7.3 are sorted based on the number of data points from the largest (Magset A dataset) to the smallest (Pig dataset). The best and worst results for each dataset in the tables are shown in blue and red, respectively. As it can be seen from the MSLL results (Table 7.2), smooth box hyper-prior kernel works better than the other methods for 8 out of 11 datasets. In general, it shows a good performance for all-size datasets. In addition ION results (Table 7.3) also confirms that hyper-prior smooth-box results in better

prediction performances for majority of the datasets (6 out of 11 datasets). The results show that this algorithm produces good results independently of the performance measure. The inconsistency between the MSLL and ION results may be related to the small size of the datasets, because more investigation shows that the predictive variance which is part of *MSLL* (but not *ION*), could be so much variable in small datasets.

A notable thing from Table 7.3 is that using Evolutionary Algorithm (EA) to optimise the hyper-parameters works well for the larger datasets (human F, Mag-set A and ER sets) as they have the best performance among the other methods. However, as the size of the datasets decreases, the performance of EA measured by *ION* also decreases as it can be noted in Table 7.3. The worst ION performance predictions are obtained using the EA method in the 3 smallest datasets (pig, human C and mouse datasets). We have also initialised the population with the best hyper-parameters obtained from the grid search and unfortunately it did not change the results noticeably.

Second thing to be noted in the results is that the grid search and random search hyper-parameter optimisation methods have similar performance which confirms the results obtained from the research by Bergstra and Bengio (2012). Interestingly, these two methods are not the best methods to optimise the hyper-parameters. Most probably it is due to the limitation of these methods in searching three hyper-parameter spaces with $\mathcal{O}(m^3)$ which is expensive and limited to the certain number of hyper-parameters (in our case 20 values for each hyper-parameters which makes the $20\times20\times20$=8000 times). It seems that, changing a bit in the hyper-parameter values can lead to better results which is not always possible with grid / random search.

As we can see from the obtained results, using the hyper-prior optimisation method outperforms the Conjugate Gradient method. Figure 7.1 shows a comparison of these two techniques prediction performances (MSLL results used for this purpose). From this figure we can see that except from human C and pig datasets, for all the other datasets using hyper-prior smooth-box, a better or same prediction performance was obtained compared to conjugate gradient hyper-parameter selection method. A smaller standard deviation of MSLL is obtained when hyper-prior method was applied.

It it important to emphasise that the natures of the permeability measurements in datasets human C, B, mouse, rat and pig are different from the Magnusson datasets. So comparisons between performance of the models constructed from K_p and J_{max} might be limited especially from pharmaceutical point of view.

All the experiments have been performed with a windows system processor 3.6 GHz and 8 GB Ram and the time each experiment took was short for all the methods and datasets. It took between 2 to 7 seconds to run the experiments for the Conjugate gradient and all hyper-prior methods depending on the size of datasets. In addition the time taken to run the Evolutionary algorithm was between 60 to 100 seconds for different datasets and between 200 to 350 seconds for the grid/random search through the hyper-parameters space. The hyper-parameter values ranges obtained for each of the methods on each dataset show that almost all hyper-parameters are in the same range from 0.001 to 5 for the datasets human C, human F, mouse, rat, pig and ER datasets. Larger ranges from 5 to 15 are obtained for signal variance σ_f^2 of Magnusson datasets (Mag- sets A to E).

Table 7.2: MSLL performance using 11 datasets, hyper-parameter optimisation methods

Dataset	Grid search	Random search	Conjugate Gradient	Hyp-prior_Gauss	Hyp-prior_Lap	Hyp-prior_Smooth Box	Evolutionary_Alg
Mag-set A	-1.33±0.21	-1.32±0.02	-1.35±0.14	-0.97±0.06	-0.99±0.04	-1.35±0.10	-1.12±0.02
Human F	-0.22±0.35	-0.15±0.07	1.17±2.90	-0.16±0.07	-0.15±0.07	-0.27±0.10	-0.27±0.01
ER	-0.39±0.23	-0.34±0.03	-0.32±0.33	-0.29±0.30	-0.23±0.45	-0.39±0.27	-0.31±0.07
Mag-set B	-0.95±0.28	-0.98±0.02	-0.98±0.21	-0.56±0.14	-0.62±0.11	-0.99±0.18	-0.86±0.06
Mouse	0.07±0.56	0.74±0.48	0.72±0.86	-0.02±0.07	-0.06±0.11	-0.13±0.28	-0.13±0.01
Mag-set D	-0.22±0.22	-0.18±0.02	-0.18±0.19	-0.12±0.12	-0.23±0.21	-0.15±0.15	-0.18±0.01
Mag-set E	-0.30±0.61	-0.25±0.09	-0.43±0.33	-0.21±0.10	-0.35±0.23	-0.46±0.25	-0.46±0.02
Mag-set C	-0.20±0.80	-0.17±0.17	-0.20±0.82	-0.15±0.06	-0.12±0.43	-0.40±0.32	-0.32±0.04
Rat	-0.04±0.76	-0.10±0.14	-0.31±0.30	-0.11±0.07	-0.37±0.29	-0.43±0.19	0.16±0.15
Human C	-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	-0.98±0.37	-1.01±0.09	-0.90±0.36	-0.50±0.15	-0.93±0.43	-0.72±0.31	-0.05±0.42

Table 7.3: ION performance using 11 datasets, hyper-parameter optimisation methods

Dataset	Grid search	Random search	Conjugate Gradient	Hyp-prior_Gauss	Hyp-prior_Lap	Hyp-prior_Smooth Box	Evolutionary_Alg
Mag-set A	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 F	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
ER	0.52±0.20	0.49±0.04	0.46±0.27	0.45±0.18	0.43±0.24	0.54±0.20	0.54±0.03
Mag-set B	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	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
Mag-set D	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
Mag-set E	0.63±0.25	0.61±0.01	0.64±0.26	0.41±0.17	0.55±0.23	0.64±0.26	0.59±0.03
Mag-set C	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	0.10±0.58	0.08±0.04	0.24±0.25	0.31±0.21	0.29±0.34	0.40±0.20	-0.0±0.22
Human C	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	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

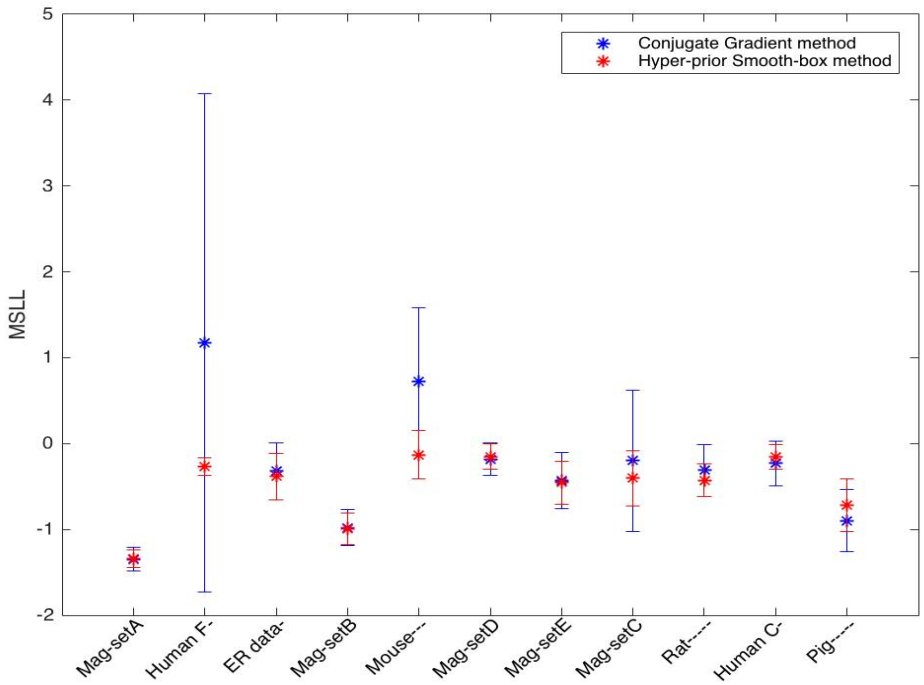


Figure 7.1: Comparing the MSLL performance using conjugate gradient and hyper-prior smooth-box optimisation methods (lower MSLL values show better performance of the models)

7.3.2 Data features analysis

To have a better visualisation on the features and their ranges in each dataset, the first 5 datasets with the same features (human C, human F, rat, mouse and pig sets) plus their log K_p ranges are plotted in box-plot separately in Figure 7.2. Similarly, the same features of the Magnusson datasets (Mag-sets A-E) and each dataset log J_{max} ranges are plotted together in Figure 7.3. Figure 7.4 shows all the features plus their target (ER_Q) value ranges in the ER dataset. From Tables 7.2 and 7.3, the best prediction performance among the first 5 datasets is for Pig which has the smallest number of data points (size=14). From the Figure 7.2, one can see that the Pig dataset covers a larger range of the target values

compared to the other datasets. Comparing Mag-set B and Mag-set D in Figure 7.3 also confirms the same relationship; as we can see with almost the same number of data points, Mag-set B covers a larger target range and its prediction performance is so much better than Mag-set D datasets. My hypothesis is that if the datasets that are used for training the model, cover a large range of target values, one can expect to obtain good prediction performance.

To investigate the effect of size and molecular feature ranges on the performance, another examination is performed in the following section.

7.3.3 Effect of size and chemical feature ranges on predictions

This experiment explores the effect of training data size and molecular feature ranges on prediction performances. In this investigation, Mag-set A is investigated as the largest datasets. It also performs better than the other datasets (using the hyper-prior smooth-box optimisation method). The aim of this experiment is observing the performance of the model by reducing the size of the dataset, while the range of molecular features are kept maximised.

To do this, in separate experiments the ranges of one important molecular feature (MW or $\log K_{ow}$) each time is kept as same as the Mag set-A (by keeping both minimum and maximum values for that feature). Then four different sized subsets are generated. The subsets sizes are varied from size=44 to size=9 based on the following steps:

- First experiment:
  - I use Mag-set A data and the aim is to choose different size data by keeping the MW range as maximum as possible. 4 different subsets are chosen randomly (sizes are 44, 33, 17 and 9). So the MW are selected in all ranges from minimum to maximum values.
  - The GP models are trained with the 4 datasets and the 5-fold cross validation is used to measure the prediction performance of the models. The hyper-prior-smoothbox optimisation method is used to set the best hyper-parameters of the model.

Table 7.4: ION and MSLL performance using 4 different size subsets from Mag-Set A, *MW* ranges are maximum

Performance/Subsets	Mag-set A	Subset 1	Subset 2	Subset 3	Subset 4
Size	85	44	33	17	9
ION	0.93	0.92	0.91	0.88	0.89
MSLL	-1.35	-1.20	-1.06	-0.88	-0.99

Table 7.5: ION and MSLL performance using 4 different size subsets from Mag-Set A, $\log K_{ow}$ ranges are maximum

Performance/Subsets	Mag-set A	Subset 1	Subset 2	Subset 3	Subset 4
Size	85	44	33	17	9
ION	0.93	0.90	0.93	0.94	0.72
MSLL	-1.35	-1.04	-1.1	-1.02	-0.98

- Second experiment:
  - The same as first experiment, but this time by considering maximum value ranges for $\log K_{ow}$ and 4 different size datasets (sized from 9 to 44) are obtained and the experiments are performed on the datasets similar to the first experiment.

The results are shown in Table 7.4 for the first experiment and Table 7.5 for the second experiment. Interestingly, the results from Table 7.4 show that decreasing the dataset Mag-Set A size, considering to keep the maximum range for *MW* do not affect the good performance of the model, especially in subsets 1 and 2 (with a slightly decrease in *ION* and *MSLL* performances). In the subsets 3 and 4 with sizes 17 and 9, respectively, the performances are more fallen (e.g. *ION* from 0.93 to 0.88 and 0.89 for subsets 3 and 4, respectively).

Similarly, the results in Table 7.5 confirm that reducing the dataset Mag-Set A size by keeping the maximum range for $\log K_{ow}$ do not affect the good performance of the model and it seems that there is even a small increase in the *ION* performance in the subset 3 (from 0.93 to 0.94). However, the model performance is declined with a larger proportion in subset 4 (*ION* form 0.93 to 0.72). Further investigation shows that the major difference between this subset and the subset 4 in the first experiment is the range of MW which is larger in subset 4 of the first experiment (maximised MW range). This shows that range of

the important features such as MW may affect the model's performance and they should be as large as possible in training sets.

7.4 Conclusion

The main finding of this study is that using hyper-prior smooth-box method to optimise the GPR hyper-parameters works good independently of the data and the performance measure method. Therefore, this approach is recommended to be used in this field.

In addition, investigation in the datasets features reveal that the range of the target values in the dataset can affect the model so that the larger the range of the target values in a dataset, the better prediction performance is obtained for the unseen data.

It is also important to consider the effect of size of the datasets on the model prediction performance. This study shows that decreasing the size of the datasets by keeping the chemical features maximum ranges, does not highly affect the performance of the model. In other words, to choose the training set for the model prediction, it is more important to pay attention choosing the wide ranges for important chemical features rather than gathering very large datasets which is difficult as discussed previously.

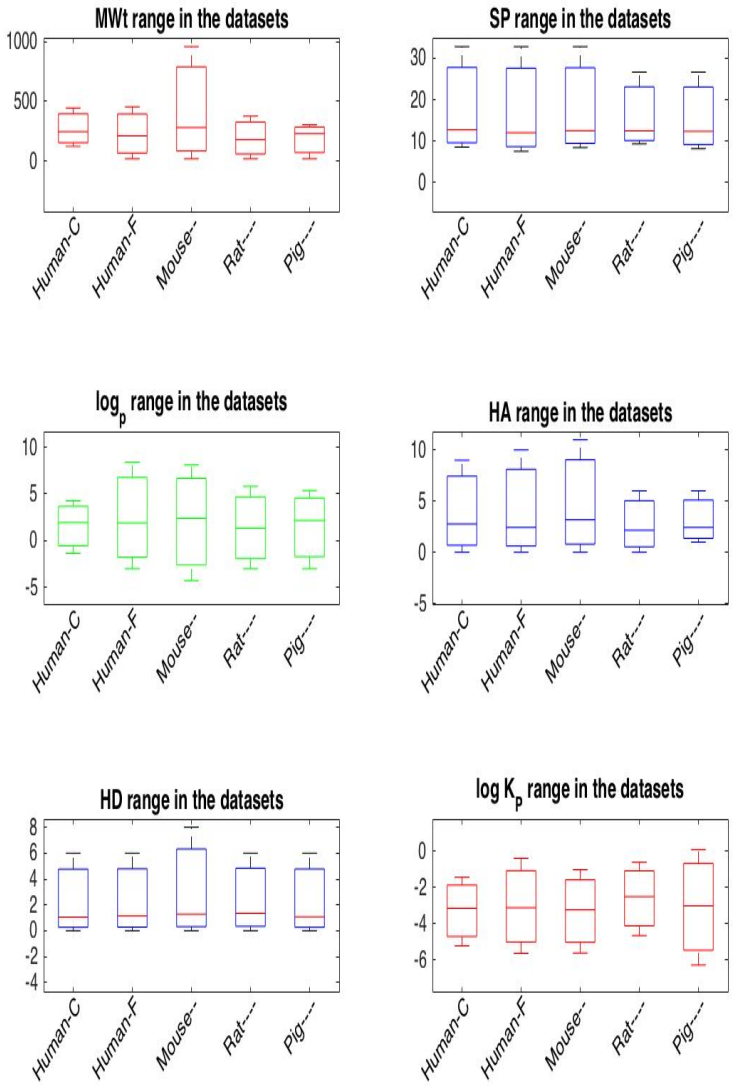


Figure 7.2: Range of features and targets in the human C, human F, Mouse, Rat and Pig datasets

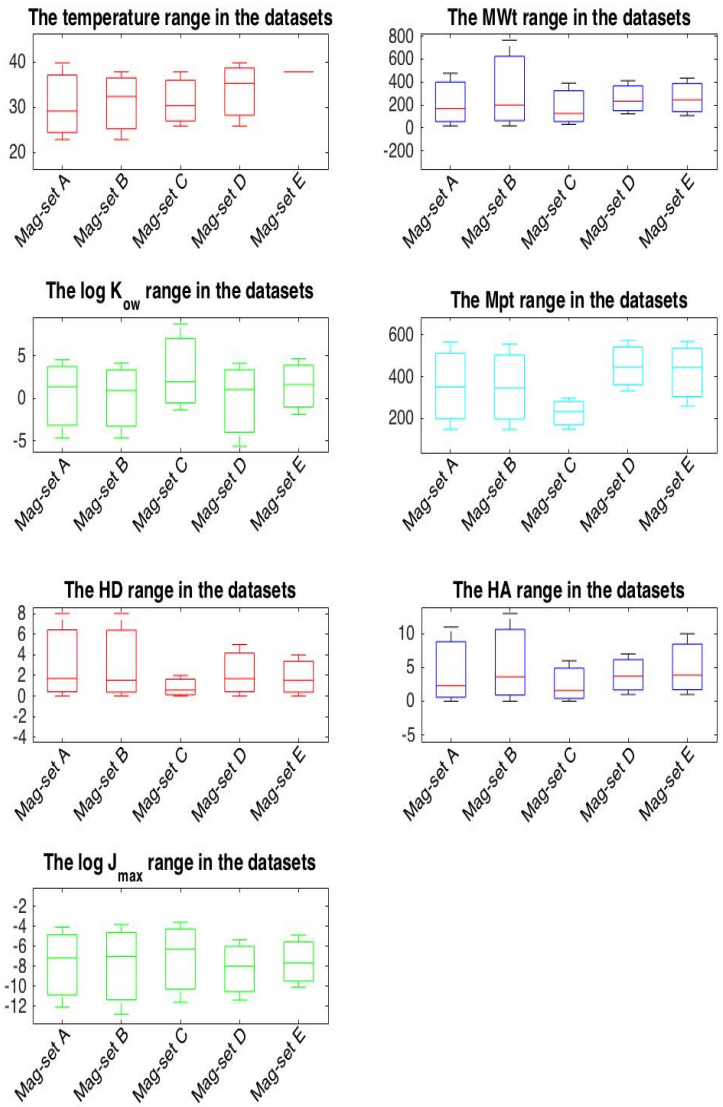


Figure 7.3: Range of features and targets in the Magnusson datasets

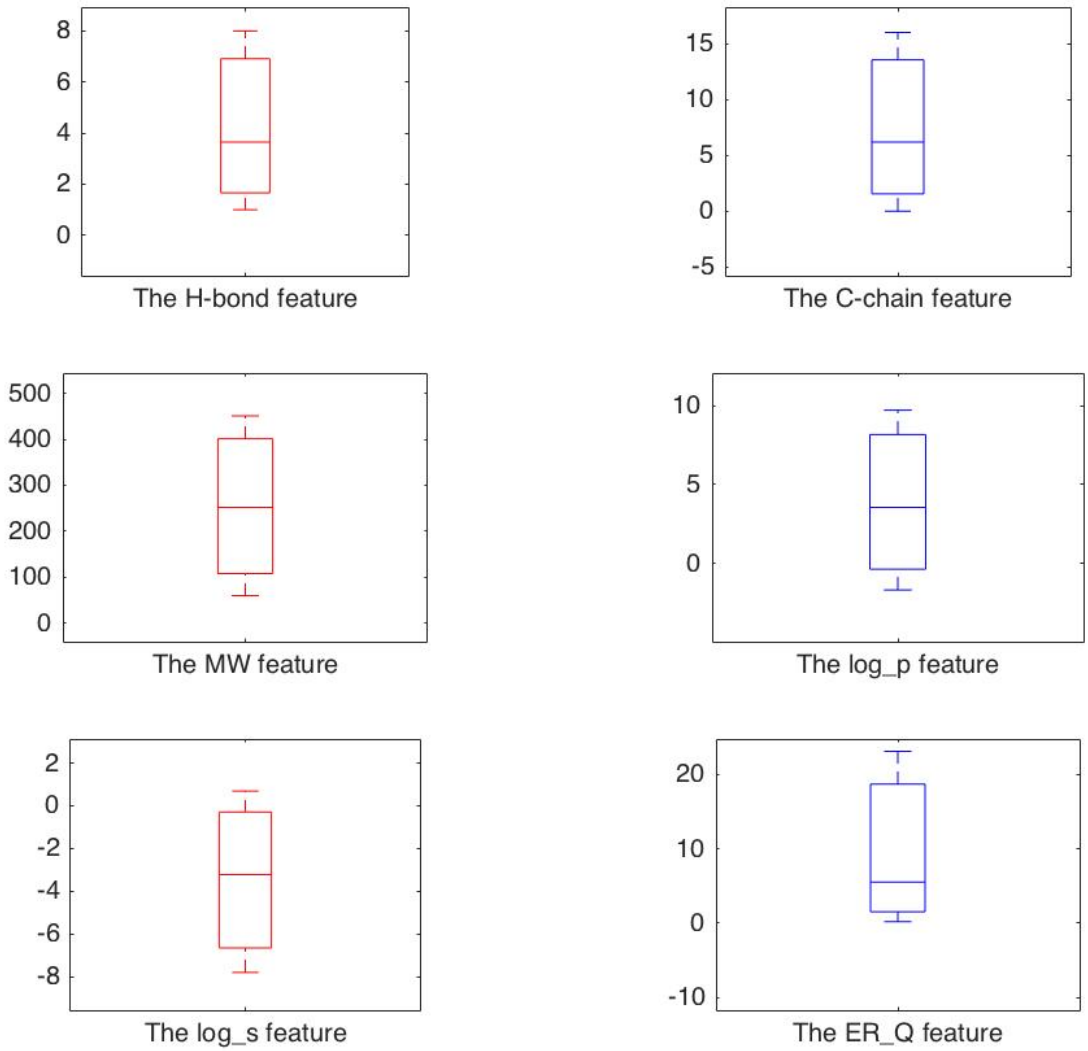


Figure 7.4: Range of features and targets in the ER dataset

Chapter 8

Conclusion and Future Work

This chapter discusses the major findings and contribution of all the work in this thesis. In addition, it demonstrates the future work. It is worth mentioning that the presented work is an interdisciplinary research between the Computer Science and Pharmacy. Bringing these two domains together and performing research on both sides has been a challenging experience.

8.1 Chapter summary

Chapter 2 of the thesis gives an overview of the skin structure and its functionalities. In addition, it specifies that human and animal skins consist of three main layers called the epidermis, the dermis and the subcutaneous layers from outside to inside. The outermost part of the epidermis is called the *stratum corneum* which works as the main barrier for the external chemicals to enter the skin (Williams (2003); Moss and Cronin (2002)). There are a number of pathways for passing the chemicals through the skin. It also discusses the importance of predicting the permeability of the chemicals through the skin in pharmaceutical and cosmetic industries. The permeability values are usually estimated based on the physicochemical properties of the compounds applied to the skin and the experimental conditions (e.g. temperature) under which the experiments have been performed in the labs. The most important pharmaceutical QSAR/QSPR approaches used for this purpose are explained and their performance are compared and evaluated (Flynn (1990); El Tayar

et al. (1991); Potts and Guy (1992); Moss and Cronin (2002)). Most of these models define a linear relationship between the permeability values and the chemical descriptors. The first reason for this is that physical scientists do not usually use computational models to obtain and evaluate the efficiency of non-linear methods and second, that linear equations are so easy to be used. However, visualising the datasets reveals that there is not a linear relationship between the permeability factors and the chemical features. Consequently, computational models and specifically machine learning algorithms are used and they outperform the QSAR/QSPR models which have been used in this domain (Sun et al. (2011); Moss et al. (2009)).

In Chapter 3 all the datasets employed for the experiments in this thesis are illustrated. These datasets consist of different sized human and animal datasets collated from various sources (Prapopoulou (2012); Moss and Cronin (2002); Flynn (1990)). There are 8 numerical features for each chemical including molecular descriptors and the temperature of the skin/environment at the time of the experiment. In addition, categorical features including the experimental conditions (nominal data) for each dataset are also explained.

In Chapter 4, all the computational methodologies applied to the datasets are explained. It includes Gaussian Processes, SVM, linear regression methods, KNN for regression, GNG clustering algorithm and performance measurement methods used in this thesis.

Chapter 5 explains the PCA visualisation method for both numerical and nominal data. In this chapter using PCA, the first and second principal components of both numerical and nominal features of dataset 'human D' are visualised and compared together. Three various methods are used to plot the PCA plot for categorical data. The first one is a novel method used by Wakelam et al. (2016), the second one uses Hamming distance Couto (2005) to generate the kernel and the last one employs the MCA method to obtain the principal components (Linting and van der Kooij (2012)). None of the PCA plots for nominal data shows any clear relationship with numerical data PCA plot and it is most probably because the number of categorical features are only limited to three.

Chapter 6 illustrates the major experiments (all the methods from Chapter 4) performed for this research. The major findings are listed as the following:

- It starts with Gaussian Process as the major regression model in this thesis. The GP is applied to the 6 human datasets using 5 and 7 numerical features. The results shows

that employing 5 molecular features *(MW, SP, logP, HA and HD)* performed better than 7 features in predicting the permeability rate of compounds through skin.

- In addition, using various covariance functions in GP with various parameters settings, the results confirm that GP with *Matérn* covariance funct*ion*, $\nu = 3/2$ outperforms the other methods in most of the human datasets.

- Comparing the GP prediction performances with the linear regression methods in including the QSAR and SLN techniques and KNN as a non linear regression algorithms illustrates that GP performs better than these methods in estimation the target values. However, SVM technique shows comparable performance to the GP and a mix of these two methods may be result in very good quality predictions (Shah et al. (2016)).

- In the next experiment of Chapter 6 the models are trained with one of the human, rat and mouse data and the targets are predicted for the other two datasets. The finding is that training the model with mouse data to estimate the human data permeability values, could result in better performance than training the model with rat data but it still does not perform better than using human datasets.

- Furthermore, the effect of experimental conditions such as temperature and cell types were examined. The results show that including temperature in the features improves the performance specially if the dataset contains a large range of temperature values. Additionally, the experiments show that the best predictive models are always obtained when static diffusion cell data permeability is predicted compared to models constructed from flow-through cell experiments.

- Continuing the visualisation of numerical and categorical data and to apply a novel covariance function in GP, a kernel is constructed based on both numerical (using Matérn covariance funct*ion*) and categorical data (using Hamming distance by Couto (2005)) and it could enhance the performance of the model slightly for a few datasets. As discussed in Chapter 5, using more categorical features (more than three that used for this experiment) may improve the performance of the model.

- Some of the datasets contain more than one target value (e.g. permeability coefficient) assigned to a single chemical and it is referred as data inconsistency. It is normal in biological systems like skin to take multiple readings and it is common to have variability on these measurements. This needs to be addressed in a way that not only suits the model but also to consider the requirements of both parties in a collaborative piece of work. Therefore, the final experiment in this Chapter, deals with the inevitable inconsistent data in the biological systems. To do so, the Monte Carlo method is applied to the complete human dataset and the prediction performance from the data states that this method can be used to generate a better training set. It can be applied to the datasets with inconsistency before the regression model is trained.

In Chapter 7 various hyper-parameter optmisation methods are applied to the 11 datasets to obtain the best setting for GP hyper-parameters. The results show that hyper-prior smooth-box method works well for most of the datasets independently of the data and the performance measure method and it can be used in further studies in this domain. The other finding of this chapter is that to choose the training set for the model, we should pay attention to use a dataset with a large ranges of important chemical features regardless the size of the data.

8.2 Contribution to knowledge

This thesis can be used as a source that compares various computational methods to estimate the target values (e.g. permeability) of compounds with chemical and experimental descriptors. It seeks to find the best technique to enhance the prediction performances with small datasets. The significant contribution that this thesis made to knowledge are:

- To predict the permeability of the compounds and similar applications, GP method using the hyper-prior smooth-box algorithm for selecting the suitable hyper-parameters shows so much better performance than the traditional linear QSAR models which are currently used in the pharmacy domain. GP models can be an outstanding replacement for QSAR/QSPR methods.

- In estimating the chemical's permeability including the experimental conditions such as temperature and diffusion cells as the model features shows to improve the efficiency of the model and they should be considered when the models are trained for regression purpose.
- The datasets that include inconsistent data, may cause problem in predicting the target values. This study confirms that Monte Carlo algorithm can generate better training set and can be used in similar studies with data inconsistency issue.
- Local non-linear models are obtained by clustering which can further improve the predictions.
- A new model including both numerical and categorical features is investigated and evaluated.
- Better experimental designs are obtained which allow the biological and physical scientists to work with computational models to build better models that more accurately reflect the aim of the models. This could result in producing better predictions and a better mechanistic understanding of the biological process.
- The efficiency of ML methods to predict the unseen chemical's permeation, show the faults and flaws in previous models.
- This work can be extended to other biological and environmental systems.

8.3 Future work

The future work to be performed after this research include:

- Using more datasets (e.g. the eye, environmental and language datasets) to evaluate the performance of the hyper-prior smooth-box hyper-parameter optimisation method.
- To include more categorical features and evaluate the performance of the model using the novel kernel function (considering both numerical and categorical kernels) in the GP.

- Using Monte Carlo algorithm suggested in Chapter 6 on more datasets and evaluate their prediction performances.
- In GNG clustering, more investigations may be performed on the compounds in each new cluster to be used in future studies in this domain.
- To apply the other methods rather than Hamming distance technique to obtain the categorical kernel. The other methods of combining the numerical and categorical kernels apart from weighted additive such as weighted multiplying can be also used to evaluate the model.
- The ML methods usually provide models as black box with 'lack of an equation'. Therefore, these techniques are mainly difficult to be employed by pharmaceutical and biological scientists. This issue is to be addressed in the further studies in this domain (e.g. by generating local effective polynomial equations/models).
- This study might have a very large impact on the pharmaceutical industry. However, applying statistical and computational methods to perform the regression (prediction), may be difficult to be employed by pharmacists and chemists. In order to use the methods suggested in this thesis, designing and implementing a web application for prediction purpose, could bring a large benefit in these domains. In the application, various skin data such as human and non-human datasets along with their known experimental conditions may be stored to be used to train the model to predict the required chemical's permeability through the skin. GP and SVM with optimised parameter settings can be employed to automatically predict the permeability for non-computer scientists.

Appendix A

Mathematical Concepts

A.1 Gaussian Process prior

In this section, I explain an example that helps to understand the GP prior. The below Gaussian Process is given with its mean and covariance function (inspired by an example in Rasmussen (2004)):

$$f \sim \mathscr{GP}(m,k), \text{ where } m(\mathbf{x}) = 0, \text{ and } k(\mathbf{x},\mathbf{x}^{'}) = exp(-\frac{1}{2}(\mathbf{x}-\mathbf{x}^{'})^2) \tag{A.1}$$

Drawing samples of function f could be helpful to understand the process. To be able to work with the finite number of quantities, the values of f should be obtained at a distinct finite number n of locations. Having the known $\mathbf{x}$ values, the vector of means and covariance function values could be evaluated and a regular Gaussian distribution is obtained:

$$\mu_i = m(\mathbf{x_i}) = 0, forall\, \mathbf{x_i}, i = 1...,n \tag{A.2}$$

$$\sum_{i,j} = k(\mathbf{x_i},\mathbf{x_j}) = exp(-\frac{1}{2}(\mathbf{x}_i-\mathbf{x}_j)^2),\ \text{i,j=1,...,n} \tag{A.3}$$

μ and $\sum$ are chosen to show the Gaussian distribution. Consequently, a random vector could be generated from this distribution and the function values $f(\mathbf{x})$ for corresponding $\mathbf{x}$'s are defined by :

$$f \sim \mathcal{N}(\mu, \Sigma) \tag{A.4}$$

As mentioned earlier, f is a generalisation of a probability Gaussian Process obtained from stochastic process. In this case it denotes the finite-dimensional random variable to functions. The plot of this distribution obtained by considering the values of f as the function of $\mathbf{x}$ and could be seen in Chapter 4 Figure 4.2. Here we have only used random functions and did not make inference about functions with given training examples. This example showed us how to use GP to define distributions over functions. The GP obtained here, can be used as a *prior* for Bayesian inference. As it is noted, the prior does not need to know anything about the training data. It only specifies some properties of the functions such as smoothness and the shape. The next step is to update this prior considering the training data (Rasmussen (2004)).

A.2 Conjugate definition

If A is a positive definite matrix, two non-zero vectors $\mathbf{u}$ and $\mathbf{v}$ are conjugate (with respect to A) if $\mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{v} = 0$, which means that conjugate vectors are orthogonal with respect to this inner product ($\langle \mathbf{u}, \mathbf{v} \rangle_\mathbf{A}$). Being conjugate is a symmetric relation: if $\mathbf{u}$ is conjugate to $\mathbf{v}$, then $\mathbf{v}$ is conjugate to $\mathbf{u}$.

If we have $P = \{p_1, ..., p_n\}$ as a set of n mutually conjugate vectors, then we may express the solution x_0 of $\mathbf{A}\mathbf{x} = \mathbf{b}$ on this basis:

$$x_* = \sum_{i=1}^{n} \alpha_i \mathbf{p}_i$$

Multiplying A and $\mathbf{p}_k^\mathsf{T}$ (for $k = i + 1$) to both sides of this equation we obtain:

$$\mathbf{A}x_* = \sum_{i=1}^{n} \alpha_i \mathbf{A}\mathbf{p}_i$$

$$\mathbf{p}_k^\mathsf{T}\mathbf{A}x_* = \sum_{i=1}^{n} \alpha_i \mathbf{p}_k^\mathsf{T}\mathbf{A}\mathbf{p}_i$$

$$\mathbf{p}_k^\top \mathbf{b} = \sum_{i=1}^{n} \alpha_i \langle \mathbf{p}_k, \mathbf{p}_i \rangle_{\mathbf{A}}$$

$$\langle \mathbf{p}_k, \mathbf{b} \rangle = \alpha_k \langle \mathbf{p}_k, \mathbf{p}_i \rangle_{\mathbf{A}}$$

Then we obtain each step α as:

$$\alpha_k = \frac{\langle \mathbf{p}_k, \mathbf{b} \rangle}{\langle \mathbf{p}_k, \mathbf{p}_i \rangle_{\mathbf{A}}}. \tag{A.5}$$

As it explained, the directions p_i are defined to be conjugate to each other. The next search direction should also be built out of the current residue and all previous search directions.

The following expression acquired from this assumption and also the conjugation constraint and Gram-Schmidt orthonormalisation (Shewchuk (1994)):

$$\mathbf{p}_k = \mathbf{r}_k - \sum_{i=1}^{n} \frac{\mathbf{p}_i^\top \mathbf{A} \mathbf{r}_k}{\mathbf{p}_i^\top \mathbf{A} \mathbf{p}_i} \mathbf{p}_i \tag{A.6}$$

Following this direction, the next optimal location is given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k p_k, \tag{A.7}$$

$$\alpha_k = \frac{\mathbf{p}_k^\top \mathbf{b}}{\mathbf{p}_k^\top \mathbf{A} \mathbf{p}_k} = \frac{\mathbf{p}_k^\top (\mathbf{r}_{k-1} + \mathbf{A}\mathbf{x}_{k-1})}{\mathbf{p}_k^\top \mathbf{A} \mathbf{p}_k} = \frac{\mathbf{p}_k^\top \mathbf{r}_{k-1}}{\mathbf{p}_k^\top \mathbf{A} \mathbf{p}_k}, \tag{A.8}$$

where holds the last equality, due to the reason that p_k and x_{k-1} are conjugate.

To calculate the iterations, we may initialise the input vector $x_0 = 0$ and $p_0 = r_0$ and continue to calculate the directions and steps values in each iteration (Shewchuk (1994)).

A.3 Using nonlinear PCA (in SPSS)

Nonlinear principal component analysis (NLPCA) (Linting and van der Kooij (2012)) is a method to explore the possible patterns and nonlinear relationships in the datasets. This technique can be used for all kind of the data and it is particularly useful for categorical

data. Similar to the PCA method, the main purpose of using NLPCA is to reduce the dimensionality of the data to a smaller number of uncorrelated variables (Principal components) which show the most information of the data. This is possible if a small number of linear combinations can be found that illustrate as much as possible of the variance in the data. Therefore, visualising the possible relational structures among the observed variables becomes possible.

The difference between PCA and NLPCA is that PCA can only recognises the linear relationships; however, NLPCA can define both linear and nonlinear relationships by quantifying nonlinearity related variables to be optimal for the PCA aim. To do so, PCA is applied to variables with numerical features and multiple correspondence analysis (MCA) to nominal variables. In order to perform this method, two models can be used called the vector model and centroid model, respectively. The vector model shows a variable as a straight line (vector), thus representing a variable as a direction in the component space, whereas the centroid model depicts a variable as a set of category points (centroids).

For dimension reduction of nominal data, MCA, first transforms the nominal variables to numeric values. To do this, optimal scaling technique is used. In this method, the centroid model is useful when one wants to find the location of the separate categories in the principal components space. Alternatively, one may be interested to examine the variable as a whole. That is when the vector model can be employed. Details of this method can be found in Linting and van der Kooij (2012). To perform this experiment, CATPCA in SPSS is used and the PCA plot is shown in Figure A.1. It can be seen that, similar to the other approaches, this technique does not also reveal a specific relationship between the numerical and nominal data.

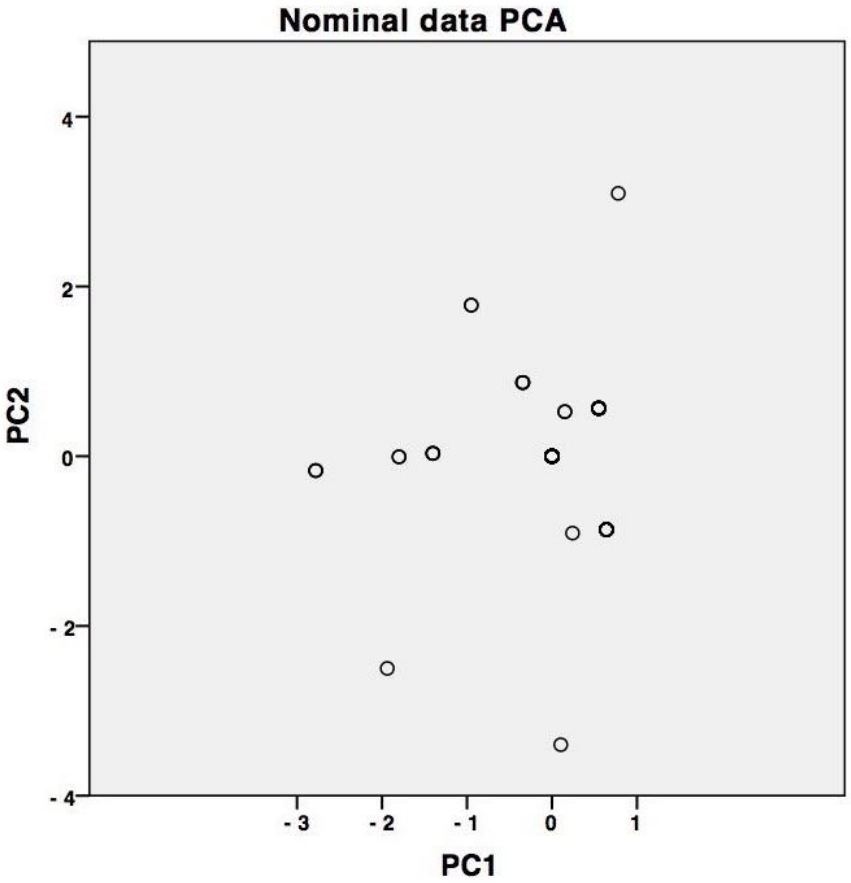


Figure A.1: The PCA plot of human dataset D (nominal data) , using NLPCA in CATPCA (Linting and van der Kooij (2012)).

Appendix B

All Datasets

B.1 Human datasets

B.1.1 Dataset human A

B.1.1.1 Experimental conditions

- Skin Thickness: 0.2-0.5 mm
- Site: Abdominal Cell
- type: Flow through cells
- Temperature: 32°C at the surface of the skin
- Time: of 24 – 48 h
- Vehicle: Water, physiological buffer, propylene glycol, ethanol, methanol
- Receptor Fluid: Normal saline

B.1.1.2 Data

Name	Mwt	MR	Mpt	SP	logP	Ha	HD	logKp	Temp	membrane	Site	Cell type
caffeine	194.2	49.2834	238	32.83	0.16	4	0	-3.68	32 (skin)	dermatomed 0.35 mm	abdomen	flow-through
caffeine	194.2	49.2834	238	32.83	0.16	4	0	-3.68	33 (skin)	dermatomed 0.35 mm	abdomen	flow-through
cannabinol	310.46	95.4483	76	10.9	7.23	2	1	-3.82	32 (skin)	dermatom. 0.2 mm	abdomen	flow-through
cannabidiol	314.46	97.9573	67	11.11	8.01	2	2	-3.62	32 (skin)	dermatom. 0.2 mm	abdomen	flow-through
Diazinon	304.35	25.9609	88	14.98	3.86	2	1	-2.02	37	dermatomed 0.5 mm	abdomen	flow-through
DDT	354.49	85.3157	108.50	9.45	6.79	0	%0	-2.23	37	dermatomed 0.5 mm	abdomen	flow-through
mannitol	182.17	38.4036	138.97	18.63	-3.01	6	6	-4.60	32	isol.epidermis	abdomen	flow-through
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-2.65	37	dermatomed 0.5 mm	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-3.74	32 (skin)	dermatom. 0.35 mm	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-2.17	32 (skin)	dermatom. 0.35 mm	abdomen	flow-through
Δ-Tetrahydrocannabinol (THC)	314.46	96.4453	304.5	9.75	7.6	2	1	-4.55	32 (skin)	dermatom. 0.2 mm	abdomen	flow-through

B.1.2 Dataset human B

B.1.2.1 Experimental conditions

- Skin Thickness: All types of thickness
- Site: All types
- type: Flow through cells
- Temperature: 32°C at the surface of the skin
- Time: of 24 – 72 h
- Vehicle: Water, physiological buffer, propylene glycol, ethanol, methanol
- Receptor Fluid: Normal saline

B.1.2.2 Data

Name	Mwt	MR	Mpt	SP	logP	Ha	HD	logKp	Temp	membrane	Site	Cell type

caffeine	194.2	49.2834	238	32.83	0.16	4	0	-3.68	32 (skin)	dermatomed 0,35 mm	abdomen	flow-through
caffeine	194.2	49.2834	238	32.83	0.16	4	0	-3.29	32 (skin)	dermatomed 0,35 mm	abdomen	flow-through
cannabinol	310.46	95.4483	76	10.9	7.23	2	1	-3.82	32 (skin)	dermatom. 0,2 mm	abdomen	flow-through
cannabidiol	314.46	97.9573	67	11.11	8.01	2	2	-3.62	32 (skin)	dermatom. 0,2 mm	abdomen	flow-through
Diazinon	304.35	25.9609	88	14.98	3.86	2	1	-2.02	37	dermatomed 0,5 mm	abdomen	flow-through
DDT	354.49	85.3157	108.50	9.45	6.79	0	0	-2.23	37	dermatomed 0,5 mm	abdomen	flow-through
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-2.65	37	dermatomed 0,5 mm	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-1.92	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-1.93	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-2.04	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-1.92	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-1.40	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-3.74	32 (skin)	dermatom. 0,35 mm	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-2.17	32 (skin)	dermatom. 0,35 mm	abdomen	flow-through

Δ-Tetrahydrocannabinol (THC)	314.46	96.4453	304.5	9.75	7.6	2	1	-4.55	32 (skin)	dermatom. 0,2 mm	abdomen	flow-through
isosorbide dinitrate (ISDN)	236.14	0	70	10.36	0.76	4	0	-2.35	37	SCE	abdomen	flow-through
nicorandil	211.18	0	92.5	14.40	0.43	3	1	-4.30	37	SCE	abdomen	flow-through
Bisphenol A diglycidyl ether (BADGE)	340.8	95.2347	10	10.38	3.84	4	0	-6.32	32(skin)	dermatomed skin	breast	flow-through
chlorpyrifos	350.59	0	42	10.10	4.66	1	0	-3.96	32(skin)	full-thickness	breast	flow-through
o-cresyl glycidyl ether (oCGE)	164.2	46.3071	30.25	10.38	2.16	2	0	-4.03	32 (skin)	dermatommed	breast	flow-through
dodecyl glycidyl ether (C12GE)	242.2	72.4384	58	8.41	5.01	2	0	-5.48	32 (skin)	dermatomed skin	breast	flow-through
1,6-hexanediol diglycidyl ether (HDDGE)	230.2	60.2566	84.83	9.36	0.84	4	0	-3.87	32(skin)	dermatomed skin	breast	flow-through
methiocarb	225.31	63.7027	119	10.92	2.87	1	1	-2.96	32 (skin)	dermatom. 0,6-0,9 mm	breast	flow-through

pirimicarb	238.29	66.9037	254.6	11.05	1.4	4	0	-2.57	32 (skin)	dermatom. 0,6-0,9 mm	breast	flow-through
prochloraz	376.7	91.2691	48	10.69	4.13	3	0	-3.15	32 (skin)	dermatom. 0,6-0,9 mm	breast	flow-through
benzoic acid	122.1	32.8164	122.4	11.94	1.87	1	1	-2.88	37	dermatomed 0,6 mm	not given	flow-through
boric acid	61.83	0	171	44.06	-0.22	3	3	-3.54	32(skin)	dermatomed 0,5 mm	thigh	flow-through
boric acid	61.83	0	171	44.06	-0.22	3	3	-5.85	32(skin)	dermatomed 0,5 mm	thigh	flow-through
boric acid	61.83	0	171	44.06	-0.22	3	3	-3.92	32(skin)	dermatomed 0,5 mm	thigh	flow-through
boric acid	61.83	0	171	44.06	-0.22	3	3	-3.30	32(skin)	dermatomed 0,5 mm	thigh	flow-through
coumarin	146.15	41.5486	70.6	11.91	1.51	1	0	-3.42	32 (skin)	full-thickness	breast	flow-through
coumarin	146.15	41.5486	70.6	11.91	1.51	1	0	-3.14	32 (skin)	full-thickness	breast	flow-through
glyphosate	169.10	0	230	12.73	-4.47	3	4	-3.34	37	dermatom. 1mm	not given	flow-through
malathion	330.36	0	2.85	10.61	2.29	2	0	-0.69	37	dermat. 1mm	not given	flow-through
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-4.60	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-3.49	37	dermatom. 0,45 mm	dorsal	flow-through

methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-3.54	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-3.27	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-2.53	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-2.38	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-2.32	37	dermatom. 0,32 mm	dorsal	flow-through
theophylline	180.17	44.3867	273	14.05	-0.39	4	1	-3.36	37	dermatomed 0,2 mm	dorsal	flow-through

B.1.3 Dataset human C

B.1.3.1 Experimental conditions

- Skin Thickness: All types of thickness
- Site: All types
- Cell type: static
- Temperature: 32°C at the surface of the skin
- Time: of 24 – 72 h
- Vehicle: Water, physiological buffer, propylene glycol, ethanol, methanol
- Receptor Fluid: Normal saline

B.1.3.2 Data

Name	Mwt	MR	Mpt	SP	logP	Ha	HD	logKp	Temp	membrane	Site	Cell type
benzoic acid	122.1	32.8164	122.4	11.94	1.87	1	1	-1.60	37	full-thickness	abdomen	static
caffeine	194.2	49.2834	238	32.83	0.16	4	0	-3.59	37	is.epidermis	upper leg	static
codeine	299.4	84.6037	155	12.09	1.28	4	1	-4.31	37	isol.epidermis	abdomen	static
coumarin	146.15	41.5486	70.6	11.91	1.51	1	0	-2.04	37	full-thickness	abdomen	static
coumarin	146.15	41.5486	70.6	11.91	1.51	1	0	-1.90	37	full-thickness	scalp	static
dichlofenac	296.16	75.4614	337.5	11.13	4.02	2	2	-3.00	37	full-thickness	abdomen	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-5.32	37	full-thickness	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-3.68	37	full-thickness	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-3.58	37	full-thickness	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-2.6	37	epidermis	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-2.6	37	epidermis	not given	static
etorphine	411.55	116.2256	215	11.76	3.02	5	2	-2.44	37	full-thickness	abdominal	static
griseofulvin	352.77	87.9616	220	10.44	1.92	6	0	-2.89	37	full-thickness	abdomen	static
griseofulvin	352.77	87.9616	220	10.44	1.92	6	0	-2.71	37	full-thickness	abdomen	static
hydromorphone	285.34	78.0905	266	10.96	1.6	4	1	-4.82	37	isol.epidermis	abdomen	static

ibuprofen	206.3	0	76	10.21	3.97	1	1	-1.44	37	full-thickness	abdomen	static
lindane	290.83	54.081	112.5	8.54	4.26	0	0	-5.23	32 (skin)	dermat. 0,23 mm	breast/abdomen	static
methyl nicotinate	137.14	35.4286	42.5	11.80	0.64	2	0	-2.41	37	full-thickness	abdomen	static
methyl parathion	263.21	2.75	35.5	10.45	2.75	1	0	-4.87	32(skin)	full-thickness	abdomen	static
morphine	285.3	79.8346	255	13.68	0.72	4	2	-5.03	37	isol.epidermis	abdomen	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.24	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.22	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-4.00	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-4.20	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.92	37	full-thickness	not given	static

N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.98	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-2.92	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.40	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-4.80	37	full-thickness	not given	static
naproxene	230.3	64.8535	153	11.42	3.1	2	1	-2.54	37	full-thickness	abdomen	static
nicorandil	211.18	0	92.5	14.40	0.43	3	1	-3.58	37	full-thickness	breast	static
nicotine	162.3	49.6542	-7.9	11.25	1	2	0	-1.99	32 (skin)	dermatom. 0,408 mm	abdomen/breast	static
nicotine	162.3	49.6542	-7.9	11.25	1	2	0	-2.48	37	full-thickness	abdomen	static
propoxur	209.25	56.3432	87	10.31	1.9	2	1	-4.54	37	full-thickness	abdomen	static
propoxur	209.25	56.3432	87	10.31	1.9	2	1	-1.39	37	full-thickness	abdomen	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-1.89	32 (skin)	dermatom. 0,408 mm	abdomen/breast	static

salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-1.86	37	full-thickness	abdomen	static
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-2.30	32 (skin)	dermatom. 0,408 mm	abdomen/breast	static

B.1.4 Dataset human D

B.1.4.1 Experimental conditions

- Skin Thickness: All types
- Site: All types
- Cell type: Flow through/Static
- Temperature: 37°C
- Time: Not specified
- Vehicle: Any
- Receptor Fluid: Any

B.1.4.2 Data

Name	Mwt	MR	Mpt	SP	logP	Ha	HD	logKp	Temp	membrane	Site	Cell type
2-ethoxyethanol	90	24.0528	-90	10.33	-0.42	2	1	-3.05	37	full-thickness	abdomen	flow-through
2-phenoxyethanol	138.17	38.8135	14	11.49	1.1	2	1	-2.87	37	dermat.0,33	breast/abdomen/leg	flow-through
2-phenylethanol	122.2	37.6285	-27	11.38	1.57	1	1	-1.51	37	epidermis	abdomen	static
2,4 dimethy-lamine	266.13	0	86	9.52	0.84	3	2	-3.02	37	dermatom.0,3 mm	abdomen	flow-through

2,4 dimethy-lamine	266.13	0	86	9.52	0.84	3	2	-3.09	37	dermatom.0,3 mm	abdomen	flow-through
4-n-butylaniline	149.24	49.6026	-14	9.51	3.1	1	1	-0.39	37	epidermis	abdomen	static
4-phenylbutanol	150	46.8305	16.20	10.8	2.55	1	1	-1.06	37	epidermis	abdomen	static
5-Fluorouracil	130.01	26.1222	281	13.46	-0.81	3	2	-3.22	37	epidermis	abdomen	static
Aniline	93.1	30.7584	-6.2	10.83	1.08	1	1	-1.21	37	epidermis	abdomen	static
b-estradiol	272.4	79.6175	173	11.9	3.94	2	2	-2.39	37	isol.epidermis	various	static
benzoic acid	122.1	32.8164	122.4	11.94	1.87	1	1	-1.60	37	full-thickness	abdomen	static
benzoic acid	122.1	32.8164	122.4	11.94	1.87	1	1	-2.88	37	dermatomed 0,6 mm	not given	flow-through
benzyl nicotinate	213.24	60.0412	24	11.55	2.35	2	0	-1.80	37	isol.epidermis	abdomen/breast	static
benzyl nicotinate	213.24	60.0412	24.00	11.55	2.35	2	0	-4.69	37	isol.epidermis		static
butyl paraben	194.23	53.1528	68.5	11.45	3.47	2	1	-1	37	epidermis	abdomen	static
caffeine	194.2	49.2834	238	32.83	0.16	4	0	-3.59	37	is.epidermis	upper leg	static
codeine	299.4	84.6037	155	12.09	1.28	4	1	-4.31	37	isol.epidermis	abdomen	static
coumarin	146.15	41.5486	70.6	11.91	1.51	1	0	-2.04	37	full-thickness	abdomen	static
coumarin	146.15	41.5486	70.6	11.91	1.51	1	0	-1.9	37	full-thickness	scalp	static
DDT	354.49	85.3157	108.5	9.45	6.79	0	0	-2.23	37	dermatomed 0,5 mm	abdomen	flow-through

Diazinon	304.35	25.9609	88	14.98	3.86	2	1	-2.02	37	dermatomed 0,5 mm	abdomen	flow-through
dichlofenac	296.16	75.4614	337.5	11.13	4.02	2	2	-3	37	full-thickness	abdomen	static
dimethylformamide	73.1	19.6669	-61	10.63	-0.93	1	0	-2.02	37	full-thickness	abdomen	flow-through
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-5.32	37	full-thickness	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-3.68	37	full-thickness	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-3.58	37	full-thickness	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-2.6	37	epidermis	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-2.6	37	epidermis	not given	static
ethylaniline	121.2	40.9998	-64	9.73	2.11	1	1	-0.54	37	epidermis	abdomen	static
etodolac	287.26	81.1565	146.5	10.86	3.93	3	2	-2.13	37	full-thickness	abdomen	static
etorphine	411.55	116.2256	215	11.76	3.02	5	2	-2.44	37	full-thickness	abdominal	static
famotidine	337.43	0	163.5	16.08	-0.64	5	4	-4.79	37	full-thickness	abdomen	franz
fentanyl	336.50	0.00	84	10.3	4.05	2	0	-2.25	37	isol.epidermis	abdomen	static
glyphosate	169.1	0	230	12.73	-4.47	3	4	-3.34	37	dermatom. 1mm	not given	flow-through
griseofulvin	352.77	87.9616	220	10.44	1.92	6	0	-2.89	37	full-thickness	abdomen	static
griseofulvin	352.77	87.9616	220	10.44	1.92	6	0	-2.71	37	full-thickness	abdomen	static

hydrocortisone	362.5	97.6308	220	12.75	1.61	5	3	-3.8	37	full-thickness	abdomen	flow-through
hydrocortisone	362.5	97.6308	220	12.75	1.61	5	3	-4.12	37	full-thickness	abdomen	flow-through
hydrocortisone	362.5	97.6308	220	12.75	1.61	5	3	-4.02	37	full-thickness	abdomen	flow-through
hydromorphone	285.34	78.0905	266	10.96	1.6	4	1	-4.82	37	isol.epidermis	abdomen	static
ibuprofen	206.3	0	76.00	10.21	3.97	1	1	-1.44	37	full-thickness	abdomen	static
isosorbide dinitrate	236.14	0	70	10.36	0.76	4	0	-2.35	37	SCE	abdomen	flow-through
ITF 296	238.00	0	55.5	12.91	1.85	3	0	-2.45	37	SCE	abdomen	flow-through
ketoprofen	254.29	0	94	11.75	3.12	2	1	-3.21	37	full-thickness	abdomen	static
lidocaine	234.34	72.1476	67	8.78	1.66	2	1	-3.46	37	dermat. 0,15 mm	leg	flow-through
lidocaine	234.34	72.1476	67	8.78	1.66	2	1	-3.41	37	dermat. 0,15 mm	leg	flow-through
lidocaine	234.34	72.1476	67	8.78	1.66	2	1	-1.97	37	dermat. 0,15 mm	leg	flow-through
linoleic acid	289.45	88.5188	-5	9.05	7.51	1	1	-4.97	37	full-thickness	abdomen	static
malathion	330.36	0	2.85	10.61	2.29	2	0	-0.69	37	dermat. 1mm	not given	flow-through
meperidine	247.4	72.4823	270	9.82	3.03	2	0	-2.43	37	isol.epidermis	abdomen	static
methotrexate	454.45	0	195	15.05	-1.28	10	5	-3.50	37	full-thickness	abdomen	static
methotrexate	454.45	0	195	15.05	-1.28	10	5	-3.95	37	full-thickness	abdomen	static

methotrexate	454.45	0	195	15.05	-1.28	10	5	-3.25	37	full-thickness	abdomen	static
methotrexate	454.45	0	195	15.05	-1.28	10	5	-2.95	37	full-thickness	abdomen	static
methotrexate	454.45	0	195	15.05	-1.28	10	5	-3.28	37	full-thickness	abdomen	static
methyl nicotinate	137.14	35.4286	42.5	11.80	0.64	2	0	-2.41	37	full-thickness	abdomen	static
methyl nicotinate	137.14	35.4286	42.5	11.80	0.64	2	0	-2.51	37	isol.epidermis	abdomen/breast	static
methyl nicotinate	137.14	35.4286	42.5	11.80	0.64	2	0	-2.47	37	isol.epidermis	abdomen/breast	static
methyl paraben	152.15	39.2796	131	12.50	2	2	1	-1.67	37	epidermis	abdomen	static
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-4.6	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-3.49	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-3.54	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-3.27	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-2.53	37	dermatom. 0,45 mm	dorsal	flow-through

methyl-4-hydroxy benzoate	152.14	39.2796	131	12.50	2	2	1	-2.38	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy benzoate	152.14	39.2796	131	12.5	2	2	1	-2.32	37	dermatom. 0,32 mm	dorsal	flow-through
morphine	285.3	79.8346	255	13.68	0.72	4	2	-5.03	37	isol.epidermis	abdomen	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-2.65	37	dermatomed 0,5 mm	abdomen	flow-through
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.24	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.22	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-4.00	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-4.20	37	full-thickness	not given	static

N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.92	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.98	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-2.92	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.40	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-4.80	37	full-thickness	not given	static
naproxene	230.3	64.8535	153	11.42	3.1	2	1	-2.54	37	full-thickness	abdomen	static
nicorandil	211.18	0	92.5	14.40	0.43	3	1	-3.58	37	full-thickness	breast	static
nicorandil	211.18	0	92.5	14.40	0.43	3	1	-4.30	37	SCE	abdomen	flow-through
nicotine	162.3	49.6542	-7.9	11.25	1	2	0	-2.48	37	full-thickness	abdomen	static
nicotinic acid	123.11	30.6595	236.6	13.23	0.69	2	1	-4.62	37	isol.epidermis	abdomen/breast	static

nimesulide	308.31	0	143	15.25	2.22	3	2	-2.995	37	full-thickness	abdomen	static
nizatidine	331.45	0	130	12.36	-0.43	5	2	-4.43	37	full-thickness	abdomen	static
parathion	291.26	0	6.1	10.76	3.73	1	0	-3.72	37	dermatom. 0,5 mm	back	flow-through
propoxur	209.25	56.3432	87	10.31	1.9	2	1	-4.54	37	full-thickness	abdomen	static
propoxur	209.25	56.3432	87	10.31	1.9	2	1	-1.39	37	full-thickness	abdomen	static
ranitidine	314.1	0	133	11.41	0.29	5	2	-4.05	37	full-thickness	abdomen	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-2.88	37	dermatom. 0,5 mm	breast	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-2.88	37	dermat. 0,5 mm	abdomen	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-1.86	37	full-thickness	abdomen	static
sufentanil	386.6	113.3917	97	10.47	3.62	3	0	-1.92	37	isol.epidermis	abdomen	static
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-1.92	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-1.93	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-2.04	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-1.92	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-1.4	37	full-thickness	abdomen	flow-through

theophylline	180.17	44.3867	273	14.05	-0.39	4	1	-3.36	37	dermatomed 0,2 mm	dorsal	flow-through
triclosan	289.55	68.4073	265.6	10.02	2.47	2	1	-4.47	37	isol.epidermis	breast and abdomen	static

B.1.5 Dataset human E

B.1.5.1 Experimental conditions

- Skin Thickness: All types
- Site: All types
- Cell type: Flow through
- Temperature: Not specified
- Time: Not specified
- Vehicle: Any
- Receptor Fluid: Any

B.1.5.2 Data

Name	**Mwt**	**MR**	**Mpt**	**SP**	**logP**	**Ha**	**HD**	**logKp**	**Temp**	**membrane**	**Site**	**Cell type**
atenolol	266.3	73.50	147	12.49	-0.03	4	3	-4.30	32	dermatomed 1,2 mm	abdominal	flow-through
benzoic acid	122.1	32.82	122.4	11.94	1.87	1	1	-2.88	37	dermatomed 0,6 mm	not given	flow-through
bisoprolol	325.5	92.15	100	10.01	1.84	5	2	-3.57	32	dermatom. 1,2 mm	abdominal	flow-through
Bisphenol A dig	340.8	95.23	10	10.38	3.84	4	0	-6.32	32 (skin)	dermatomed skin	breast	flow-through

boric acid	61.83	0.00	171	44.06	-0.22	3	3	-3.54	32 (skin)	dermatomed 0,5 mm	thigh	flow-through
boric acid	61.83	0.00	171	44.06	-0.22	3	3	-5.85	32 (skin)	dermatomed 0,5 mm	thigh	flow-through
boric acid	61.83	0.00	171	44.06	-0.22	3	3	-3.92	32 (skin)	dermatomed 0,5 mm	thigh	flow-through
boric acid	61.83	0.00	171	44.06	-0.22	3	3	-3.30	32 (skin)	dermatomed 0,5 mm	thigh	flow-through
caffeine	194.2	49.28	238	32.83	0.16	4	0	-3.14	32 (skin)	dermatomed 0,35 mm	abdomen	flow-through
caffeine	194.2	49.28	238	32.83	0.16	4	0	-7.39	32 (skin)	dermatomed 0,35 mm	abdomen	flow-through
caffeine	194.2	49.28	238	32.83	0.16	4	0	-3.68	32 (skin)	dermatomed 0,35 mm	abdomen	flow-through
caffeine	194.2	49.28	238	32.83	0.16	4	0	-3.29	32 (skin)	dermatomed 0,35 mm	abdomen	flow-through
cannabinol	310.46	95.45	76	10.9	7.23	2	1	-3.82	32 (skin)	dermatom. 0,2 mm	abdomen	flow-through
cannabidiol	314.46	97.96	67	11.11	8.01	2	2	-3.62	32 (skin)	dermatom. 0,2 mm	abdomen	flow-through
celiprolol	379.5	106.41	111	11.51	1.93	5	3	-3.23	32	dermatom. 1,2 mm	abdominal	flow-through
chlorpyrifos	350.59	0.00	42	10.10	4.66	1	0	-3.60	32 (skin)	full-thickness	breast	flow-through
chlorpyrifos	350.59	0.00	42	10.10	4.66	1	0	-3.96	32 (skin)	full-thickness	breast	flow-through
coumarin	146.15	41.55	70.6	11.91	1.51	1	0	-3.42	32 (skin)	full-thickness	breast	flow-through
coumarin	146.15	41.55	70.6	11.91	1.51	1	0	-3.14	32 (skin)	full-thickness	breast	flow-through

o-cresyl glycidyl	164.2	46.31	30.25	10.38	2.16	2	0	-4.03	32 (skin)	dermatommed	breast	flow-through
Diazinon	304.35	25.96	88	14.98	3.86	2	1	-2.02	37	dermatomed 0,5 mm	abdomen	flow-through
DDT	354.49	85.32	108.50	9.45	6.79	0	0	-2.23	37	dermatomed 0,5 mm	abdomen	flow-through
dimethylformamide	73.1	19.67	-61	10.63	-0.93	1	0	-2.02	37	full-thickness	abdomen	flow-through
2,4 dimethy-lamine	266.13	0.00	86	9.52	0.84	3	2	-3.02	37	dermatom.0,3 mm	abdomen	flow-through
2,4 dimethy-lamine	266.13	0.00	86	9.52	0.84	3	2	-3.09	37	dermatom.0,3 mm	abdomen	flow-through
dodecyl glycidyl	242.2	72.44	58	8.41	5.01	2	0	-5.48	32 (skin)	dermatomed skin	breast	flow-through
2-ethoxyethanol	90.12	24.05	-90	10.33	-0.42	2	1	-3.05	37	full-thickness	abdomen	flow-through
2-ethoxyethanol	90.12	24.05	-90	10.33	-0.42	2	1	-4.23	32	dermat. 0,28 mm	breast	flow-through
2-ethoxyethanol	90.12	24.05	-90	10.33	-0.42	2	1	-4.13	32	dermat. 0,28 mm	breast	flow-through
flufenamic acid	281.2	0.00	133	10.96	4.88	5	2	-3.28	32	sc 0,012 mm	abdomen	flow-through
flufenamic acid	281.2	0.00	133	10.96	4.88	5	2	-3.27	32	sc 0,012 mm	abdomen	flow-through
flufenamic acid	281.2	0.00	133	10.96	4.88	5	2	-3.41	32	sc 0,012 mm	abdomen	flow-through
flufenamic acid	281.2	0.00	133	10.96	4.88	5	2	-3.26	32	dermatom. 0,075 mm	abdomen	flow-through

flufenamic acid	281.2	0.00	133	10.96	4.88	5	2	-3.29	32	dermatom. 0,075 mm	abdomen	flow-through
flufenamic acid	281.2	0.00	133	10.96	4.88	5	2	-3.42	32	dermatom. 0,075 mm	abdomen	flow-through
flufenamic acid	281.2	0.00	133	10.96	4.88	5	2	-4.40	32	full-thickness	abdomen	flow-through
flufenamic acid	281.2	0.00	133	10.96	4.88	5	2	-4.03	32	full-thickness	abdomen	flow-through
flufenamic acid	281.2	0.00	133	10.96	4.88	5	2	-4.43	32	full-thickness	abdomen	flow-through
5-Fluorouracil	130.01	26.12	281	13.46	-0.81	3	2	-4.78	31	dermatomed 0,42 mm	abdomen	flow-through
glyphosate	169.10	0.00	230	12.73	-4.47	3	4	-3.34	37	dermatom. 1mm	not given	flow-through
1,6-hexanediol dig	230.2	60.26	84.83	9.36	0.84	4	0	-3.87	32 (skin)	dermatomed skin	breast	flow-through
hydrocortisone	362.5	97.63	220	12.75	1.61	5	3	-3.80	37	full-thickness	abdomen	flow-through
hydrocortisone	362.5	97.63	220	12.75	1.61	5	3	-4.12	37	full-thickness	abdomen	flow-through
hydrocortisone	362.5	97.63	220	12.75	1.61	5	3	-4.02	37	full-thickness	abdomen	flow-through
ITF 296	238	0.00	55.5	12.91	1.61	3	0	-2.45	37	SCE	abdomen	flow-through
isosorbide dinitrate	236.14	0.00	70	10.36	0.76	4	0	-2.35	37	SCE	abdomen	flow-through
lidocaine	234.34	72.15	67	8.78	1.66	2	1	-3.46	37	dermat. 0,15 mm	leg	flow-through

lidocaine	234.34	72.15	67	8.78	1.66	2	1	-3.41	37	dermat. 0,15 mm	leg	flow-through
lidocaine	234.34	72.15	67	8.78	1.66	2	1	-1.97	37	dermat. 0,15 mm	leg	flow-through
malathion	330.36	0.00	2.85	10.61	2.29	2	0	-0.69	37	dermat. 1mm	not given	flow-through
mannitol	182.17	38.40	138.97	18.63	-3.01	6	6	-4.60	32	isol.epidermis	abdomen	flow-through
mannitol	182.17	38.40	138.97	18.63	-3.01	6	6	-4.60	not given	dermat.0,13 mm	abdomen	flow-through
methiocarb	225.31	63.70	119	10.92	2.87	1	1	-2.96	32 (skin)	dermatom. 0,6-0,9 mm	breast	flow-through
methyl-4-hydroxy	152.14	39.28	131	12.50	2	2	1	-4.60	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy	152.14	39.28	131	12.50	2	2	1	-3.49	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy	152.14	39.28	131	12.50	2	2	1	-3.54	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy	152.14	39.28	131	12.50	2	2	1	-3.27	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy	152.14	39.28	131	12.50	2	2	1	-2.53	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy	152.14	39.28	131	12.50	2	2	1	-2.38	37	dermatom. 0,45 mm	dorsal	flow-through
methyl-4-hydroxy	152.14	39.28	131	12.50	2	2	1	-2.32	37	dermatom. 0,32 mm	dorsal	flow-through
1-methoxylpropan-2	90.123	23.72	-142	10.16	-0.49	2	1	-2.84	32	dermatom. 0,5 mm	breast	flow-through

1-methoxylpropan-2	90.123	23.72	-142	10.16	-0.49	2	1	-4.19	32	dermatom. 0,5 mm	breast	flow-through
1-methoxylpropan-2	90.123	23.72	-142	10.16	-0.49	2	1	-3.19	32	dermatom. 0,5 mm	breast	flow-through
metoprolol	267.4	76.70	124	10.39	1.69	4	2	-3.08	32	dermatomed 1,2 mm	abdominal	flow-through
N-N-Diethyl m-tolua	191.28	58.97	-45	10.70	2.26	1	0	-2.65	37	dermatomed 0,5 mm	abdomen	flow-through
Naltrexone (NTX)	341	91.33	166	13.75	1.39	5	2	-2.02	32	dermatomed 0,2 mm	abdominal	flow-through
NTX-3-acetate	383	100.77	114.00	12.96	1.47	5	1	-2.11	32	dermatomed 0,2 mm	abdominal	flow-through
NTX-3-propionate	397	105.39	147.00	12.06	1.96	5	1	-2.6	32	dermatomed 0,2 mm	abdominal	flow-through
NTX-3-butyrate	411	110.00	106.00	11.88	2.45	5	1	-2.89	32	dermatomed 0,2 mm	abdominal	flow-through
NTX-3-valerate	425	110.00	83.00	11.72	2.45	5	1	-3.15	32	dermatomed 0,2 mm	abdominal	flow-through
NTX-3-hexanoate	439	114.60	62.00	11.57	2.94	5	1	-2.92	32	dermatomed 0,2 mm	abdominal	flow-through
NTX-3 heptanoate	453	123.80	58.00	11.44	3.92	5	1	-3.05	32	dermatomed 0,2 mm	abdominal	flow-through
nicorandil	211.18	0.00	92.5	14.40	0.43	3	1	-4.30	37	SCE	abdomen	flow-through
oxprenolol	265.4	76.00	110	10.52	1.83	4	2	-2.81	32	dermatomed 1,2 mm	abdominal	flow-through

parathion	291.26	0.00	6.1	10.76	3.73	1	0	-3.72	37	dermatom. 0,5 mm	back	flow-through
2-phenoxyethanol	138.17	38.81	14	11.49	1.1	2	1	-2.87	37	dermat.0,33	breast/abdomen/leg	flow-through
pirimicarb	238.29	66.90	254.6	11.05	1.4	4	0	-2.57	32 (skin)	dermatom. 0,6-0,9 mm	breast	flow-through
prochloraz	376.7	91.27	48	10.69	4.13	3	0	-3.15	32 (skin)	dermatom. 0,6-0,9 mm	breast	flow-through
propranolol	295.3	76.82	166	11.13	2.6	3	2	-2.75	32	dermatomed 1,2 mm	abdominal	flow-through
testosterone	288.4	84.55	155	10.66	3.27	2	1	-1.92	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.55	155	10.66	3.27	2	1	-1.93	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.55	155	10.66	3.27	2	1	-2.04	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.55	155	10.66	3.27	2	1	-1.92	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.55	155	10.66	3.27	2	1	-1.40	37	full-thickness	abdomen	flow-through
testosterone	288.4	84.55	155	10.66	3.27	2	1	-4.57	32 (not given)	dermatom. 0,28 mm	breast	flow-through
testosterone	288.4	84.55	155	10.66	3.27	2	1	-3.55	32 (skin)	dermatom. 0,35 mm	abdomen	flow-through
testosterone	288.4	84.55	155	10.66	3.27	2	1	-3.74	32 (skin)	dermatom. 0,35 mm	abdomen	flow-through
testosterone	288.4	84.55	155	10.66	3.27	2	1	-2.17	32 (skin)	dermatom. 0,35 mm	abdomen	flow-through
testosterone	288.4	84.55	155	10.66	3.27	2	1	-3.62	27	dermatom. 0,6 mm	not given	flow-through

Δ-Tetrahydrocann	314.46	96.45	304.5	9.75	7.6	2	1	-4.55	32 (skin)	dermatom. 0,2 mm	abdomen	flow-through
theophylline	180.17	44.39	273	14.05	-0.39	4	1	-3.36	37	dermatomed 0,2 mm	dorsal	flow-through
water	18.02	0.00	0	26.68	-1.38	1	1	-2.89	31 (skin)	dermatomed 0,42 mm	abdomen	flow-through

B.1.6 Dataset human F

B.1.6.1 Experimental conditions

- Skin Thickness: All types
- Site: All types
- Cell type: Static
- Temperature: Not specified
- Time: Not specified
- Vehicle: Any
- Receptor Fluid: Any

B.1.6.2 Data

Name	**Mwt**	**MR**	**Mpt**	**SP**	**logP**	**Ha**	**HD**	**logKp**	**Temp *C***	**membrane**	**Site**	**Cell type**
water	18.02	0	0	26.68	-1.38	1	1	-3.12	Not given	full-thickness	abdomen	static
water	18.02	0	0	26.68	-1.38	1	1	-3.16	Not given	isol.epidermis	abdomen	static
water	18.02	0	0	26.68	-1.38	1	1	-3.04	Not given	isol.epidermis	abdomen	static
water	18.02	0	0	26.68	-1.38	1	1	-2.93	Not given	dermatomed 0,13 mm	abdomen	static

water	18.02	0	0	26.68	-1.38	1	1	-3.15	Not given	dermatomed 0,13 mm	abdomen	static
water	18.02	0	0	26.68	-1.38	1	1	-2.77	30 (not given)	dermatomed skin	abdomen	static
ethanol	46.07	13.0093	-114.1	10.92	-0.14	1	1	-3.5	30	full-thickness	abdomen	static
2-methoxyethanol	76.096	19.3048	-85.1	10.67	-0.91	2	1	-2.54	30(not given)	is.epidermis	abdomen	static
benzene	78.115	26.058	5.5	9.19	1.99	0	0	-0.95	31 (skin)	isol.epidermis	abdomen	static
benzene	78.115	26.058	5.5	9.19	1.99	0	0	-3.03	31 (skin)	isol.epidermis	abdomen	static
benzene	78.115	26.058	5.5	9.19	1.99	0	0	-0.78	31 (skin)	isol.epidermis	abdomen	static
benzene	78.115	26.058	5.5	9.19	1.99	0	0	-2.62	31 (skin)	isol.epidermis	abdomen	static
benzene	78.115	26.058	5.5	9.19	1.99	0	0	-3.96	31 (skin)	isol.epidermis	abdomen	static
2-ethoxyethanol	90.12	24.0528	-90	10.33	-0.42	2	1	-3.07	30 (not given)	full-thickness	abdomen	static
1-methoxylpropan-2-ol	90.123	23.723	-142	10.16	-0.49	2	1	-2.9	30 (not given)	isol.epidermis	abdomen	static
Aniline	93.1	30.7584	-6.2	10.83	1.08	1	1	-1.21	37	epidermis	abdomen	static
phenol	94.11	27.7521	40.9	12.33	1.51	1	1	-2.09	25 (not given)	is.epidermis	abdominal	static
phenol	94.11	27.7521	40.9	12.33	1.51	1	1	-3.83	22 (not given)	stratum corneum	abdomen	static
2-cresol (o-cresol)	108.1	32.7933	30.9	11.89	2.06	1	1	-1.8	25 (not given)	isolated epidermis	abdominal	static
4-cresol (p-cresol)	108.1	32.7933	33.00	11.89	2.06	1	1	-1.76	25 (not given)	isolated epidermis	abdominal	static
3-cresol (m-cresol)	108.14	32.7933	11.50	11.89	2.06	1	1	-1.82	25 (not given)	isolated epidermis	abdominal	static
hydroquinone	110.11	29.4462	170	15.18	1.03	2	2	-5.03	30	isol.epidermis	abdomen	static

resorcinol	110.11	29.4462	110	15.18	1.03	2	2	-3.62	25	is.epidermis	abdominal	static
squaric acid	114.06	25.027	293	20.47	-0.44	4	2	-5.12	not given	dermatom.	thigh	static
2-butoxyethanol	118.18	33.178	-70	9.88	0.57	2	1	-3.67	30 (not given)	is.epidermis	abdomen	static
2-(2-methoxyethoxy)ethanol	120.15	30.3483	-70	10.25	-1.18	3	1	-3.69	30(not given)	isol.epidermis	abdomen	static
ethylaniline	121.2	40.9998	-64	9.73	2.11	1	1	-0.54	37	epidermis	abdomen	static
benzoic acid	122.1	32.8164	122.4	11.94	1.87	1	1	-1.6	37	full-thickness	abdomen	static
2-phenylethanol	122.2	37.6285	-27	11.38	1.57	1	1	-1.51	37	epidermis	abdomen	static
nicotinic acid	123.11	30.6595	236.6	13.23	0.69	2	1	-4.62	37	isol.epidermis	abdomen/breast	static
nonane	128.3	43.2112	-53.5	7.51	4.76	0	0	-4.38	32 (skin)	dermatom.0,56 mm	back	static
5-Fluorouracil	130.01	26.1222	281	13.46	-0.81	3	2	-3.22	37	epidermis	abdomen	static
n-octanol	130.23	40.5385	-15.5	9.45	2.81	1	1	-1.21	22 (not given)	strarum corneum	abdomen	static
2-ethoxyethyl acetate	132.16	33.2043	-61.7	9.22	0.59	2	0	-3.09	30 (not given)	isol.epidermis	abdomen	static
n-nitrosodiethanolamine	134.13	32.6351	81.52	15.67	-1.28	4	2	-2.39	32 (skin)	isol.epidermis	abdomen	static
2-(2-ethoxyethoxy)ethanol	134	35.0963	-76	10.04	-0.69	3	1	-3.88	30 (not given)	isol.epidermis	abdomen	static
methyl nicotinate	137.14	35.4286	42.5	11.8	0.64	2	0	-2.41	37	full-thickness	abdomen	static
methyl nicotinate	137.14	35.4286	42.5	11.8	0.64	2	0	-2.51	37	isol.epidermis	abdomen/breast	static

methyl nicotinate	137.14	35.4286	42.5	11.8	0.64	2	0	-2.47	37	isol.epidermis	abdomen/breast	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	0.34	32	dermatom. 0,6 mm	abdomen	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-1.89	32 (skin)	dermatom. 0,408 mm	abdomen/breast	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-2.88	37	dermatom. 0,5 mm	breast	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-1.92	25	full-thickness	breast	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-3.13	25	full-thickness	breast	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-2.88	37	dermat. 0,5 mm	abdomen	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-1.86	37	full-thickness	abdomen	static
3-nitrophenol	139.1	0	98.00	13.02	1.91	2	1	-2.25	25 (not given)	isol.epidermis	abdomen	static
2-naphthol	144.16	0	123	12.69	2.69	1	1	-1.55	25 (not given)	isol.epidermis	abdomen	static
coumarin	146.15	41.5486	70.6	11.91	1.51	1	0	-2.04	37	full-thickness	abdomen	static
coumarin	146.15	41.5486	70.6	11.91	1.51	1	0	-1.90	37	full-thickness	scalp	static
4-n-butylaniline	149.24	49.6026	-14	9.51	3.1	1	1	-0.39	37	epidermis	abdomen	static
4-phenylbutanol	150	46.8305	16.2	10.8	2.55	1	1	-1.06	37	epidermis	abdomen	static
thymol	150.2	46.9841	49	10.81	3.52	1	1	-1.28	25 (not given)	is.epidermis	abdominal	static

ethyl nicotinate	151.17	40.1766	-8.5	11.06	1.13	2	2	-2.22	37(not given)	isol.epidermis	abdomen/breast	static
ethyl nicotinate	151.17	40.1766	-8.5	11.06	1.13	2	2	-2.18	37(not given)	isol.epidermis	abdomen/breast	static
ethyl nicotinate	151.17	40.1766	-8.5	11.06	1.13	2	2	-3.65	37(not given)	isol.epidermis	abdomen/breast	static
methyl paraben	152.15	39.2796	131	12.50	2	2	1	-2.38	23	epidermis	abdomen	static
methyl paraben	152.15	39.2796	131	12.50	2	2	1	-2.03	30	epidermis	abdomen	static
methyl paraben	152.15	39.2796	131	12.50	2	2	1	-1.67	37	epidermis	abdomen	static
methyl paraben	152.15	39.2796	131	12.50	2	2	1	-1.49	45	epidermis	abdomen	static
chloroxylenol	156.6	42.6393	115	11.20	3.25	1	1	-1.23	25	isol.epidermis	abdominal	static
2-(2-butoxyethoxy)ethanol	162.23	44.2215	-68	9.74	0.29	3	1	-4.45	30 (not given)	is.epidermis	abdomen	static
nicotine	162.3	49.6542	-7.9	11.25	1	2	0	-1.99	32 (skin)	dermatom. 0,408 mm	abdomen/breast	static
nicotine	162.3	49.6542	-7.9	11.25	1	2	0	-2.48	37	full-thickness	abdomen	static
diethyl squarate	170.16	44.0254	131.55	11.5	4.07	4	0	-3.92	not given	dermatomed skin	thigh	static
2-phenylphenol	170.21	52.8883	116.1	12.24	3.28	1	1	-1.58	32	is.epidermis	dorsal/flank	static
2-phenylphenol	170.21	52.8883	116.1	12.24	3.28	1	1	-2.8	32	full-thickness	abdomen	static
2-phenylphenol	170.21	52.8883	116.10	12.24	3.28	1	1	-1.74	32	is.epidermis	abdomen	static

4-bromophenol	173.01	35.3749	66.40	11.48	2.4	1	1	-1.44	25 (not given)	is.epidermis	abdominal	static
butyl nicotinate	179.22	49.3018	69.59	10.57	2.11	2	0	-1.78	37 (not given)	is.epidermis	abdomen/breast	static
butyl nicotinate	179.22	49.3018	69.59	10.57	2.11	2	0	-4.1	37 (not given)	is.epidermis	abdomen/breast	static
mannitol	182.17	38.4036	138.97	18.63	-3.01	6	6	-4.96	not given	full-thickness	abdomen	static
mannitol	182.17	38.4036	138.97	18.63	-3.01	6	6	-4.96	not given	isol.epidermis	abdomen	static
mannitol	182.17	38.4036	138.97	18.63	-3.01	6	6	-4.21	30	full-thickness	abdomen	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.24	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.22	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.7	2.26	1	0	-4	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.7	2.26	1	0	-4.2	37	full-thickness	not given	static

N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-3.92	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.7	2.26	1	0	-3.98	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.7	2.26	1	0	-2.92	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.7	2.26	1	0	-3.4	37	full-thickness	not given	static
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.7	2.26	1	0	-4.8	37	full-thickness	not given	static
caffeine	194.20	49.2834	238	32.83	0.16	4	0	-3.59	37	is.epidermis	upper leg	static
butyl paraben	194.23	53.1528	68.5	11.45	3.47	2	1	-1.56	23	epidermis	abdomen	static
butyl paraben	194.23	53.1528	68.5	11.45	3.47	2	1	-1.25	30	epidermis	abdomen	static
butyl paraben	194.23	53.1528	68.5	11.45	3.47	2	1	-1	37	epidermis	abdomen	static
butyl paraben	194.23	53.1528	68.5	11.45	3.47	2	1	-0.56	45	epidermis	abdomen	static
3,4 xylenol	202.55	37.8345	62.50	11.54	2.61	1	1	-1.44	25	isol.epidermis	abdominal	static

propoxur	209.25	56.3432	87	10.31	1.9	2	1	-4.54	37	full-thickness	abdomen	static
propoxur	209.25	56.3432	87	10.31	1.9	2	1	-1.39	37	full-thickness	abdomen	static
propoxur	209.25	56.3432	87	10.31	1.9	2	1	-3.05	32	full-thickness	abdomen	static
nicorandil	211.18	0	92.5	14.4	0.43	3	1	-3.58	37	full-thickness	breast	static
benzyl nicotinate	213.24	60.0412	24	11.55	2.35	2	0	-1.8	37	isol.epidermis	abdomen/breast	static
benzyl nicotinate	213.24	60.0412	24	11.55	2.35	2	0	-4.69	37	isol.epidermis	not given	static
DEP	222.24	58.609	-3	10.51	2.65	2	0	-4.94	30 (not given)	isol.epidermis	abdomen	static
dibutyl squarate	226.27	62.2758	176.71	10.58	2.45	4	0	-4.70	not given	dermatomed skin	thigh	static
naproxene	230.3	64.8535	153	11.42	3.1	2	1	-2.54	37	full-thickness	abdomen	static
meperidine	247.4	72.4823	270	9.82	3.03	2	0	-2.43	37	isol.epidermis	abdomen	static
paraquat	257.16	0	300	10.45	-2.71	0	0	-5.06	30	full-thickness	abdomen	static
methyl parathion	263.21	2.75	35.5	10.45	2.75	1	0	-4.87	32 (skin)	full-thickness	abdomen	static
methyl parathion	263.21	2.75	35.5	10.45	2.75	1	0	-4.42	32 (skin)	full-thickness	abdomen	static
estrone	270.4	78.7956	254.5	11.55	3.43	2	1	-2.44	26 (not given)	isol.epidermis	not given	static
b-estradiol	272.4	79.6175	173	11.9	3.94	2	2	-3.52	26 (not given)	isol.epidermis	not given	static

b-estradiol	272.4	79.6175	173	11.9	3.94	2	2	-3	32	dermatomed 0,5 mm	abdomen	static
b-estradiol	272.4	79.6175	173	11.9	3.94	2	2	-2.00	32	dermatomed 0,5 mm	abdomen	static
b-estradiol	272.4	79.6175	173	11.9	3.94	2	2	-2.39	37	isol.epidermis	various	static
dibutylphthalate	278.35	0	-35	10.86	5.11	2	0	-5.64	30 (not given)	isol.epidermis	abdomen	static
morphine	285.3	79.8346	255	13.68	0.72	4	2	-5.03	37	isol.epidermis	abdomen	static
hydromorphone	285.34	78.0905	266	10.96	1.6	4	1	-4.82	37	isol.epidermis	abdomen	static
etodolac	287.26	81.1565	146.5	10.86	3.93	3	2	-2.13	37	full-thickness	abdomen	static
estriol	288.4	80.979	282	12.95	2.81	3	3	-4.4	26 (not given)	isol.epidermis	not given	static
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-3.92	32	full-thickness	abdomen	static
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-2.30	32 (skin)	dermatom. 0,408 mm	abdomen/breast	static
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-3.4	26 (not given)	is.epidermis	not given	static
linoleic acid	289.45	88.5188	-5	9.05	7.51	1	1	-4.97	37	full-thickness	abdomen	static
triclosan	289.55	68.4073	265.6	10.02	2.47	2	1	-4.47	37	isol.epidermis	breast and abdomen	static
lindane	290.83	54.081	112.5	8.54	4.26	0	0	-5.23	32 (skin)	dermat. 0,23 mm	breast/abdomen	static
lindane	290.83	54.081	112.5	8.54	4.26	0	0	-5.23	32 (skin)	dermat. 0,23 mm	breast/abdomen	static
dichlofenac	296.16	75.4614	337.5	11.13	4.02	2	2	-3	37	full-thickness	abdomen	static
codeine	299.4	84.6037	155	12.09	1.28	4	1	-4.31	37	isol.epidermis	abdomen	static

nimesulide	308.31	0	143	15.25	2.22	3	2	-2.995	37	full-thickness	abdomen	static
ranitidine	314.1	0	133	11.41	0.29	5	2	-4.05	37	full-thickness	abdomen	static
progesterone	314.5	92.8212	121	10.05	3.67	2	0	-2.82	26 (not given)	is.epidermis	not given	static
progesterone	314.5	92.8212	121	10.05	3.67	2	0	-1.52	37 (not given)	is.epidermis	various	static
pregnenolone	316.5	93.7559	192	10.36	3.89	2	1	-2.82	26 (not given)	is.epidermis	not given	static
nizatidine	331.45	0	130	12.36	-0.43	5	2	-4.43	37	full-thickness	abdomen	static
clotrimazole	344.85	102.1434	148	11.17	6.26	2	0	-2.70	32	dermatom. 0,6 mm	abdomen	static
cortexolone	346.47	96.0389	208	11.91	3.15	4	2	-4.12	26 (not given)	isol.epidermis	not given	static
corticosterone	346.5	96.2312	183	11.91	1.99	4	2	-4.22	26 (not given)	isol.epidermis	not given	static
griseofulvin	352.77	87.9616	220	10.44	1.92	6	0	-2.89	37	full-thickness	abdomen	static
griseofulvin	352.77	87.9616	220	10.44	1.92	6	0	-2.71	37	full-thickness	abdomen	static
Aldosterone	360.45	96.4918	164	12.31	1.63	4	2	-5.52	26 (not given)	isol.epidermis	not given	static
Aldosterone	360.45	96.4918	164	12.31	1.63	4	2	-4.24	26 (not given)	isol.epidermis	various	static
cortisone	360.5	96.7062	220	12.1	1.81	5	2	-5	26 (not given)	isol.epidermis	not given	static
sufentanil	386.6	113.3917	97	10.47	3.62	3	0	-1.92	37	isol.epidermis	abdomen	static

DEHP	390.57	113.409	-50	9.39	8.39	2	0	-5.24	30 (not given)	isol.epidermis	abdomen	static
etorphine	411.55	116.2256	215	11.76	3.02	5	2	-2.44	37	full-thickness	abdominal	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-5.32	37	full-thickness	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-3.68	37	full-thickness	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-3.58	37	full-thickness	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-2.6	37	epidermis	not given	static
doxycycline HCL	444.44	113.7208	866	16.55	-1.36	9	6	-2.6	37	epidermis	not given	static
methotrexate	454.45	0	195	15.05	-1.28	10	5	-3.5	37	full-thickness	abdomen	static
methotrexate	454.45	0	195	15.05	-1.28	10	5	-3.95	37	full-thickness	abdomen	static
methotrexate	454.45	0	195	15.05	-1.28	10	5	-3.25	37	full-thickness	abdomen	static
methotrexate	454.45	0	195	15.05	-1.28	10	5	-2.95	37	full-thickness	abdomen	static
methotrexate	454.45	0	195	15.05	-1.28	10	5	-3.28	37	full-thickness	abdomen	static

B.2 Animal datasets

B.2.1 Mouse dataset

B.2.1.1 Experimental conditions

- **Skin thickness: Full thickness**
- **Site: All body cites**
- **Cell type: Flow-through/Static**
- **Temperature:** $25^{\circ}C$ **and** $37^{\circ}C$
- **Time: Not specific**
- **Vehicle: Any**
- **Receptor fluid: Mainly Saline**

B.2.1.2 Data

Name	Mwt	MR	Mpt	SP	logP	Ha	HD	logKp	Temp C	membrane	Site	Cell type
Alachlor	269.77	73.9275	40	9.80	3.37	2	0	-3.42	not given	full-thickness	dorsal	flow-through
Alachlor	269.77	73.9275	40	9.80	3.37	2	0	-3.18	not given	full-thickness	dorsal	flow-through
atenolol	266.3	73.5041	147	12.49	-0.03	4	3	-1.32	37	full-thickness	abdomen	flow-through
atenolol	266.3	73.5041	147	12.49	-0.03	4	3	-1.32	37	full-thickness	dorsal	flow-through
atenolol	266.3	73.5041	147	12.49	-0.03	4	3	-1.52	37	full-thickness	abdomen	flow-through
atenolol	266.3	73.5041	147	12.49	-0.03	4	3	-1.60	37	full-thickness	dorsal	flow-through

atenolol	266.3	73.5041	147	12.49	-0.03	4	3	-1.45	37	full-thickness	abdomen	flow-through
atenolol	266.3	73.5041	147	12.49	-0.03	4	3	-1.67	37	full-thickness	dorsal	flow-through
Atrazine	215.69	62.214	175	11.77	2.82	5	2	-3.11	not given	full-thickness	dorsal	flow-through
Atrazine	215.69	62.214	175	11.77	2.82	5	2	-3.77	not given	full-thickness	dorsal	flow-through
Bisphenol A diglycidyl ether	340.8	95.2347	10	10.38	3.84	4	0	-5.07	not given	full-thickness	dorsal	flow-through
n-butanol	74.14	22.1345	-89.8	10.13	0.84	1	1	-2.19	not given	full-thickness	abdomen/dorsal	static
n-butanol	74.14	22.1345	-89.8	10.13	0.84	1	1	-2.09	not given	full-thickness	abdomen/dorsal	static
n-butanol	74.14	22.1345	-89.8	10.13	0.84	1	1	-1.93	not given	full-thickness	abdomen/dorsal	static
n-butanol	74.14	22.1345	-89.8	10.13	0.84	1	1	-1.91	not given	full-thickness	abdomen/dorsal	static
n-butanol	74.14	22.1345	-89.8	10.13	0.84	1	1	-1.92	not given	full-thickness	abdomen/dorsal	static
n-butanol	74.14	22.1345	-89.8	10.13	0.84	1	1	-1.94	not given	full-thickness	abdomen/dorsal	static
n-butanol	74.14	22.1345	-89.8	10.13	0.84	1	1	-1.90	not given	full-thickness	abdomen	static
n-butanol	74.14	22.1345	-89.8	10.13	0.84	1	1	-1.63	not given	full-thickness	back	static
caffeine	194.2	49.2834	238	32.83	0.16	4	0	-3.59	37	full-thickness	dorsal/abdominal	static

corticosterone	346.5	96.2312	183	11.91	1.99	4	2	-3.28	37	full-thickness	abdomen	static
coumarin	146.15	41.5486	70.60	11.91	1.51	1	0	-3.00	32 (skin)	full-thickness	dorsal	flow-through
o-cresyl glycidyl ether (oCGE)	164.20	46.3071	34.98	10.38	2.16	2	0	-3.75	32 (skin)	full-thickness	dorsal	flow-through
decabromodiphenyl oxide (DBDPO)	959.17	128.5268	302.5	12.11	8.10	1	0	-5.93	not given	full-thickness	dorsal	flow-through
decabromodiphenyl oxide (DBDPO)	959.17	128.5268	302.5	12.11	8.10	1	0	-5.79	not given	full-thickness	dorsal	flow-through
decabromodiphenyl oxide (DBDPO)	959.17	128.5268	302.5	12.11	8.10	1	0	-5.15	not given	full-thickness	dorsal	flow-through
deoxycortisone	330.47	94.6393	141.5	11.03	3.12	3	1	-2.47	37	full-thickness	abdomen	static
dibutyl squarate	226.27	62.2758	176.71	10.58	2.45	4	0	-3.07	not given	full-thickness	not given	static
diethyl squarate	170.16	44.0254	131.55	11.50	4.07	4	0	-3.00	not given	full-thickness	not given	static
dodecyl glycidyl ether (C12GE)	242.2	72.4384	58	8.41	5.01	2	0	-4.34	32 (skin)	full-thickness	dorsal	flow-through
ethanol	46.07	13.0093	-114.1	10.92	-0.14	1	1	-2.68	not given	full-thickness	abdomen/dorsal	static

ethanol	46.07	13.0093	-114.1	10.92	-0.14	1	1	-2.66	not given	full-thickness	abdomen/dorsal	static
ethanol	46.07	13.0093	-114.1	10.92	-0.14	1	1	-2.64	not given	full-thickness	abdomen/dorsal	static
ethanol	46.07	13.0093	-114.1	10.92	-0.14	1	1	2.66	not given	full-thickness	abdomen/dorsal	static
ethanol	46.07	13.0093	-114.1	10.92	-0.14	1	1	-2.68	not given	full-thickness	abdomen/dorsal	static
ethanol	46.07	13.0093	-114.1	10.92	-0.14	1	1	-2.70	not given	full-thickness	abdomen/dorsal	static
ethanol	46.07	13.0093	-114.1	10.92	-0.14	1	1	-2.72	not given	full-thickness	abdomen/dorsal	static
ethanol	46.07	13.0093	-114.1	10.92	-0.14	1	1	-3.05	not given	full-thickness	abdomen	static
ethanol	46.07	13.0093	-114.1	10.92	-0.14	1	1	-2.62	not given	full-thickness	back	static
epikote YX4000	354.4	101.7748		11.00	5.19	4	0	-5.58	32 (skin)	full-thickness	dorsal	flow-through
erioglaucine	793.86		283		-1.5			-4.52	37	full-thickness	abdomen	static
etorphine	411.55	116.2256	215	11.76	3.02	5	2	-2.34	37	full-thickness	abdominal/dorsal	static
5-fluorouracil	130.01	26.1222	281	13.46	-0.81	3	2	-4.22	31	dermatomed 0,42 mm	abdomen	flow-through
n-hexanol	102.18		-52	9.71	2.03	1	1	-1.71	not given	full-thickness	Abdomen	static
n-hexanol	102.18		-52	9.71	2.03	1	1	-1.54	not given	full-thickness	Abdomen	static
n-hexanol	102.18		-52	9.71	2.03	1	1	-1.41	not given	full-thickness	Abdomen	static

n-hexanol	102.18		-52	9.71	2.03	1	1	-1.43	not given	full-thickness	Abdomen	static
n-hexanol	102.18		-52	9.71	2.03	1	1	-1.43	not given	full-thickness	Abdomen	static
n-hexanol	102.18		-52	9.71	2.03	1	1	-2.42	not given	full-thickness	Abdomen	static
n-hexanol	102.18		-52	9.71	2.03	1	1	-1.07	not given	full-thickness	back	static
1,6-hexanediol diglycidyl ether	230.2	60.2566	84.83	9.36	0.84	4	0	-3.24	32 (skin)	full-thickness	dorsal	flow-through
n-heptanol	116.2	31.3365	-34	9.57	1.82	1	1	-1.18	not given	full-thickness	Abdomen	static
n-heptanol	116.2	31.3365	-34	9.57	1.82	1	1	-1.09	not given	full-thickness	Abdomen	static
n-heptanol	116.2	31.3365	-34	9.57	1.82	1	1	-0.99	not given	full-thickness	Abdomen	static
n-heptanol	116.2	31.3365	-34	9.57	1.82	1	1	-1.01	not given	full-thickness	Abdomen	static
n-heptanol	116.2	31.3365	-34	9.57	1.82	1	1	-0.98	not given	full-thickness	Abdomen	static
n-heptanol	116.2	31.3365	-34	9.57	1.82	1	1	-0.99	not given	full-thickness	Abdomen	static
hydrocortisone	362.5	97.6308	220	12.75	1.62	5	3	-4.00	37	full-thickness	Abdomen	static
hydrocortisone	362.5	97.6308	220	12.75	1.62	5	3	-3.92	37	full-thickness	dorsal	static
hydrocortisone	362.5	97.6308	220	12.75	1.62	5	3	-4.22	37	full-thickness	Abdomen	static

17a-hydroxyprogesterone	330.5	94.3367	276	10.98	3.08	3	1	-3.06	37	full-thickness	Abdomen	static
lidocaine	234.34	72.1476	67	8.78	1.66	2	1	-2.05	37	full-thickness	Dorsal/Ventral	flow-through
lidocaine	234.34	72.1476	67	8.78	1.66	2	1	-1.75	37	full-thickness	Dorsal/Ventral	flow-through
lidocaine	234.34	72.1476	67	8.78	1.66	2	1	-1.74	37	full-thickness	Dorsal/Ventral	flow-through
lidocaine	234.34	72.1476	67	8.78	1.66	2	1	-1.59	37	full-thickness	Dorsal/Ventral	flow-through
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-2.46	37	full-thickness	back	flow-through
methanol	32.04	8.2613	-98	11.68	-0.63	1	1	-2.74	not given	full-thickness	Abdomen/dorsal	static
methanol	32.04	8.2613	-98	11.68	-0.63	1	1	-2.80	not given	full-thickness	Abdomen/dorsal	static
methanol	32.04	8.2613	-98	11.68	-0.63	1	1	-2.77	not given	full-thickness	Abdomen/dorsal	static
methanol	32.04	8.2613	-98	11.68	-0.63	1	1	-2.72	not given	full-thickness	Abdomen/dorsal	static
methanol	32.04	8.2613	-98	11.68	-0.63	1	1	-2.74	not given	full-thickness	Abdomen/dorsal	static
methanol	32.04	8.2613	-98	11.68	-0.63	1	1	-2.74	not given	full-thickness	Abdomen/dorsal	static
methanol	32.04	8.2613	-98	11.68	-0.63	1	1	-3.00	not given	full-thickness	Abdomen	static
morphine	285.3	79.8346	255	13.68	0.72	4	2	-3.82	37	full-thickness	dorsal/abdominal	flow-through

nicorandil	211.18	0	92.5	14.40	0.43	3	1	-3.00	37	full-thickness	abdomen	static
n-octanol	130.23	40.5385	-15.5	9.45	2.81	1	1	-1.11	not given	full-thickness	abdomen/dorsal	static
n-octanol	130.23	40.5385	-15.5	9.45	2.81	1	1	-0.92	not given	full-thickness	abdomen/dorsal	static
n-octanol	130.23	40.5385	-15.5	9.45	2.81	1	1	-1.02	not given	full-thickness	abdomen/dorsal	static
n-octanol	130.23	40.5385	-15.5	9.45	2.81	1	1	-1.03	not given	full-thickness	abdomen/dorsal	static
n-octanol	130.23	40.5385	-15.5	9.45	2.81	1	1	-1.02	not given	full-thickness	abdomen/dorsal	static
n-octanol	130.23	40.5385	-15.5	9.45	2.81	1	1	-1.01	not given	full-thickness	abdomen/dorsal	static
prednisolone 21-heptanoate	472.62	130.9065	186	12.02	4.6	5	2	-4.11	25 (not given)	not given	abdomen	not given
prednisolone 21-octanoate	486.65	135.5075	159	11.89	5.09	5	2	-3.96	25 (not given)	not given	abdomen	not given
prednisolone 21-nonanoate	500.68	140.1085	131	11.78	5.58	5	2	-4.04	25 (not given)	not given	abdomen	not given
prednisolone 21-decanoate	514.7	144.7095	145	11.67	6.07	5	2	-3.95	25 (not given)	not given	abdomen	not given
prednisolone 21-undecanoate	528.73	0	129	11.57	5.54	5	2	-3.83	25 (not given)	not given	abdomen	not given

prednisolone 21-tridecanoate	556.78	0	140	11.39	6.02	5	2	-3.97	25 (not given)	not given	abdomen	not given
prednisolone 21-pentadecanoate	584.84	0	139	11.23	6.79	5	2	-4.01	25 (not given)	not given	abdomen	not given
progesterone	314.5	92.8212	121	10.05	3.67	2	0	-1.96	37	full-thickness	abdomen	static
propranolol HCL	295	0	233.2	11.96	0.74	3	2	-4.30	37	full-thickness	abdomen	flow-through
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-1.71	25	full-thickness	dorsal	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-1.93	25	full-thickness	dorsal	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-3.13	25	full-thickness	dorsal	static
squaric acid	114.06	25.027	293	20.47	-0.44	4	2	-3.15	not given	full-thickness	not given	static
sucrose	342.3	68.7741	190	18.06	-4.27	11	8	-2.66	37	full-thickness	not given	flow-through
thiourea	76.12	21.3277	176	15.23	-1.31	0	2	-4.02	37	full-thickness	abdomen	static
TDCPP	430.91	0	27	8.67	3.65	1	0	-3.34	not given	full-thickness	dorsal	flow-through
TDCPP	430.91	0	27	8.67	3.65	1	0	-3.25	not given	full-thickness	dorsal	flow-through
trifluralin	335.28	0	49	9.49	5.31	6	0	-3.55	not given	full-thickness	dorsal	flow-through
trifluralin	335.28	0	49	9.49	5.31	6	0	-3.91	not given	full-thickness	dorsal	flow-through

water	18.02	0	0	26.68	-1.38	1	1	-2.66	31 (skin)	full-thickness	dorsal	flow-through
water	18.02	0	0	26.68	-1.38	1	1	-2.80	not given	full-thickness	abdomen/dorsal	static
water	18.02	0	0	26.68	-1.38	1	1	-2.82	not given	full-thickness	abdomen/dorsal	static
water	18.02	0	0	26.68	-1.38	1	1	-2.85	not given	full-thickness	abdomen/dorsal	static
water	18.02	0	0	26.68	-1.38	1	1	-2.89	not given	full-thickness	abdomen/dorsal	static
water	18.02	0	0	26.68	-1.38	1	1	-2.92	not given	full-thickness	abdomen/dorsal	static
water	18.02	0	0	26.68	-1.38	1	1	-4.96	not given	full-thickness	abdomen/dorsal	static
water	18.02	0	0	26.68	-1.38	1	1	-2.53	35 (skin)	full-thickness	not given	flow-through
water	18.02	0	0	26.68	-1.38	1	1	-2.72	not given	full-thickness	abdomen	static
water	18.02	0	0	26.68	-1.38	1	1	-1.70	not given	full-thickness	back	static
water	18.02	0	0	26.68	-1.38	1	1	-1.74	37 (not given)	not given	dorsal	flow-through
water	18.02	0	0	26.68	-1.38	1	1	-2.00	37 (not given)	not given	dorsal	flow-through
urea	60.6	13.0926	135	14.36	-1.56	1	2	-3.52	37 (not given)	not given	dorsal	flow-through
urea	60.6	13.0926	135	14.36	-1.56	1	2	-3.15	37 (not given)	not given	dorsal	flow-through

B.2.2 Rat dataset

B.2.2.1 Experimental conditions

- **Skin thickness: All types**
- **Site: All body cites**
- **Cell type: Flow-through/Static**
- **Temperature:** $30^{\circ}C$ **and** $37^{\circ}C$
- **Time: Not specific**
- **Vehicle: Any**
- **Receptor fluid: Any**

B.2.2.2 Data

Name	Mwt	MR	Mpt	SP	logP	Ha	HD	logKp	Temp *C*	membrane	Site	Cell type
Alizapride	339.9	89.8971	207	12.06	1.8	6	2	-2.24	37	full-thickness	dorsal	flow-through
1,6-hexanediol diglycidyl ether (HDDGE)	230.2	60.2566	84.83	9.36	0.84	4	0	-3.40	32 (skin)	dermatomed	dorsal	flow-through
2-(2-methoxyethoxy)ethanol	120.15	30.3483	-70	10.25	-1.18	3	1	-1.10	32 (skin)	dermat. 0,56 mm	back	static
2,4 dimethy-lamine	266.13	0	86	9.52	0.84	3	2	-3.51	37	dermat. 0,3 mm	back	flow-through
2-ethoxyethanol	90.12	0	-90	10.33	-0.42	2	1	-4.11	32	dermat. 0,33 mm	dorsal	flow-through

2-ethoxyethanol	90.12	0	-90	10.33	-0.42	2	1	-3.64	32	full-thickness	dorsal	flow-through
2-ethoxyethanol	90.12	0	-90	10.33	-0.42	2	1	-3.83	32	dermat. 0,33 mm	dorsal	flow-through
2-ethoxyethanol	90.12	0	-90	10.33	-0.42	2	1	-4.12	32	full-thickness	dorsal	flow-through
2-phenoxyethanol	138.17	38.8135	14	11.49	1.1	2	1	-2.57	37	dermatom.0,244 mm	dorsal	static
2-phenoxyethanol	138.17	38.8135	14	11.49	1.1	2	1	-4.57	37	dermatom.0,244 mm	dorsal	static
2-phenoxyethanol	138.17	38.8135	14	11.49	1.1	2	1	-4.75	37	dermatom.0,244 mm	dorsal	flow-through
2-phenoxyethanol	138.17	38.8135	14	11.49	1.1	2	1	-2.75	37	dermatom.0,244 mm	dorsal	flow-through
2-phenoxyethanol	138.17	38.8135	14	11.49	1.1	2	1	-2.13	37	dermatom.0,244 mm	dorsal	static
2-phenylphenol	170.21	52.8883	59	12.24	3.28	1	1	-3.01	32	full-thickness skin	dorsal/Flank	static
2-phenylphenol	170.21	52.8883	59	12.24	3.28	1	1	-1.58	32	Isolated Epidermis	dorsal/Flank	static
4-methylaniline	107.2	35.7996	43.7	10.58	1.62	1	1	-1.07	37	isol.epidermis		flow-through
4-n-butylaniline	149	49.6026	-14	9.51	3.1	1	1	-0.64	37	isol.epidermis		flow-through
4-n-hexylaniline	177.3	58.8046		9.68	4.08	1	1	-0.64	37	isol.epidermis		flow-through
4-n-pentylaniline	163.3	54.2036		9.79	3.59	1	1	-0.61	37	isol.epidermis		flow-through

4-n-propylaniline	135	45.0016		10.10	2.61	1	1	-0.77	37	isol.epidermis		flow-through
Aminopyrene	231	70.1104	108	10.7	0.6	3	0	-1.48	32	full-thickness		static
Aniline	93	30.7584	-6.2	10.83	1.08	1	1	-1.17	37	isol.epidermis		flow-through
benzoic acid	122.1	32.8164	122	11.94	1.87	1	1	-1.54	31,5 (skin)	isol.epidermis	back	flow-through
benzoic acid	122.1	32.8164	122	11.94	1.87	1	1	-2.02	31,5 (skin)	isol.epidermis	back	flow-through
benzoic acid	122.1	32.8164	122	11.94	1.87	1	1	-3.46	32	full-thickness	back	static
benzoic acid	122.1	32.8164	122	11.94	1.87	1	1	-3.54	32	full-thickness	abdomen	static
benzyl acetate	150.18	42.0254	-51.3	10.10	2.08	1	0	-3.57	32(skin)	full-thickness	dorsal	flow-through
benzyl acetate	150.18	42.0254	-51.3	10.10	2.08	1	0	-4.24	32(skin)	full-thickness	dorsal	flow-through
Bisphenol A diglycidyl ether (BADGE)	340.8	95.2347	10	10.38	3.84	4	0	-5.26	32 (skin)	dermatomed skin	dorsal	flow-through
bromopride	344.26	85.8368	152.5	10.74	1.94	4	2	-2.11	37	full-thickness	dorsal	flow-through
bufexamac	223.3	61.002	154	12.43	1.98	3	2	-0.57	32	full-thickness		static
butyl salicylate	194.23	53.1528	-6	11.45	4.08	2	1	-4.7	32 (skin)	full-thickness	not given	static
caffeine	194.2	49.2834	238	32.83	0.16	4	0	-3.51	32	full-thickness	back	static

caffeine	194.2	49.2834	238	32.83	0.16	4	0	-2.99	37	full-thickness	not given	flow-through
clebopride	373.9	105.5674	162	11.47	3.21	4	2	-2.13	37	full-thickness	dorsal	flow-through
clotrimazole	344.85	102.1434	148	11.17	6.26	2	0	-2.26	32	full-thickness	unknown	static
cortisone	360.5	96.7062	220	12.10	1.81	5	2	-3.77	32	full-thikness skin	back	static
cortisone	360.5	96.7062	220	12.10	1.81	5	2	-2.91	32	full-thikness skin	abdomen	static
cortisone	360.5	96.7062	220	12.10	1.81	5	2	-3.33	32	full-thikness skin	abdomen	static
coumarin	146.15	41.5486	70.6	11.91	1.51	1	0	-3.12	32(skin)	full-thikness skin	dorsal	flow-through
coumarin	146.15	41.5486	70.6	11.91	1.51	1	0	-3.1	32(skin)	full-thikness skin	dorsal	flow-through
DDT	354.49	85.3157	108.50	9.45	6.79	0	0	-2.17	37	dermatomed 0,5mm	back	flow-through
decane	142.3	47.8122	-29.7	7.58	5.25	0	0	-4.03	not given	dermatom. 0,56 mm	back	static
DEHP	390.57	113.409	-50	9.39	8.39	2	0	-3.02	31,5 (skin)	isol.epidermis	back	flow-through
DEHP	390.57	113.409	-50	9.39	8.39	2	0	-4.01	31,5 (skin)	isol.dermis	back	flow-through
DEHP	390.57	113.409	-50	9.39	8.39	2	0	-4.89	31,5 (skin)	isol.epidermis	back	flow-through
DEHP	390.57	113.409	-50	9.39	8.39	2	0	-4.32	31,5 (skin)	isol.dermis	back	flow-through
DEP	222.24	58.609	-3	10.51	2.65	2	0	-3.43	30 (not given)	isol.epidermis	abdomen	static
dibutylphthalate	278.35	0	-35	10.86	5.11	2	0	-4.43	36	full-thikness skin	dorsal	static
dibutylphthalate	278.35	0	-35	10.86	5.11	2	0	-4.05	30 (not given)	isol.epidermis	abdomen	static

diethylene glycol monomethyl ether	120.2	30.3483	-14.6	10.25	-1.5	3	1	-1.09	32 (skin)	dermatomed 0,56 mm	back	static
dinoseb	240.22	0	40	11.10	3.67	2	1	-3.06	32	dermatomed 0,35	dorsal	static
dinoseb	240.22	0	40	11.10	3.67	2	1	-2.94	32	dermatomed 0,35	dorsal	static
dodecane	170.34	57.0142	-9.6	7.69	6.23	0	0	-4.85	32 (skin)	dermatomed 0,56 mm	back	static
dodecyl decaethoxy-late	482.9	0			4.33			-3.37	37	dermatomed 0,244 mm	dorsal	static
dodecyl decaethoxy-late	482.9	0			4.33			-5.31	37	dermatomed 0,244 mm	dorsal	flow-through
dodecyl decaethoxy-late	482.9	0			4.33			-3.37	35	dermatomed 0,28 mm	dorsal	static
dodecyl glycidyl ether (C12GE)	242.2	72.4384	58	8.41	5.01	2	0	-4.54	32 (skin)	dermatomed skin	dorsal	flow-through
dodecyl mo-noethoxy-late	230.4	69.986		9.16	4.5	2	1	-2.26	37	dermatomed 0,244 mm	dorsal	static
dodecyl mo-noethoxy-late	230.4	69.986		9.16	4.5	2	1	-6.49	37	dermatomed 0,244 mm	dorsal	flow-through

domperidone	425.92	115.7121	242.5	12.48	3.35	3	2	-2.55	37	full-thickness	dorsal	flow-through
epikote YX4000	354.4	101.7748		11.00	5.19	4	0	-1.01	32	dermatomed skin	dorsal	flow-through
erioglaucine	793.86	0	283		-1.5			-5.4	37	full-thickness	abdomen	static
erioglaucine	793.86	0	283		-1.5			-4.88	37	full-thickness	abdomen	static
erioglaucine	793.86	0	283		-1.5			-5.27	37	full-thickness	abdomen	static
erioglaucine	793.86	0	283		-1.5			-4.12	37	full-thickness	abdomen	static
erioglaucine	793.86	0	283		-1.5			-5.00	37	full-thickness	abdomen	static
erioglaucine	793.86	0	283		-1.5			-4.70	37	full-thickness	abdomen	static
erioglaucine	793.86	0	283		-1.5			-4.70	37	full-thickness	abdomen	static
ethanol	46.07	13.0093	-114.1	10.92	-0.14	1	1	-3.38	30	full-thickness	back	static
ethyl benzene	106.2	35.7002	-94.9	9.04	3.03	0	0	-3.51	32 (skin)	dermat. 0,56 mm	back	static
ethylaniline	121.2	40.9998	-64	9.73	2.11	1	1	-0.94	37	isol.epidermis	Not given	flow-through
felodipine	384.3	0	283	10.43	3.86	3	1	-2.40	37	full-thickness	dorsal	flow-through
fenoxapropethyl	361.78	0	84	11.06	4.95	5	0	-3.15	37	dermatom. 0,6 mm	back	flow-through
fenoxapropethyl	361.78	0	84	11.06	4.95	5	0	-3.46	37	dermatom. 0,9 mm	back	flow-through

fenoxapropethyl	361.78	0	84	11.06	4.95	5	0	-3.46	37	dermatom. 0,8 mm	back	flow-through
haloperidol	375.9	102.5915	151.5	10.78	4.22	4	1	-1.70	37	full-thickness	abdomen	flow-through
hydroquinone	110.11	29.4462	170	15.18	1.03	2	2	-4.66	30	full-thickness	abdomen	static
ketoprofen	254.29	0	94	11.75	3.12	2	1	-1.73	32	full-thickness		static
linoleic acid	289.45	88.5188	-5	9.05	7.51	1	1	-4.54	37	full-thickness	sides,abdomen and back	static
lorazepam	321.16	80.9184	167	12.91	2.41	3	2	-4.29	37	full-thickness	abdomen	static
lorazepam	321.16	80.9184	167	12.91	2.41	3	2	-3.94	37	full-thickness	abdomen	static
lorazepam	321.16	80.9184	167	12.91	2.41	3	2	-4.11	37	full-thickness	abdomen	static
lorazepam	321.16	80.9184	167	12.91	2.41	3	2	-3.70	37	full-thickness	abdomen	static
lorazepam	321.16	80.9184	167	12.91	2.41	3	2	-3.44	37	full-thickness	abdomen	static
lorazepam	321.16	80.9184	167	12.91	2.41	3	2	-3.82	37	full-thickness	abdomen	static
lorazepam	321.16	80.9184	167	12.91	2.41	3	2	-3.25	37	full-thickness	abdomen	static
mannitol	182.17	38.4036	138.97	18.53	-3.01	6	6	-3.49	30	full-thickness	dorsal	static
mannitol	182.17	38.4036	138.97	18.53	-3.01	6	6	-3.64	30	isol.epidermis	dorsal	static
mannitol	182.17	38.4036	138.97	18.53	-3.01	6	6	-3.24	not given	full-thickness	dorsal	static
mannitol	182.17	38.4036	138.97	18.53	-3.01	6	6	-3.06	not given	isol.epidermis	dorsal	static

mannitol	182.17	38.4036	138.97	18.53	-3.01	6	6	-3.04	32	isol.epidermis	dorsal	not given
mannitol	182.17	38.4036	138.97	18.63	-3.01	6	6	-3.40	not given	dermat. 0,23 mm	dorsal	static
mannitol	182.17	38.4036	138.97	18.63	-3.01	6	6	-3.06	32	dermat. 0,23 mm	dorsal	flow-through
mannitol	182.17	38.4036	138.97	18.63	-3.01	6	6	-2.87	30	full-thickness	back	static
mefenamic acid	241.29	71.385	230	11.85	5.28	2	0	-2.11	32	full-thickness		static
methyl nicotinate	137.14	35.4286	42.5	11.80	0.64	2	0	-2.49	30(skin)	full-thickness skin	Abdomen	static
methyl nicotinate	137.14	35.4286	42.5	11.80	0.64	2	0	-2.56	30(skin)	full-thickness skin	Abdomen	static
methyl nicotinate	137.14	35.4286	42.5	11.80	0.64	2	0	-2.65	30(skin)	full-thickness skin	Abdomen	static
methyl nicotinate	137.14	35.4286	42.5	11.80	0.64	2	0	-2.18	30(skin)	full-thickness skin	Abdomen	static
metochloropramide	354.3	83.0188	182	11.13	1.69	4	2	-2.04	37	full-thickness	dorsal	flow-through
metopimazine	445.61	122.3837	170.5	13.55	2.42	3	1	-2.29	37	full-thickness	dorsal	flow-through
naphthalene	128.2	42.5082	80.6	10.42	3.17	0	0	-3.29	32(skin)	dermat.0,56mm	back	static
nicardipine	479.54	0	136	10.87	3.9	5	1	-2.31	37	full-thickness	dorsal	flow-through
nicorandil	211.18	0	92.5	14.40	0.43	3	1	-3.14	37	full-thickness	abdomen	static

nifedipine	346.3	0	172	10.92	4.04	4	0	-2.77	37	full-thickness	dorsal	flow-through
nimodipine	418.4	0	125	10.50	3.13	3	1	-2.59	37	full-thickness	dorsal	flow-through
nitrendipine	360.4	0	184.13	11.44	2.99	5	1	-2.41	37	full-thickness	dorsal	flow-through
N-N-Diethyl m-toluamide	191.28	58.9692	-45	10.70	2.26	1	0	-2.77	37	dermat.0,5mm	back	flow-through
nonane	128.3	43.2112	-53.5	7.51	4.76	0	0	-4.38	32 (skin)	dermatom. 0,56 mm	back	static
o-cresyl glycidyl ether (oCGE)	164.2	46.3071	30.25	10.38	2.16	2	0	-3.87	32	dermatomed skin	dorsal	flow-through
paraquat	257.16	0	300	10.45	-2.71	0	0	-3.46	30	full-thickness	back	static
propoxur	209.25	56.3432	87	10.31	1.9	2	1	-3.28	32	full-thickness	dorsal	static
propoxur	209.25	56.3432	87	10.31	1.9	2	1	-2.43	32	isol.epidermis	dorsal	static
salicylamide	137.14	36.3327	140	16.60	1.03	2	2	-4.77	32 (skin)	full-thickness	not given	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-1.98	25.	full-thickness	dorsal	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-2.14	25.	full-thickness	dorsal	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-2.88	25.	full-thickness	dorsal	static

salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-1.61	25.	full-thickness	dorsal	static
salicylic acid	138.1	34.5105	158	14.39	2.24	2	2	-4.62	32 (skin)	full-thickness	not given	static
scopolamine	303.4	79.7213	59	11.53	0.39	4	1	-2.39	37	full-thickness	dorsal	flow-through
terbinafine	291.4	98.0752		9.33	5.81	1	0	-1.26	32	full-thickness	unknown	static
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-3.70	32	isol.epidermis	abdomen	static
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-2.74	37(not given)	full-thickness	dorsal	static
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-3.89	37(not given)	full-thickness	dorsal	static
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-4.15	37(not given)	full-thickness	dorsal	static
testosterone	288.4	84.5453	155	10.66	3.27	2	1	-4.70	37(not given)	full-thickness	dorsal	static
theophylline	180.17	44.3867	273	14.05	-0.39	4	1	-2.03	37	dermatom. 0,45 mm	dorsal	flow-through
theophylline	180.17	44.3867	273	14.05	-0.39	4	1	-1.96	37	dermatom. 0,45 mm	dorsal	flow-through
theophylline	180.17	44.3867	273	14.05	-0.39	4	1	-1.86	37	dermatom. 0,45 mm	dorsal	flow-through
theophylline	180.17	44.3867	273	14.05	-0.39	4	1	-2.66	37	dermatom. 0,45 mm	dorsal	flow-through
toluene (methyl benzene)	92.1	31.0992	-94.9	9.14	2.54	0	0	-2.96	32(skin)	dermatom. 0,56 mm	back	static
triclosan	289.55	68.4073	56	10.02	2.47	2	1	0.13	32	dermatomed 0,28 mm	dorsal	flow-through

tridecane	185.4	61.6152	-5.5	7.74	6.73	0	0	-4.82	32(skin)	dermatomed 0,56 mm	back	static
undecane	156.31	52.4132	-25.6	7.64	5.74	0	0	-4.60	32(skin)	dermatom. 0,56 mm	back	static
urea	60.6	13.0926	135	14.36	-1.56	1	2	-4.80	32	full-thickness	back	static
urea	60.6	13.0926	135	14.36	-1.56	1	2	-3.80	32	full-thickness	back	static
urea	60.6	13.0926	135	14.36	-1.56	1	2	-2.73	32	full-thickness	abdomen	static
water	18.02	0	0	26.68	-1.38	1	1	-2.13	31,5 (skin)	isol.epidermis	back	flow-through
water	18.02	0	0	26.68	-1.38	1	1	-2.77	31,5 (skin)	isol.epidermis	back	flow-through
water	18.02	0	0	26.68	-1.38	1	1	-1.29	31,5 (skin)	isol.dermis	back	flow-through
water	18.02	0	0	26.68	-1.38	1	1	-2.84	30	full-thickness	dorsal	static
water	18.02	0	0	26.68	-1.38	1	1	-2.94	30	isol.epidermis	dorsal	static
water	18.02	0	0	26.68	-1.38	1	1	-2.71	not given	full-thickness	dorsal	static
water	18.02	0	0	26.68	-1.38	1	1	-2.85	not given	isol.epidermis	dorsal	static
water	18.02	0	0	26.68	-1.38	1	1	-2.7	32	dermatomed 0,23 mm	dorsal	flow-through
xylene (dimethyl benzene)	106.2	36.1404	-50	9.10	3.09	0	0	-3.77	32(skin)	dermatom. 0,56 mm	back	static

B.2.3 Pig dataset

B.2.3.1 Experimental conditions

- **Skin thickness: Full thicknessepidermal**
- **Site: Outer ear**

- **Cell type: Flow-through/Static**
- **Temperature:** $30^{\circ}C$ **and** $37^{\circ}C$
- **Time: 4, 24, 48h**
- **Vehicle: Water, physiological buffer, propylene glycol, ethanol, aqueous saturated solution, isotonic phosphate buffered saline (pH7.4) with ethanol, propylene glycol, Azone**
- **Receptor fluid: phosphate buffered saline PH 7.4**

B.2.3.2 Data

Name	Mwt	MR	Mpt	SP	logP	Ha	HD	logKp	Temp *C*	membrane	Site	Cell type
water	18.02	0.00	0.00	26.68	-1.38	1	1	-2.65	not given	full-thickness	outer ear	static
salicylic acid	138.1	34.5105	158.00	14.39	2.24	2	2	0.10	32	dermatomed 0,6 mm	unknown	static
2-phenylphenol	170.21	52.8883	59	12.24	3.28	1	1	-1.80	30	full-thickness	ear	perfused pig ear
mannitol	182.17	38.4036	138.97	18.63	-3.01	6	6	-2.82	32	isol.epidermis	outer ear	flow-through
caprtopril	217.29	54.7325	106	11.55	0.84	2	2	-3.53	37	fresh epidermal membrane	pig ears	static
caprtopril	217.29	54.7325	106	11.55	0.84	2	2	-2.83	37	frozen epidermal membrane	pig ears	static
methyl ester	231.29	52.9845	119	9.31	2.9	2	0	-2.75	37	fresh epidermal membrane	pig ears	static

methyl ester	231.29	52.9845	119	9.31	2.9	2	0	-2.44	37	frozen epidermal membrane	pig ears	static
ethyl ester	245.29	57.7325	105	9.25	3.39	2	0	-1.66	37	fresh epidermal membrane	pig ears	static
ethyl ester	245.29	57.7325	105	9.25	3.39	2	0	-2.09	37	frozen epidermal membrane	pig ears	static
propyl ester	259.29	62.2567	116	9.20	3.89	2	0	-1.8	37	fresh epidermal membrane	pig ears	static
propyl ester	259.29	62.2567	116	9.20	3.89	2	0	-1.91	37	frozen epidermal membrane	pig ears	static
butyl ester	273.29	66.8577	127	8.14	4.38	2	0	-1.84	37	fresh epidermal membrane	pig ears	static
butyl ester	273.29	66.8577	127	8.14	4.38	2	0	-1.97	37	frozen epidermal membrane	pig ears	static
tropicamide	284.4	82.5341	96.50	12.53	1.19	3	1	-6.15	37	epidermal membrane	pig ears	flow-through
pentyl ester	287.29	71.4587	138	9.11	4.87	2	0	-2.16	37	fresh epidermal membrane	pig ears	static
pentyl ester	287.29	71.4587	138	9.11	4.87	2	0	-2.97	37	frozen epidermal membrane	pig ears	static

atropine	289.38	80.8156	116	11.04	1.91	3	1	-5.75	37	epidermal membrane	pig ears	flow-through
hexyl ester	301.29	76.0597	150	9.07	5.36	2	0	-2.91	37	frozen epidermal membrane	pig ears	static
scopolamine	303.4	79.7213	59.00	11.53	0.39	4	1	-6.29	37	epidermal membrane	pig ears	flow-through

B.3 Enhancement Ratio (ER) dataset

B.3.1 Data

Compound	H bond	C chain	MW	Mean logP	Mean logS	ER Q	Group	Formula
696.01	7	0	60.06	-1.692	0.565	1.5	0	C1H4N2O1
695.04	3	0	85.11	-0.658	0.595	1.2	0	C4H7N1O1
695.13	1	1	85.15	0.72	0.385	1.4	0	C5H11N1
695.05	2	1	99.13	-0.328	0.685	1	0	C5H9N1O1
695.06	3	1	99.13	-0.164	0.34	1.3	0	C5H9N1O1
695.14	3	1	113.12	-0.688	0.465	1.4	0	C5H7N1O2
695.08	2	2	113.16	0.228	0.445	1.1	0	C6H11N1O1
695.09	6	0	129.12	-1.102	-0.065	1.1	0	C5H7N1O3
694.1	3	1	129.16	-0.598	0.645	1.3	0	C6H11N1O2
694.09	3	1	155.2	0.58	-0.285	4.6	0	C8H13N1O2
695.11	5	2	157.17	-0.226	-0.13	1.1	0	C7H11N1O3
695.16	4	1	159.23	0.214	-0.013	2	0	C8H17N1O2
695.12	2	2	167.25	1.756	-1.02	1.2	0	C10H17N1O1
695.15	2	6	169.27	2.276	-1.365	1.2	0	C10H19N1O1
697.01	2	1	181.26	2.263	-2.54	0.74	0	C9H11N1O1S1
698.15	3	1	198.25	2.52	-3.32	1.47	0	C8H10N2O2S1
698.13	4	1	199.23	1.54	-2.86	9.03	0	C7H9N3O2S1

443.06	2	9	200.32	4.23	-3.64	8	0	C12H24O2
690.01	4	1	210.28	0.242	-0.335	0.8	0	C11H18N2O2
696.07	5	2	212.25	2.842	-3.37	2	0	C13H12N2O1
698.14	3	1	212.27	2.88	-3.49	1.17	0	C9H12N2O2S1
697.02	2	1	215.7	2.75	-3.05	1.44	0	C9H10N1O1S1Cl1
698.01	3	1	217.31	2.007	-2.18	0.48	0	C8H11N1O2S2
694.06	1	10	224.39	5.77	-5.23	6.7	0	C15H28O1
698.1	4	1	226.26	2.175	-3.61	0.77	0	C9H10N2O3S1
696.08	5	2	228.32	2.752	-4.33	3.7	0	C13H12N2S1
698.12	2	1	228.32	3.49	-5.69	1.16	0	C13H12N2S1
696.02	6	12	228.38	4.466	-3.3	2.8	0	C13H28N2O1
698.06	3	1	231.34	2.492	-2.46	0.93	0	C9H13N1O2S2
694.05	1	10	238.41	6.058	-5.475	7.9	0	C16H30O1
695.03	1	12	239.44	6.492	-5.45	5.2	0	C16H33N1
696.03	5	12	242.4	4.908	-3.295	1.8	0	C14H30N2O1
698.11	3	1	248.31	3.575	-4.92	1.28	0	C12H12N2O2S1
698.08	3	1	251.76	2.6	-2.9	0.4	0	C8H10N1O2S2Cl1
694.02	2	11	253.43	5.19	-4.325	15.6	1	C16H31N1O1
695.01	2	12	253.43	5.494	-4.41	23	1	C16H31N1O1
698.18	1	1	256.58	4.292	-4.69	2.21	0	C8H8N1S1Cl3
698.16	3	2	257.36	3.775	-4.06	2.81	0	C14H15N3S1
696.04	5	12	258.47	5.18	-4.31	5.3	0	C14H30N2S1
698.09	2	1	260.15	2.738	-2.83	23.12	1	C9H10N1O1S1Br1
698.04	5	1	262.31	1.968	-3.33	0.68	0	C8H10N2O4S2
694.04	3	11	267.41	4.946	-3.895	10.1	1	C16H29N1O2
694.01	2	11	267.46	5.65	-4.56	14.7	1	C17H33N1O1
694.14	2	12	267.46	5.82	-4.62	22.2	1	C17H33N1O1
690.07	6	1	268.31	0.156	-0.89	0.6	0	C13H20N2O4
694.12	3	11	269.43	4.488	-3.62	15.1	1	C16H31N1O2

443.05	5	12	273.46	4.332	-3.62	6.4	0	C16H35N1O2
698.07	5	1	275.35	2.117	-2.94	0.19	0	C10H13N1O4S2
690.02	4	6	280.41	2.69	-2.29	2.1	0	C16H28N2O2
694.03	3	11	281.44	5.36	-4.165	7.7	0	C17H31N1O2
694.99	2	12	281.48	6.254	-4.85	22.1	1	C18H35N1O1
443.03	3	12	285.47	5.22	-4.375	7.8	0	C17H35N1O2
694.07	3	11	285.49	5.422	-4.49	21	1	C16H31N1O1S1
696.09	5	12	304.48	6.675	-3.94	1.1	0	C19H32N2O1
695.02	4	12	311.46	5.308	-4.105	11	1	C18H33N1O3
443.04	3	12	313.52	6.002	-5.045	6.1	0	C19H39N1O2
696.1	5	12	320.54	6.614	-5.085	3.4	0	C19H32N2S1
690.03	4	10	336.52	4.682	-3.945	8.8	0	C20H36N2O2
690.08	6	6	338.45	2.706	-2.605	1	0	C18H30N2O4
698.17	7	1	338.45	1.777	-3.83	0.83	0	C14H18N4O2S2
698.19	8	1	340.43	1.495	-4.3	0.72	0	C13H16N4O3S2
690.04	4	12	364.57	5.688	-4.46	11	1	C22H40N2O2
690.09	6	8	366.5	3.704	-3.505	2.2	0	C20H34N2O4
690.05	4	14	392.63	6.432	-4.97	18.6	1	C24H44N2O2
690.1	6	10	394.55	4.648	-4.16	4	0	C22H38N2O4
694.08	2	12	395.71	9.684	-6.55	8.9	0	C26H53N1O1
696.05	5	12	396.7	9.484	-6.85	1.9	0	C25H52N2O1
696.06	5	12	412.77	9.702	-7.81	1.6	0	C25H52N2S1
690.06	4	16	420.68	7.382	-5.455	9.6	0	C26H48N2O2
690.11	6	12	422.61	5.652	-4.72	9.1	0	C24H42N2O4
690.12	6	14	450.66	6.394	-5.29	9.6	0	C26H46N2O4

B.4 Magnusson datasets

B.4.1 Magnusson set A

B.4.1.1 Experimental conditions

- **Temperature**: 298 to 312 Kelvin

B.4.1.2 Data

Compound	logJmaxb	Texpc	Mwd	logKowe	Mptf	Saqg	Saq(T)h	Hdk	Hal
Benzene	-5.61	304	78.1	2.22	279	2.30E-05	2.30E-05	0	0
Benzoic acid	-5.9	308	122.1	1.9	395	4.10E-05	4.10E-05	1	2
Benzoic acid	-5.9	308	122.1	1.9	395	4.10E-05	4.10E-05	1	2
Benzyl alcohol	-5.62	298	108.1	1.04	258	4.00E-04	4.00E-04	1	1
Betamethasone-17-valerate	-10.65	298	476.6	3.98	457	1.90E-08	1.90E-08	2	6
p-Bromophenol	-5.5	298	173	2.49	337	8.70E-05	8.70E-05	1	1
2,3-Butanediol	-6.25	303	90.1	-0.99	298	1.1E-2n	1.1E-2n	2	2
Butanol	-5.59	303	74.1	0.88	184	8.50E-04	8.50E-04	1	1
Butanol	-5.67	298	74.1	0.88	184	8.50E-04	8.50E-04	1	1
2-Butanone	-4.86	303	72.1	0.37	187	3.10E-03	3.10E-03	0	1
Chlorocresol	-5.72	298	142.6	2.89	340	3.50E-05	3.50E-05	1	1
p-Chlorophenol	-5.17	298	128.6	2.43	317	1.90E-04	1.90E-04	1	1
o-Chlorophenol	-5.25	298	128.6	2.04	282	1.70E-04	1.70E-04	1	1
Chloroxylenol	-6.95	298	156.6	3.35	389	1.90E-06	1.90E-06	1	1
Cortexone	-9.87	299	330.5	3.41	415	1.80E-07	1.30E-07	1	3
Corticosterone	-9.51	300	346.5	1.76	454	5.70E-07	6.20E-07	2	4
Corticosterone	-10.54	299	346.5	1.76	454	5.70E-07	5.90E-07	2	4
Corticosterone	-10.89	298	346.5	1.76	454	5.70E-07	5.70E-07	2	4
Corticosterone	-8.83	312	346.5	1.76	454	5.70E-07	9.10E-07	2	4
Cortisone	-11.19	299	360.5	1.24	495	7.80E-07	8.10E-07	2	5
p-Cresol	-5.47	298	108.1	1.94	309	1.90E-04	1.90E-04	1	1

p-Cresol	-4.62	310	108.1	1.94	309	2.00E-04	2.50E-04	1	1
o-Cresol	-5.44	298	108.1	1.94	303	2.30E-04	2.30E-04	1	1
m-Cresol	-5.45	298	108.1	1.94	285	2.30E-04	2.30E-04	1	1
Decanol	-7.73	298	158.3	4.06	279	2.30E-07	2.30E-07	1	1
2,4-Dichlorophenol	-5.73	298	163	3	318	3.10E-05	3.10E-05	1	1
beta-Estradiol	-9.89	310	272.4	4.13	449	1.30E-08	1.80E-08	2	2
beta-Estradiol	-11.88	299	272.4	4.13	449	1.30E-08	1.30E-08	2	2
beta-Estradiol	-10.2	305	272.4	4.13	449	1.30E-08	1.60E-08	2	2
Estriol	-11.23	299	288.4	2.94	555	1.10E-07	9.30E-08	3	3
Estrone	-10.76	299	270.4	3.69	528	1.10E-07	1.20E-07	1	2
Ethanol	-4.87	298	46	-0.19	159	1.7E-2n	1.7E-2n	1	1
2-Ethoxy ethanol	-5.58	303	90.1	-0.27	183	1.0E-2n	1.0E-2n	1	2
Ethyl ether	-4.88	303	74.1	0.98	157	8.20E-04	8.20E-04	0	1
p-Ethylphenol	-5.85	298	122.2	2.47	318	4.10E-05	4.10E-05	1	1
5-Fluorouracil (+ - + -)	-8.57	305	130.1	-0.78	556	8.50E-05	1.30E-04	2	4
Heptanol	-6.27	303	116.2	2.47	238	1.40E-05	1.40E-05	1	1
Heptanol	-6.34	298	116.2	2.47	238	1.40E-05	1.40E-05	1	1
Hexanol	-6.13	298	102.2	1.94	228	5.80E-05	5.80E-05	1	1
Hydrocortisone (HC)	-11.64	299	362.5	1.43	493	8.80E-07	9.20E-07	3	5
Hydrocortisone (HC)	-11.6	298	362.5	1.43	493	8.80E-07	8.80E-07	3	5
4-Hydroxybenzyl alcohol	-6.97	310	124.1	0.3	393	5.40E-05	8.90E-05	2	2
alfa-(4-Hydroxyphenyl) acetamide	-7.37	310	151.2	-0.29	450	9.50E-05	1.40E-04	3	3
17-alfa-Hydroxyprogesterone	-10.77	299	330.5	2.89	496	2.00E-08	1.90E-08	1	3
Mannitol	-7.05	312	182.2	-4.67	440	9.10E-04	1.80E-03	6	6
Mannitol	-7.26	300	182.2	-4.67	440	9.10E-04	1.00E-03	6	6
Mannitol	-6.93	303	182.2	-4.67	440	9.10E-04	1.20E-03	6	6
Methanol	-4.81	298	32	-0.72	175	3.1E-2n	3.1E-2n	1	1
Methanol	-4.3	303	32	-0.72	175	3.1E-2n	3.1E-2n	1	1
Methyl-4-hydroxy benzoate	-6.92	298	152.1	1.87	401	1.30E-05	1.30E-05	1	3

beta-Naphthol	-6.71	298	144.2	2.71	396	6.90E-06	6.90E-06	1	1
Nicotinate, ethyl	-5.65	310	151.2	1.41	282	3.70E-04	3.70E-04	0	3
Nicotinate, methyl	-5.97	310	137.1	0.88	316	3.50E-04	5.20E-04	0	3
p-Nitrophenol	-6.25	298	139.1	1.57	387	1.00E-04	1.00E-04	1	4
m-Nitrophenol	-6.28	298	139.1	1.93	370	9.30E-05	9.30E-05	1	4
Nonanol	-7.23	298	144	3.53	268	9.70E-07	9.70E-07	1	1
Octanol	-6.67	298	130.2	3	258	4.10E-06	4.10E-06	1	1
Octanol	-6.6	303	130.2	3	258	4.10E-06	4.10E-06	1	1
Pentanol	-5.82	298	88.2	1.41	194	2.50E-04	2.50E-04	1	1
Phenol	-4.77	310	94.1	1.48	314	8.80E-04	1.20E-03	1	1
Phenol	-6.88	295	94.1	1.48	314	8.80E-04	8.20E-04	1	1
Phenol	-5.17	298	94.1	1.48	314	8.30E-04	8.30E-04	1	1
2-Phenylethanol	-5.86	298	122.2	1.36	259	1.80E-04	1.80E-04	1	1
o-Phenylenediamine	-6.74	305	108.1	0.05	377	4.10E-04	3.80E-04	4	2
p-Phenylenediamine	-7.09	305	108.1	-0.85	419	3.40E-04	4.50E-04	4	2
Prednisolone	-10.56	298	360.4	1.69	514	6.20E-07	6.20E-07	3	5
Pregnenolone	-10.09	299	316.5	4.52	466	2.20E-08	1.50E-08	1	2
Progesterone	-10.37	299	314.5	4.04	394	2.80E-08	2.90E-08	0	2
Propanol	-4.65	303	60	0.34	147	1.3E-2n	1.3E-2n	1	1
Propanol	-4.8	298	60	0.34	147	1.3E-2n	1.3E-2n	1	1
Resorcinol	-5.81	298	110.1	0.76	384	6.50E-03	6.50E-03	2	2
Sucrose	-7.24	310	342.3	-3.85	459	6.10E-03	9.20E-03	8	11
Testosterone	-10.46	299	288.4	3.48	428	8.10E-08	8.40E-08	1	2
Testosterone	-10.16	298	288.4	3.48	428	8.10E-08	8.10E-08	1	2
Thymol	-6.45	298	150.2	3.28	325	6.70E-06	6.70E-06	1	1
Toluene	-5.32	310	92.1	2.68	178	5.70E-06	5.70E-06	0	0
Triamcinolone	-12.09	298	394.5	1.03	543	2.00E-07	2.00E-07	4	6
Triamcinolone acetonide	-12.01	298	434.5	2.6	566	4.80E-08	4.20E-08	2	6
2,4,6-Trichlorophenol	-6.57	298	197.5	3.58	342	4.60E-06	4.60E-06	1	1

Urea	-5.6	312	60.1	-2.11	406	9.10E-03	1.40E-02	4	3
Urea	-5.76	300	60.1	-2.11	406	9.10E-03	9.60E-03	4	3
Urea	-5.87	310	60.1	-2.11	406	9.10E-03	1.30E-02	4	3
Water	-4.06	303	18	-1.38	273	5.6E-2n	5.6E-2n	2	1
Water	-4.07	305	18	-1.38	273	5.6E-2n	5.6E-2n	2	1
Water	-4.32	303	18	-1.38	273	5.6E-2n	5.6E-2n	2	1
Water	-4.56	298	18	-1.38	273	5.6E-2n	5.6E-2n	2	1
3,4-Xylenol	-5.83	298	122.2	2.4	334	4.10E-05	4.10E-05	1	1

B.4.2 Magnusson set B

B.4.2.1 Experimental conditions

- **Temperature**: 295 to 310 Kelvin

B.4.2.2 Data

Compound	logJmaxb	Texpc	Mwd	logKowe	Mptf	Saqg	Saq(T)h	Hdk	Hal
Aminopyrine	-6.6	310	231.3	0.76	381	2.40E-04	2.90E-04	0	4
Aniline	-5.09	303	93.1	0.94	267	3.90E-04	3.90E-04	2	1
Anisole	-5.89	303	108.1	2.13	236	9.60E-06	9.60E-06	0	1
Antipyrine	-6.53	310	188.2	0.27	385	4.30E-03	4.30E-03	0	3
Benzaldehyde	-5.37	303	106.1	1.64	247	6.20E-05	6.20E-05	0	1
Benzene	-5.49	310	78.1	2.28	279	2.30E-05	2.30E-05	0	0
Benzoic acid	-5.86	308	122.1	1.9	395	4.10E-05	4.10E-05	1	2
Benzyl alcohol	-5.25	303	108.1	1.04	258	4.00E-04	4.00E-04	1	1
Betamethasone	-10.41	310	392.5	2.06	506	1.50E-07	2.40E-07	3	5
Betamethasone-17-valerate	-9.92	310	476.6	3.98	457	1.10E-08	1.80E-08	2	6
Butobarbitone	-8.35	303	212.3	1.75	400	2.30E-05	2.70E-05	2	5
Chloroform	-5.03	299	119.4	1.76	209	6.70E-05	6.70E-05	0	0
Coumarin	-6.93	310	146.1	1.39	344	1.30E-05	2.00E-05	0	2

Cyclobarbitone	-7.99	310	236.3	2.13	446	6.80E-06	1.00E-05	2	5
Dexamethasone	-9.76	305	392.5	2.06	535	2.30E-07	3.00E-07	3	5
Diethylcarbamazine	-6.3	303	199.3	1.14	321	4.00E-03	4.00E-03	0	4
Digitoxin	-12.77	303	764.9	2.73	529	5.10E-09	6.20E-09	5	13
Ephedrine	-5.74	303	165.2	1.05	307	3.00E-04	3.00E-04	2	2
beta-Estradiol	-10.23	303	272.4	4.13	449	1.10E-08	1.10E-08	2	2
beta-Estradiol	-10.24	303	272.4	4.13	449	1.50E-08	1.60E-08	2	2
Ethanol	-5.27	303	46	-0.19	159	1.7E-2n	1.7E-2n	1	1
Ethanol	-5.3	295	46	-0.19	159	1.7E-2n	1.7E-2n	1	1
5-Fluorouracil (+ - + -)	-8.66	304	130.1	-0.78	556	8.50E-05	1.20E-04	2	4
5-Fluorouracil(+ - + -)	-8.18	310	130.1	-0.78	556	8.50E-05	1.60E-04	2	4
Formaldehyde	-4.82	303	30	0.35	155	1.30E-02	1.30E-02	0	1
Glycolic acid	-5.15	305	76.1	-1.05	353	1.8E-2n	1.8E-2n	2	3
Griseofulvin	-10.5	310	352.8	2.36	493	2.40E-08	3.80E-08	0	6
Hexanol	-5.8	304	102.2	1.94	228	5.80E-05	5.80E-05	1	1
Hydrocortisone (HC)	-9.98	303	362.5	1.43	493	8.80E-07	1.10E-06	3	5
HC-yl-propionate	-10.12	310	418.5	2.51	469	2.30E-08	3.70E-08	2	6
Isoquinoline	-6.23	303	129.2	1.96	299	3.50E-05	3.60E-05	0	1
Isosorbide dinitrate	-7.03	310	236.1	0.9	343	2.30E-06	3.20E-06	0	10
Ketorolac (S)	-8.8	305	255.3	2.08	449	1.30E-07	1.30E-07	1	4
Ketorolac (R)	-8.71	305	255.3	2.08	449	1.40E-07	1.40E-07	1	4
Ketorolac (SR50:50)	-8.29	305	255.3	2.08	429	2.50E-07	2.50E-07	1	4
Mannitol	-7.25	303	182.2	-4.67	440	9.10E-04	1.20E-03	6	6
Morphine hydrochloride(+)	-7.17	310	339.4	-2.53	473	2.40E-04	4.40E-04	2	4
Nicorandil	-7.47	309	211.2	-1.02	358	1.90E-04	1.80E-04	1	7
Nicorandil ?	-7.31	310	211.2	0.72	366	1.90E-04	1.90E-04	1	7
Nicotine	-3.89	303	162.2	0.72	265	6.7E-3n	6.7E-3n	0	2
Nitroglycerine	-7.24	303	227.1	2.22	287	5.70E-06	5.70E-06	0	12
Octanol	-6.67	295	130.2	3	258	4.10E-06	4.10E-06	1	1

Ouabain	-10.86	303	584.7	-1.35	463	1.70E-05	2.00E-05	8	12
Ouabain	-10.09	303	584.7	-1.35	463	2.10E-05	2.40E-05	8	12
Pentanol	-5.82	295	88.2	1.41	194	2.50E-04	2.50E-04	1	1
2-Phenylethanol	-5.67	303	122.2	1.36	259	1.80E-04	1.80E-04	1	1
Propanol	-4.88	295	60	0.34	147	1.3E-2n	1.3E-2n	1	1
Sucrose	-7.5	310	342.3	-3.85	459	6.10E-03	9.20E-03	8	11
Tetrachloroethylene	-7.66	299	165.9	2.95	251	1.20E-06	1.20E-06	0	0
Trichloroethylene	-6	299	131.4	2.26	188	8.40E-06	8.40E-06	0	0
Water	-4.11	304	18	-1.38	273	5.6E-2n	5.6E-2n	2	1
Water	-4.06	303	18	-1.38	273	5.6E-2n	5.6E-2n	2	1
Water	-4.45	303	18	-1.38	273	5.6E-2n	5.6E-2n	2	1
Water	-3.81	305	18	-1.38	273	5.6E-2n	5.6E-2n	2	1

B.4.3 Magnusson set C

B.4.3.1 Experimental conditions

- **Temperature**: 298 to 310 Kelvin

B.4.3.2 Data

Compound	logJmaxb	Texpc	Mwd	logKowe	Mptf	Saqg	Saq(T)h	Hdk	Hal
Aniline	-4.7	303	93.1	0.94	267	3.90E-04	3.90E-04	2	1
Anisole	-5.04	303	108.1	2.13	236	9.60E-06	9.60E-06	0	1
Benzaldehyde	-4.73	303	106.1	1.64	247	6.20E-05	6.20E-05	0	1
Benzene	-4.62	304	78.1	2.22	279	2.30E-05	2.30E-05	0	0
Benzene	-5.9	303	78.1	2.22	279	2.30E-05	2.30E-05	0	0
Benzyl alcohol	-5.3	303	108.1	1.04	258	4.00E-04	4.00E-04	1	1
Butanol	-6.19	298	74.1	0.88	184	8.50E-04	8.50E-04	1	1
Decanol	-8.38	298	158.3	4.06	279	2.30E-07	2.30E-07	1	1

Diethylene glycol monobutyl ether	-5.74	310	162.23	0.44	205	6.1E-3n	6.1E-3n	1	3
Di (2-ethylhexyl) phthalate	-9.59	303	390.56	8.7	223	6.10E-10	6.90E-10	0	4
Ethanol	-4.91	299	46	-0.19	159	1.7E-2n	1.7E-2n	1	1
Ethoxyethyl acetate	-4.97	310	132.16	0.72	211	1.90E-03	1.90E-03	0	3
Ethyl-3-ethoxypropionate	-5.3	303	146.18	1.25	198	3.80E-04	3.80E-04	0	3
Ethylene glycol	-5.72	303	62.07	-1.36	260	1.7E-2n	1.7E-2n	0	2
2-Ethylhexanol	-6.53	303	130.23	2.82	197	6.80E-06	6.80E-06	1	1
Heptanol	-6.74	298	116.2	2.47	238	1.40E-05	1.40E-05	1	1
n-Hexane	-8.02	303	86.2	3.9	178	1.10E-07	1.10E-07	0	0
Hexanol	-6.37	298	102.2	1.94	228	5.80E-05	5.80E-05	1	1
Methanol	-3.59	298	32	-0.72	175	3.1E-2n	3.1E-2n	1	1
Nicotinate, benzyl	-7.02	310	213.2	2.66	297	1.60E-06	1.60E-06	0	3
Nicotinate, ethyl	-5.85	310	151.2	1.41	282	3.70E-04	3.70E-04	0	3
Nonanol	-7.77	298	144	3.53	268	9.70E-07	9.70E-07	1	1
Octanol	-7.2	298	130.2	3	258	4.10E-06	4.10E-06	1	1
Parathion	-11.59	310	291.3	3.84	279	3.80E-08	3.80E-08	0	6
Pentanol	-6.33	298	88.2	1.41	194	2.50E-04	2.50E-04	1	1
2-Phenylethanol	-5.27	303	122.2	1.36	259	1.80E-04	1.80E-04	1	1
Propanol	-5.67	298	60	0.34	147	1.3E-2n	1.3E-2n	1	1

B.4.4 Magnusson set D

B.4.4.1 Experimental conditions

- **Temperature:** 298 to 312 Kelvin

B.4.4.2 Data

Compound	logJmaxb	Texpc	Mwd	logKowe	Mptf	Saqg	Saq(T)h	Hdk	Hal

Acetylsalicylic acid	-7.89	305	180.2	1.19	408	2.60E-05	3.20E-05	1	4
2-Amino-4-Nitrophenol	-8.4	305	154.1	1.23	415	6.00E-06	7.90E-06	3	5
Amylobarbital	-8.22	303	226.3	2.1	431	2.70E-06	3.10E-06	2	5
Aspartic acid (- + and - + -)	-8.45	310	133.1	-0.67	543	4.00E-05	6.50E-05	4	5
Aspartic acid (- + and - + -)	-8.27	310	133.1	-0.67	543	4.00E-05	6.50E-05	4	5
Atenolol	-9.45	305	266.3	-1.4	428	7.10E-06	9.70E-06	4	5
Atropine	-10.16	303	289.4	1.53	389	8.30E-06	9.60E-06	1	4
Baclofen	-9.07	310	213.7	1.56	480	2.20E-05	3.20E-05	3	3
Barbital	-8.34	303	184.2	0.69	463	4.00E-05	4.80E-05	2	5
Caffeine	-6.75	303	194.2	-0.08	511	1.10E-04	1.30E-04	0	6
Codeine (+)	-9.52	310	299.4	2.04	428	7.10E-06	1.00E-05	1	4
Diclofenac Sodium	-7.47	310	296.2	-0.96	557	1.10E-04	1.80E-04	0	3
Diclofenac	-9.82	310	296.2	3.28	430	8.00E-09	1.20E-08	0	3
Dopamine hydrochloride(+)	-6.48	310	153.2	-3.4	439	3.40E-03	3.40E-03	4	3
Etorphine (+)	-8.72	310	411.5	3	487	5.30E-07	5.30E-07	2	5
Fentanyl	-9.11	310	336.5	3.93	361	5.90E-07	8.20E-07	0	3
Fentanyl	-8.23	303	336.5	3.93	361	5.90E-07	6.80E-07	0	3
Flurbiprofen	-7.27	310	244.3	4.12	384	3.30E-08	5.00E-08	1	2
Histidine (- +)	-7.79	310	155.2	-1.19	560	2.90E-04	4.80E-04	4	5
Hydromorphone (+)	-9.95	310	285.3	1.23	540	N/A	N/A	1	4
4-Hydroxyphenyl acetic acid	-6.55	310	152.1	0.77	423	1.10E-04	1.80E-04	2	3
Ibuprofen	-6.92	310	206.3	3.72	349	1.00E-07	1.40E-07	1	2
Indomethacin	-10.15	310	357.8	3.11	431	2.60E-09	3.80E-09	1	5
Indomethacin	-8.79	310	357.8	3.11	431	2.60E-09	3.80E-09	1	5
Isoprenaline hydrocloride(+)	-7.33	310	211.3	0.25	429	1.60E-03	2.30E-03	4	4
Ketoprofen	-7.36	310	254.3	2.81	367	2.00E-07	3.00E-07	1	3
Levodopa (- +)	-8.79	310	197.2	-4.7	558	2.50E-05	2.50E-05	5	5
Lidocaine (+)	-7.14	310	234.3	2.36	342	1.70E-05	2.10E-05	1	3
Lidocaine (+)	-6.54	310	234.3	2.36	342	1.70E-05	2.10E-05	1	3

L-Lysine	-5.35	310	146.2	-1.04	498	9.0E-3n	9.0E-3n	5	4
Meperidine (+)	-8.62	310	247.3	2.81	543	1.30E-05	1.70E-05	0	3
Metoprolol	-6.6	305	267.4	-0.51	397	3.00E-04	4.00E-04	2	4
Morphine (+)	-12.08	310	285.3	1.27	528	5.20E-07	1.00E-06	2	4
Morphine (+)	-10.68	310	285.3	1.27	528	5.20E-07	1.00E-06	2	4
Naproxen	-8.57	310	230.3	3	426	6.90E-08	1.00E-07	1	3
Nicotinic acid (- +)	-8.45	310	123.1	0.82	510	1.50E-04	2.30E-04	1	3
Oxprenolol	-6.43	305	265.4	0.08	353	2.40E-04	2.40E-04	2	4
Paraquat dichloride (++)	-7.31	303	186.3	-5.65	573	3.80E-03	5.80E-03	0	2
Paraquat dichloride (++)	-7.49	303	186.3	-5.65	573	3.80E-03	5.80E-03	0	2
Phenobarbitone	-8.66	303	232.2	1.71	449	4.80E-06	5.70E-06	2	5
Piroxicam	-7.73	310	331.4	1.71	472	6.90E-08	1.20E-07	2	7
Propranolol	-6.94	305	259.4	3.1	369	6.50E-05	8.40E-05	2	3
Propranolol (+)	-8.55	310	259.4	3.1	369	2.30E-06	3.70E-06	2	3
Salicylic acid	-7.08	298	138.1	2.06	432	1.60E-05	1.60E-05	2	3
Salicylic acid	-6.99	303	138.1	2.06	432	1.60E-05	1.90E-05	2	3
Salicylic acid	-6.22	310	138.1	2.06	432	1.60E-05	2.40E-05	2	3
Salicylic acid	-5.91	310	138.1	2.06	432	1.60E-05	2.40E-05	2	3
Salicylic acid	-6.45	310	138.1	2.06	432	1.60E-05	2.40E-05	2	3
Scopolamine	-7.9	303	303.4	1.34	332	2.50E-04	2.50E-04	1	5
Sufentanil	-8.99	310	386.5	3.42	370	2.00E-07	2.70E-07	0	4
Tetraethylammonium bromide (+)	-6.41	300	130.3	-2.82	559	7.70E-03	8.40E-03	0	1
Tetraethylammonium bromide (+)	-6.17	312	130.3	-2.82	559	7.70E-03	1.40E-02	0	1

B.4.5 Magnusson set E

B.4.5.1 Experimental conditions

- **Temperature:** 310 Kelvin

B.4.5.2 Data

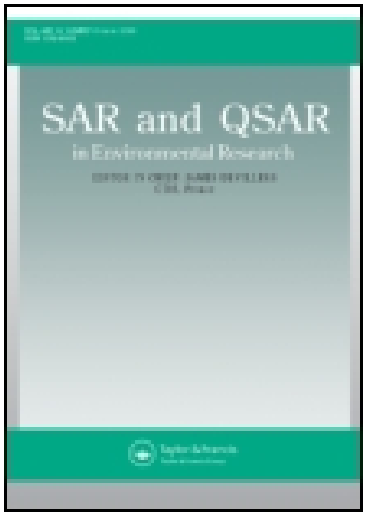
Compound	logJmaxb	Texpc	Mwd	logKowe	Mptf	Saqg	Saq(T)h	Hdk	Hal
Acetaminophen	-7.52	310	151.2	0.46	443	9.30E-05	1.37E-04	2	3
Benzoic acid	-5.23	310	122.1	1.9	395	4.10E-05	4.33E-05	1	2
Benzyl alcohol	-5.01	310	108.1	1.04	258	4.00E-04	3.97E-04	1	1
Caffeine	-8.11	310	194.2	-0.08	511	1.10E-04	1.75E-04	0	6
Clonidine	-7.22	310	230.1	1.59	403	5.90E-05	8.44E-05	2	3
Dextromethorphan	-7.43	310	271.4	4.13	383	N/A	N/A	0	2
Dextromethorphan HBr	-7.48	310	271.4	-0.05	398	N/A	N/A	0	2
Diazepam	-7.78	310	284.7	2.82	405	1.80E-07	2.51E-07	0	3
beta-Estradiol	-7.39	310	272.4	4.13	449	1.30E-08	1.84E-08	2	2
Ethacrynic acid	-5.61	310	303.1	3.69	396	N/A	N/A	1	4
5-Fluorouracil	-7.61	310	130.1	-0.78	556	8.50E-05	1.59E-04	2	4
Furosemide	-9.2	310	330.8	2.03	568	2.20E-07	2.95E-07	4	7
Griseofulvin	-9.17	310	352.8	2.36	493	2.40E-08	3.78E-08	0	6
Hydralazine HCl	-6.9	310	160.2	-1.53	548	N/A	N/A	3	4
Hydrocortisone	-8.94	310	362.5	1.43	493	8.80E-07	1.36E-06	3	5
Ibuprofen	-5.68	310	206.3	3.72	349	1.00E-07	1.39E-07	1	2
Indolyl-3-acetic acid	-7.2	310	175.2	1.41	442	8.60E-06	1.50E-05	2	3
Indomethacin	-9.16	310	357.8	3.11	431	2.60E-09	3.83E-09	1	5
Isosorbide dinitrate	-7.69	310	236.1	0.9	343	2.30E-06	3.15E-06	0	10
Ketoprofen	-7.33	310	254.3	2.81	367	2.00E-07	3.20E-07	1	3
Methyl salicylate	-5.05	310	152.1	2.55	265	4.60E-06	4.60E-06	1	3
Minoxidil	-8.41	310	209.3	1.24	521	1.10E-05	1.67E-05	4	6
Morphine sulfate	-9.18	310	285.3	-1.76	527	N/A	N/A	2	4
Naproxen	-7.68	310	230.3	3	426	6.90E-08	1.10E-07	1	3
Nicotinic acid	-7.75	310	123.1	0.82	510	1.50E-04	2.29E-04	1	3
Nifedipine	-9.69	310	346.3	2.2	446	1.60E-07	2.41E-07	1	8
Pentazocine	-8.62	310	285.4	4.64	419	N/A	N/A	1	2
Pentazocine HCl	-8.65	310	285.4	-1.7	527	N/A	N/A	1	2

Piroxicarn	-8.68	310	331.4	1.71	472	6.90E-08	1.17E-07	2	7
Propranolol HCl	-7.59	310	259.4	-0.45	436	N/A	N/A	2	3
Salicylamide	-6.41	310	137.1	0.89	415	1.50E-05	2.17E-05	3	3
Salicylic acid	-4.86	310	138.1	2.06	432	1.60E-05	2.38E-05	2	3
Sulindac	-10.07	310	356.4	3.42	456	8.40E-06	1.26E-05	1	3
Terbutaline sulfate	-9.75	310	225.3	-1.9	520	N/A	N/A	4	4
Testosterone	-6.92	310	288.4	3.48	428	8.10E-08	1.18E-07	1	2
Triamcinolone acetonide	-9.38	310	434.5	2.6	566	4.80E-08	7.01E-08	2	6

Appendix C

Peer reviewed journal and conference papers including poster abstracts

C.1 Journal paper



SAR and QSAR in Environmental Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gsar20

The application of machine learning to the modelling of percutaneous absorption: An overview and guide

P. Ashrafi[a], G.P. Moss[b], S.C. Wilkinson[c], N. Davey[a] & Y. Sun[a]
[a] School of Computer Science, University of Hertfordshire, Hatfield, UK
[b] School of Pharmacy, Keele University, Keele, UK
[c] Medical Toxicology Centre, Institute for Cellular Medicine, University of Newcastle-upon-Tyne, UK
Published online: 18 Mar 2015.

To cite this article: P. Ashrafi, G.P. Moss, S.C. Wilkinson, N. Davey & Y. Sun (2015) The application of machine learning to the modelling of percutaneous absorption: An overview and guide, SAR and QSAR in Environmental Research, 26:3, 181-204, DOI: 10.1080/1062936X.2015.1018941

To link to this article: http://dx.doi.org/10.1080/1062936X.2015.1018941

SAR and QSAR in Environmental Research, 2015
Vol. 26, No. 3, 181–204, http://dx.doi.org/10.1080/1062936X.2015.1018941

Taylor & Francis
Taylor & Francis Group

The application of machine learning to the modelling of percutaneous absorption: An overview and guide

P. Ashrafi[a], G.P. Moss[b]*, S.C. Wilkinson[c], N. Davey[a] and Y. Sun[a]

[a]*School of Computer Science, University of Hertfordshire, Hatfield, UK;* [b]*School of Pharmacy, Keele University, Keele, UK;* [c]*Medical Toxicology Centre, Institute for Cellular Medicine, University of Newcastle-upon-Tyne, UK*

(*Received 12 August 2014; in final form 21 January 2015*)

Machine learning (ML) methods have been applied to the analysis of a range of biological systems. This paper reviews the application of these methods to the problem domain of skin permeability and addresses critically some of the key issues. Specifically, ML methods offer great potential in both predictive ability and their ability to provide mechanistic insight to, in this case, the phenomena of skin permeation. However, they are beset by perceptions of a lack of transparency and, often, once a ML or related method has been published there is little impetus from other researchers to adopt such methods. This is usually due to the lack of transparency in some methods and the lack of availability of specific coding for running advanced ML methods. This paper reviews critically the application of ML methods to percutaneous absorption and addresses the key issue of transparency by describing in detail – and providing the detailed coding for – the process of running a ML method (in this case, a Gaussian process regression method). Although this method is applied here to the field of percutaneous absorption, it may be applied more broadly to any biological system.

Keywords: Gaussian process; machine learning; quantitative structure–permeability relationships (QSPRs); skin permeation; percutaneous absorption

1. Introduction

1.1 *Machine learning methods for predicting percutaneous absorption – a problem of perception?*

In developing mathematical models for percutaneous absorption, it is important to consider that the endpoint, and the end user or client, is the key driver in developing such models, rather than the mathematical modeller. In the context of skin absorption this reflects the relevance of the model to those working in transdermal or topical drug delivery, to cosmetic scientists and to toxicologists working in fields such as industrial risk assessment and the absorption of pesticides.

The most obvious application of such models is to generate an algorithm which enables prediction of absorption of chemicals for which no absorption data exist. Furthermore, mathematical models or algorithms describing percutaneous absorption may yield important mechanistic information on the absorption process. However, as machine learning (ML) methods do not yield an explicit algorithm [1], they are perceived to give limited mechanistic insight and

*Corresponding author. Email: g.p.j.moss@keele.ac.uk

therefore lack wider applicability to the field. There is also a perception that machine learning methods require the use of specialist software – either particular programmes or the expertise to develop, within existing packages, specific codecs and macros to interpret the data correctly [2]. Furthermore, Cronin and Schultz have commented on the use of non-linear methods, suggesting that they are prone to over-fitting and can often model the error inherent in the data [3]. This is a valid criticism, despite the clear understanding that the vast majority of biological systems are inherently non-linear in nature.

Models must be accessible to the field that they are applied to. This means that a model defining percutaneous absorption, whether represented by an explicit algorithm or some collation of descriptive statistics, should be comprehensible to, and usable by, researchers in the field to which the model is applied and not solely the developers of the model. This is perhaps demonstrated by the widespread applicability and use of quantitative structure–activity relationship (QSAR) models for percutaneous absorption such as Potts and Guy [4] or, to a lesser extent, the more recent models by Magnusson et al. [5] and Moss and Cronin [6], compared with those described by Patel et al. [7], Moss et al. [1], or indeed any number of machine learning methods that cannot be readily applied to new data but which simply model existing data and draw inferences from it. Thus, producing relevant models is a balance of employing powerful techniques with the ability of the model to find use in its field. This means that the ability to develop models is central to their continued use.

1.2 *Aims and objectives*

This article demonstrates the ease of use of machine learning models and aids the interpretation of the perceived limitations of models, including lack of transparency and lack of explicit outputs. Further, the Gaussian process methods recently used successfully in models of percutaneous absorption are examined in more detail and the methods used presented, as a 'tutorial', to address the perceived issue of lack of transparency so that other researchers in the field of percutaneous absorption, and related fields, may more readily employ such methods in their research.

2. Machine learning methods applied to the pharmaceutical sciences

2.1 *The first statistical models for the prediction of percutaneous absorption*

Traditionally, the prediction of percutaneous absorption using mathematical approaches has concentrated on predicting two parameters: the steady state flux (J_{ss}), defined as the rate of chemical absorption (per unit surface area of exposed skin) during the zero-order, steady-state part of the absorption profile, with dimensions of μg/cm^2/h and the permeability coefficient, k_p. The permeability coefficient is derived by dividing the flux by the concentration of the applied penetrant, typically in a saturated aqueous solution, and has dimensions of cm/h or cm/s. This allows fluxes to be compared across a wide range of chemicals, but k_p is a somewhat artificial term, being a composite of the diffusion coefficient, skin:vehicle partition coefficient and path length. This subject is reviewed in detail elsewhere [8,9].

The publication of large amounts of permeability data from a range of experiments led to the collation of data by Flynn [10]. From this he proposed a series of semi-quantitative expressions (Table 1) which related permeability across human skin to specific physicochemical

Table 1. Algorithms for calculating permeability coefficient (K_p).

	Low molecular weight compounds (<150 Da)	*High molecular weight compounds (>150 Da)*
$\log K_{ow} < 0.5$	$\log K_p = -3$	$\log K_p = -5$
$0.5 \leq \log K_{ow} \leq 3.0$	$\log K_p = \log K_{ow} - 3.5$	
$0.5 \leq \log K_{ow} \leq 3.5$		$\log K_p = \log K_{ow} - 5.5$
$\log K_{ow} > 3.0$	$\log K_p = -0.5$	
$\log K_{ow} > 3.5$		$\log K_p = -1.5$

Adapted from [10].

properties of molecules: their lipophilicity (represented by the octanol–water partition coefficient, log P or log K_{ow}) and size (as molecular weight, MW). This work was quantified [4,11] to develop quantitative structure–permeability relationships (QSPRs). The early work in this field has been extensively reviewed [8] and more recent work considered by the excellent review by Mitragotri and colleagues [12].

2.2 *New frontiers: Fuzzy logic and artificial neural networks*

One particular problem regarding the prediction of percutaneous absorption is the sparseness and ambiguity of available data, a point highlighted by Moss et al. [8]. Experimental data used to derive the early QSPR models exist for relatively few compounds and originate from a diversity of study designs, not always with the intention of deriving mathematical models. This prompted Pannier et al. [13] to use the adaptive neural fuzzy interference system to model skin permeability via the MatLab software package. Three models were developed using subtractive clustering to define structures within the data and to assign subsequent rules; the models developed were able to successfully predict skin permeability as well as, or better than, previously published algorithms with fewer inputs. Subsequently Keshwani et al. [14] developed a rule-based Takagi–Sugeno fuzzy model, which was shown to predict the permeability across human skin, using a previously published database containing 140 compounds and log P, MW and temperature as input variables. This fuzzy model was compared with a regression model for the same inputs using both the square of the correlation coefficient and root mean square error (RMSE) as measures of model quality, as well as comparison with previously published models. The results indicated that the fuzzy model performed better than the regression model with identical data and validation protocols, and was at least comparable with existing published models. However, despite the success of this modelling approach, it has not been widely adopted in the literature. This is most likely due to the complexity of the methodologies employed and their limited use in physical sciences.

Artificial neural networks (ANNs) are biologically inspired computer programs which aim to simulate the way in which the human brain processes information. They work by detecting the patterns and relationships in data, and ‘learn’ or are trained systematically through experiential modifications rather than from programming. An ANN is formed from hundreds of single processing elements (PEs) which are interconnected via a series of coefficients, or weightings, signifying the relative importance of connections within the network (Figure 1). Each PE within the network has individually weighted inputs, a transfer/transformation

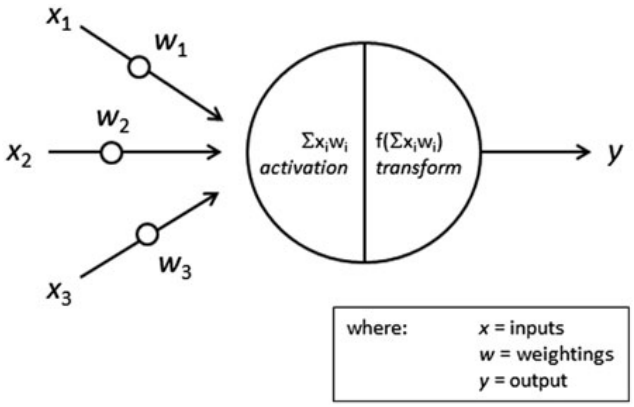


Figure 1. Model of an artificial neuron. Modified from ref. [15].

function and a single output which, along with the arrangement (or 'architecture' of the network), governs the function of the ANN and the nature of its output. Examples of the typical structures of feedback and feedforward ANNs are shown in Figure 2. The use of transformation functions may introduce non-linearity to the model. This function is optimised for each PE within the network to reduce error in predictions and to make the predictions available from the ANN as accurate as possible. Once this has been completed new data can be entered to the ANN (within the same 'chemical spaces' as the training dataset) and predictions output, that is, the same or similar domain which represents the range of physicochemical parameters of the dataset; in the case of the Potts and Guy model this was defined as $-3 < \log P < 6$ and $150 < \mathrm{MW} \leq 750$ [4,12]. This method has enormous potential for application to many fields, including pharmaceutical sciences.

For example, the broad application of ANNs to pharmaceutical research has been reviewed [15]. As such, they also indicate the potential of the Gaussian process methods to be applied to similar pharmaceutical domains. More broadly, ANNs have been applied to the interpretation of analytical data [16] to enhance the use of response surface methodology (RSM) to optimise formulations for drug release [17–22]. They have also been applied to the analysis of gene classification, and protein structure prediction and sequence classification [23,24]; they have also shown potential in the sequencing of biological molecules [25–27]. ANNs have been applied to the field of predictive drug absorption via the development of quantitative structure–permeability relationships (QSPRs). Such studies may employ related techniques, such as genetic analyses, to optimise descriptor use and subsequent analysis [28–31].

A number of researchers have applied ANN approaches to predict skin penetration [32–36]. They have also been applied to skin absorption, where a statistically significant model [$r^2 = 0.854$; mean square error (MSE) = 0.04] based on the Abraham's descriptors was reported [37]. The resulting ANN models gave very good correlations with experimental data and in many cases outperformed traditional QSPR models derived by linear regression or multiple linear regression. Indeed, one research group [35] noted that their ANN model was non-linear in nature. This is a finding that has been supported by more recent machine learning studies [1,2,38].

But despite the superior performance of these ANN models to more traditional approaches, very few of these techniques have established themselves as first choice methods

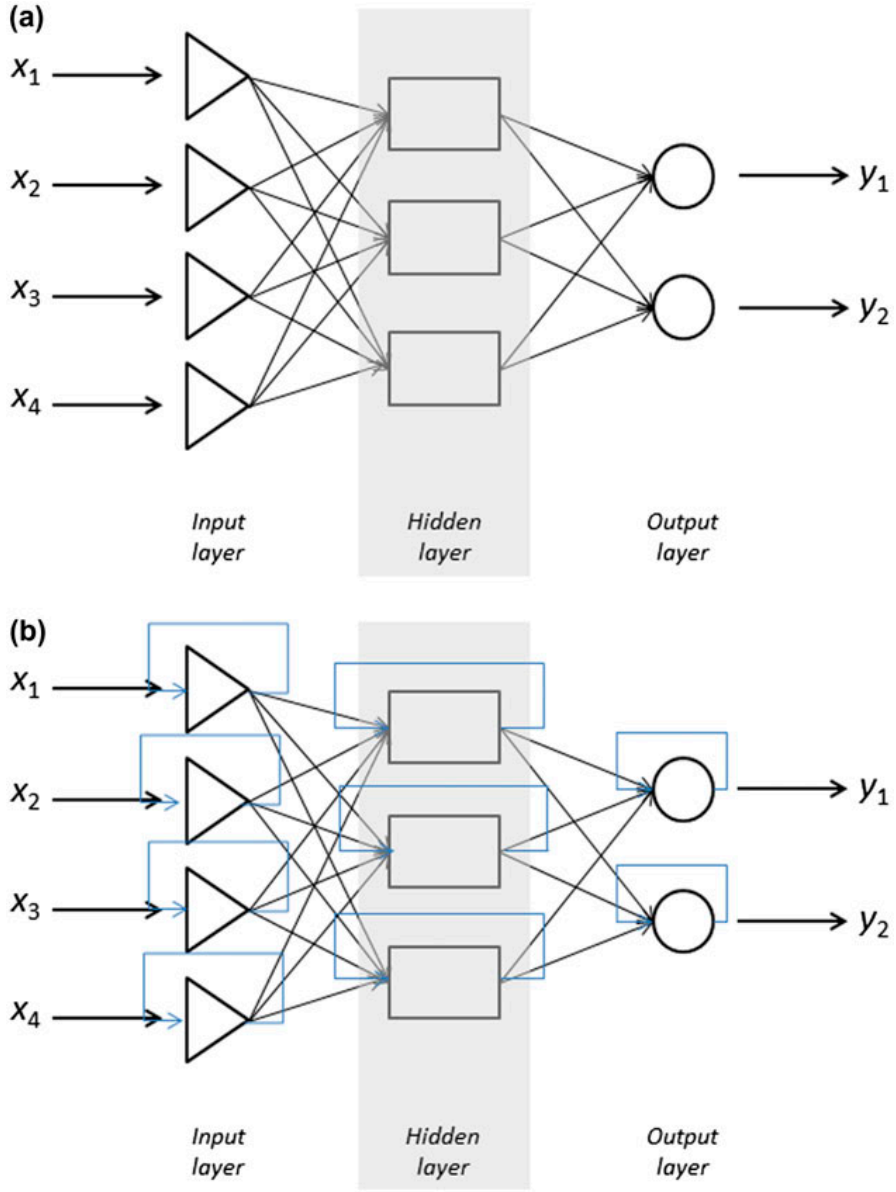


Figure 2. Schematic model of a feed–forward network (a) and a feed–back network (b) Modified from [15].

in the prediction of drug and/or exogenous chemical absorption via any number of routes. Nor has their improved performance of ANN-based formulation design resulted in it ousting response surface methodology as the key tool in this field. This is most likely due to the specialist nature of these methods and the lack of specialist expertise in the pharmaceutical sciences to apply this work correctly. This is emphasised by work of a greater clinical significance where ANN methods have been applied recently to the clinical diagnosis of skin disease based on the intelligent classification of sonogram information [39].

2.3 *Classification methods and avoiding over-parameterisation*

Classification methods were employed to analyse a dataset consisting of 116 structurally diverse compounds, mainly drugs, characterised by 1630 molecular descriptors and related to the skin permeability coefficient, represented by log k_p [40]. This approach was validated using 12 compounds not included in the original dataset. Further, stepwise multiple linear regression analysis was employed to find the optimal model. Avoidance of over-parameterisation, which may be artificially achieved by adding more terms to the model, was facilitated by the use of FIT (the Kubinyi function, which is related to the *F*-value but is less sensitive to changes in *k*, where *k* is the number of variables in the equation that describes the model, that is, the best model will have the highest FIT value) and AIC (Akaike's Information Criterion, a corrected measure of the sum of the square residuals, and thus;the best model will have the lowest value). This allowed the number of parameters to be reduced from 23 to 10, producing a model which the authors suggested was a good compromise between model complexity and model fitness. The inclusion of additional variables resulted in only minimal improvements in model quality and resulted in the following expression:

$$\begin{aligned}\log k_p(\mathrm{cm/s}) = {} & -6.243(\pm 2.12 \times 10^{-1}) - 3.14(\pm 6.17 \times 10^{-2})\mathrm{H.050} - 1.03(\pm 2.09 \\ & \times 10^{-1})\mathrm{Hypertens.50} + 1.04 \times 10^{-1}(\pm 5.73 \times 10^{-2})\mathrm{ALOGP} - 4.84 \\ & \times 10^{-4}(\pm 1.05 \times 10^{-4})\mathrm{SRW09} + 1.50 \times 10^{-1}(\pm 3.09 \times 10^{-2})\mathrm{RDF075m} \\ & - 1.39 \times 10^{-1}(\pm 2.99 \times 10^{-2})\mathrm{H.052} - 4.84 \times 10^{-1}(\pm 8.65 \times 10^{-2})\mathrm{T.(S..F)} \\ & + 4.77 \times 10^{-1}(\pm 1.10 \times 10^{-1})\mathrm{C.025} - 10.60(\pm 2.73)\mathrm{R1m} \\ & + -6.15(\pm 2.00)\mathrm{RTm+} \end{aligned} \tag{1}$$

where: H.050 (atom-centred fragment) represents the number of hydrogen atoms attached to a heteroatom; Hypertens.50 (molecular property class) is the Ghose–Viswanadhan–Wendoloski 50% antihypertensive drug-like index; SRW09 is the self-returning walk count of order 09; RDF075m is the radial distribution function 7.5, which is weighted by atomic masses (i.e. the corrected probability distribution associated with finding an atom in a spherical volume with radius *r*); H.052 is the number of hydrogen atoms attached to C(sp30 with one halogen attached to the next C; T.(S.F) is the sum of topological distances between S and F atoms; C.025 is the atom-centred fragment R–CR–R; and R1m+ and RTm+ are GETAWAY class descriptors describing the maximal autocorrelation of lag 1 and the maximal index, respectively, both of which are weighted by atomic masses.

Although this approach produced a novel, statistically robust model that was a significant improvement on others available at the time, it still lacked accessibility – given the parameters returned as significant, and their applicability and use more broadly in the field by non-experts in modelling. Sadly it has found little application within the field of percutaneous absorption.

2.4 *Embracing non-linearity: Gaussian process models applied to percutaneous absorption*

A large skin permeability dataset (n = 142) was subjected to analysis by data visualisation and both principal and canonical component analysis [1,2,38]. This demonstrated the fundamentally non-linear nature of the skin dataset, in contrast to a number of preceding studies which were based on (multiple) linear regression analysis and related methods. The lack of an explicit functional output, such as an equation, was addressed using the feature selection method [2]. This produced a series of models containing every possible permutation of the

physicochemical descriptors being modelled and therefore allows the best model – and combination of parameters – to be determined. Length-scale analysis [38] indicated that models of equal statistical quality and predictive ability could be constructed from a certain range of physicochemical parameters and that certain parameters were effectively interchangeable, or possibly co-linear. This finding suggests that the use of small datasets, or subsets, may bias the output of models to particular physicochemical parameters. It certainly reflects the importance of hydrogen bonding, as described previously [41–43] in the context of models produced by Potts and Guy and others [4,6]. Application of non-linear methods [44] produced a statistically robust model by integrating QSPRs, genetic algorithms and neural networks. The resulting model suggested that size/shape and polarity descriptors accounted for approximately 70% of the permeability information in their model.

While such methods may superficially appear to lack 'real world' relevance, particularly as they do not yield an easily digestible equation, their relevance was demonstrated when the performance of a Gaussian process model was compared with the Potts and Guy model [2]. They are characterised by the difference not only in proximity of each model to its intended experimental target, but in the overall pattern of predictions across the whole dataset. Further, Moss et al. [45] have shown the relevance of Gaussian process methods in a 'real world' situation when they examined a dataset of chemical penetration enhancers. They were able to show that the ML methods were able to: (1) provide fewer classification errors than discriminant analysis; and (2) generate predictions of enhancement, something which discriminant analysis was not able to achieve. So although still a novel method, the use of machine learning shows great potential but it does so, as all such studies should, within the framework for developing robust mathematical models of biological processes [3].

2.5 *Sources of input data: Literature sources of skin permeability data*

The underlying data inputs from which a model is constructed are vital to the relevance and validity of any model. For example, the applications for QSARs to skin permeation have been considered by Moss et al. [8]. In this review the authors discussed the range of experiments from which data were obtained and how this could introduce variance into any model subsequently derived. In particular, they focused on several key examples where skin permeability data was shown to be erroneous [46]. These data were used to develop models of skin absorption, as they were included in Flynn's dataset [10], which has been widely used by a large number of researchers. This led to inferences that, because they appeared as outliers in the dataset analysis, steroids permeated the skin by a different mechanism to other chemicals due to the comparatively large number of hydrogen bonds often found on such drugs. It was later determined that these data were erroneous by at least one order of magnitude [47] and, when they were replaced and remodelled it was found that steroids were no longer modelled as outliers [6]. Further, a large multi-centre validation study demonstrated that permeability data from a number of laboratories, for a standardised protocol, was still highly variable [48]. Thus, these studies underpin a major issue in the development of models using any statistical methods: the quality of the underlying data and its validation. In most cases, models are limited in that they are seldom able to consider such concerns due to a paucity of data and a lack of validation.

2.6 *Limitations of skin permeability models: Solvent effects*

One of the major limitations of skin permeability models is that they are almost exclusively based on permeation from saturated aqueous solutions. This is due to Flynn's description that

the majority of environmental toxicology situations involve skin permeation from saturated aqueous solutions [10]. Thus, his and subsequent models were based on this premise. While addressing a significant issue, the Flynn approach does not tackle skin permeation based on exposures to non-aqueous solvents, which may also include finite-exposure scenarios; indeed, few researchers have addressed this issue. However, significant studies have developed quantitative models of skin permeation under such circumstances [49,50].

Similarly, these authors also developed similar QSAR methods to model solvent effects on the permeation of a dataset similar to that described above [51]. The study design used a porcine membrane and considered established models of skin permeability. Stepwise regression resulted in the following QSAR:

$$\log k_p = -0.909 - 0.610\log P + 2.26x_p - 0.00918(solBP - solMP) \quad (2)$$

$$n = 288, \quad s = 0.438, \quad r^2 = 0.729, \quad F = 255.2, \quad p = 0.000]$$

where: log P is the octanol–water partition coefficient; x_p is the nineth-order path molecular connectivity index; and ($SolBP$ – $SolMP$) is the difference between the boiling point and the melting point of the solvent system

This expression features two penetrant descriptors and one solvent mixture descriptor, highlighting the significance of solvent effects in skin permeation. Log P was determined to be the most significant descriptor and had a negative effect on skin permeation, which is significantly different from the vast majority of other QSAR models which do not consider non-aqueous solvents. This was attributed to the lipophilic nature of their model compounds which, while still sitting within a similar domain to most common QSAR models, was skewed towards a higher mean log P value. The x_p descriptor relates to the presence of chains of nine atoms in a molecule and is related, in this study, to an ideal molecular weight for skin permeation of approximately 350 Da. Finally, the difference in melting and boiling points is considered significant in the context of molecular symmetry.

The influence of skin biology on the utility of QSPR models was also explored by Riviere and Brooks [52]. Solvent mixtures (16 chemicals and 384 permutations) were applied to porcine skin *in vitro* and QSPRs were thus derived. They produced a QSPR which describes the effect of solvent mixtures and which was, significantly, dependent on the methodology employed to produce the data. They also described the known parabolic relationship between skin permeability and log P, with its maxima normally at log P 2–3, and demonstrated that this plateau is formulation dependent.

It is therefore clear that the advanced methods including fuzzy logic, machine learning and ANNs have significant application to pharmaceutical sciences and, in the case of percutaneous absorption, more broadly to the absorption of exogenous chemicals in many fields. But although these methods have been applied sporadically, they have seldom taken hold and established themselves within the field. This may be due to the lack of transparency that such methods feature, which limits any mechanistic interpretation and, in fields where significant regulatory hurdles exist for product safety and efficacy, the lack of transparency and possibly validation have been extremely negative issues. Moreover, the use of advanced or specialist methods outside their normal spheres of use is a significant issue, not only in ensuring competent and appropriate use but also in ensuring correct interpretation of results. While multidisciplinary practices are recommended, some overlap between disciplines is also necessary. Therefore, the great promise of these methods appears to have slowly dissipated and, despite significant limitations, the transparent and easy-to-apply regression methods, based mostly on multiple linear regression analysis, still dominate this research field.

In the remainder of this article, we address these issues in the context of Gaussian process machine learning methods. To achieve this, the codes used to construct the models are represented in the following sections as part of an overall description of how to conduct analysis with these methods. This aims to remove the specific technical difficulties associated with the application of these methods by researchers in a range of biological science fields. Ultimately, this might be best served by the development of an online resource which carries out these analyses for the uses. However, such a resource simply places an interface onto the model and ensures that it still lacks transparency, whereas the description of the process that follows should allow all users to develop their own Gaussian process based methods on any dataset.

3. Gaussian processes

3.1 *Background to Gaussian process regression modelling.*

Over the last decade Gaussian processes (GPs) have been increasingly favoured in the machine learning community. Indeed, theoretical and practical developments in the past decade have seen GPs become the method of choice in a range of applications and fields [53]. One example of this is in the field of percutaneous absorption [1,54], where the application of GPs has been assessed for their ability to predict the permeability of new chemicals and to provide mechanistic insight to the general process of skin permeation. It was found that the GP regression can yield considerable improvements over the QSAR models previously used for this purpose.

As far as we are aware, we are the only researchers to apply GP methods to the problem domain of percutaneous absorption. This lack of uptake is a major limitation to the GP method as, in general, we are essentially discussing our own work, which limits context and applicability. The lack of uptake is a major criticism of the GP method and reflects the lack of transparency in this, and related ML methods, which manifests itself quite significantly as the method does not yield an explicit functional representation of the process being modelled. However, this paper aims to address such a shortcoming by providing sufficient information, notably the MatLab instructions used in our previous research, to allow other researchers to use these methods.

The Gaussian process works in the following manner. Suppose we are given a number of data points, each with a fixed number of features (usually called the inputs); in general, 'features' in this context refer to commonly used physicochemical descriptors of a molecule, and include lipophilicity (as log P), molecular weight, melting point and the count of hydrogen bond donors and/or acceptors on a molecule. For each input, there is a corresponding target. The aim is to model the relationship between the inputs and their targets. For example, if we consider a dataset with 100 chemical compounds and five physicochemical features (lipophilicity, molecular weight, hydrogen bonding donor and acceptor groups, and melting point). The target is the defining output for the system which, for models of skin permeability, is usually the skin permeability coefficient (k_p) or the flux (J); for the discussion below, the use of k_p in models is considered. In this case, the aim of the modelling process is to infer a function from the dataset (using the available descriptors) to their related skin permeability coefficients, and to then use this function to predict the skin permeability coefficient for a new compound.

There are two alternative approaches to solving such problems (which are generally classified as regression problems) where the outputs represent the values of continuous variables. The first approach is to use parametric methods which assume a specific functional form for

the relationship between the inputs and targets. For example, targets of ANNs can be expressed either as a single layer or multiple layers of linear combinations of inputs (or hidden units, shown in Figure 2), optionally with non-linear activation functions on some of the inputs (or hidden units). The weightings on these combinations are free parameters. The second approach involves the use of non-parametric methods. Non-parametric methods do not assume a particular functional form, but allow the form of the relationship between inputs and targets to be determined entirely by the data.

The process of determining the values for the parameters on the basis of the dataset employed is called 'supervised learning' or 'supervised training'. Since the aim of the modelling is to find the algorithm having the best performance on new data, the simplest approach to compare different algorithms is to evaluate the error function of each model using data which is independent of that used for training the model. Various models are trained to minimise their particular error function defined with respect to a training dataset. In addition, the performance of each model should be evaluated using an independent test set.

Gaussian process modelling is a non-parametric method and does not produce an explicit functional representation of the data. In a GP model it is assumed that the underlying function that produces the data will remain unknown, but that the data are produced from a (potentially infinite) set of functions, with a Gaussian distribution in the function space. Thus, a Gaussian process is completely characterised by its mean and covariance function. Generally, a simplification of the models sees the mean function being described as the 'zero everywhere' function. Thus, all targets are generated from a joint Gaussian distribution with a mean of zero and a covariance function.

The covariance (or kernel) function, $k(x_i, x_j)$ is crucial in GP modelling. It expresses the expected correlation between values of $f(x)$ at two points, x_i and x_j. Thus, it defines the nearness – or similarity – between data points and allows for specifying *a priori* information from training data for solving the regression problem posed by the modelling experiments.

To make a prediction (y_*) at a new input (x_*) we need to compute the conditional distribution (defined by $p(y*|y1, \ldots yN_{trn})$, where N_{trn} denotes the number of training examples). As the model of interest is a Gaussian process, this distribution is also Gaussian and is completely defined by its mean and variance, which can be calculated using standard linear algebra [55]. Thus, the prediction of a new input is usually approximated by the corresponding mean value and the predictive variance can be obtained from the corresponding variance value. The mean at x_* is given by:

$$E[y_*] = k_*^T (K + \sigma_n^2 I)^{-1} y \tag{3}$$

where: k_* denotes the vector of covariances between the test point and the N_{trn} training data; I is the identity matrix (with ones on the diagonal and zeros elsewhere); σ_n^2 is the variance of an independent and identically distributed Gaussian noise (which infers that observations are noisy); and y denotes the vector of training targets.

In addition, it should be noted that $(K + \sigma_n^2 I)$ is a matrix of size $N_{trn} \times N_{trn}$ and is shown as:

$$\begin{bmatrix} k(x_1,x_1)k(x_1,x_2) \ldots k(x_1,x_{Ntrn},) \\ k(x_2,x_1)k(x_2,x_2) \ldots k(x_2,x_{Ntrn},) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ k(x_{Ntrn},x_1)k(x_{Ntrn},x_2) \ldots k(x_{Ntrn},x_{Ntrn},) \end{bmatrix} \tag{4}$$

The predictive variance at x_* is given by:

$$var[y_*] = k(x_*, x_*,) - k_*^T(K + \sigma_n^2 I)^{-1} k_* \quad (5)$$

where $k(x_*, x_*)$ denotes the variance of y_*.

In the output from these models the mean is used as the predicted value and the variance as the error bars of the prediction.

GP methods are non-parametric in nature and hence are able to use as much information as possible from a training dataset. Thus, it has been found that the GP regression algorithm produces smoother and less biased results [55].

3.2 *Gaussian processes regression: A simple example*

Suppose we have two values of x, for which we know the values of the dependant variable y, shown in Table 2. We want to predict the value of y for a new value of x, shown as x_* above. This is achieved in GP modelling by using a weighted average of the known values of y, with the weighting determined by the closeness of x_* to each value for x in the data. So, if $x_* = 2$, it is equally close to the original x values (1 and 3) so each is given equal weight and the prediction is the average of y, 3. So, if $x_* = 2.5$, which is 3 times as close to 3 as to 1, then a weighted average of the known values of y would be $[(0.75 \times 4) + (0.25 \times 2)]$, or 3.5. It should also be noted that using the reciprocal of the distance between points is limited as (1 / distance) approaches zero, yielding inconstant results.

In a GP model, the actual weighting is not exactly proportional to the separation from the known values; rather, it is a Gaussian function of that distance. For example, Figure 3(a) shows the Gaussian function of distance plotted for both the data points used in above example. The blue Gaussian represents the weighting of $x = 1$ and the red Gaussian the weighting of $x = 3$, and it is apparent that they are identical Gaussian curves which only vary on their centres.

The green line in Figure 3(a) represents the prediction based on the weighted sum of the two known values of y, as: green$(x*)$ = [blue$(x*)$ × 2] + [red$(x*)$ × 4]. However, closer inspection of Figure 3(a) shows that the green line seems to pass slightly above the value it should take at both 1 and 3. This is seen more clearly if the second data point is moved left to $x = 2$, as shown in Figure 3(b). The prediction is clearly not an interpolation (that is, a function that passes through the known points). The reason for this observed effect is that both Gaussian plots are contributing relatively large weights at both data points. For example, the red Gaussian is still at a value around 0.5 for $x = 1$, meaning that it will contribute approximately (0.5 × 4) to the prediction at that point, which explains why the prediction is nearly double what it should be.

To overcome this problem, the Gaussian weighting functions are modified so that they are forced to zero as they pass through the other data points, as is shown in Figure 3(c). This results in the green function being an interpolation – it passes through both data points; at

Table 2. Example data used to demonstrate the Gaussian process model.

x	y
1	2
3	4
x_*	?

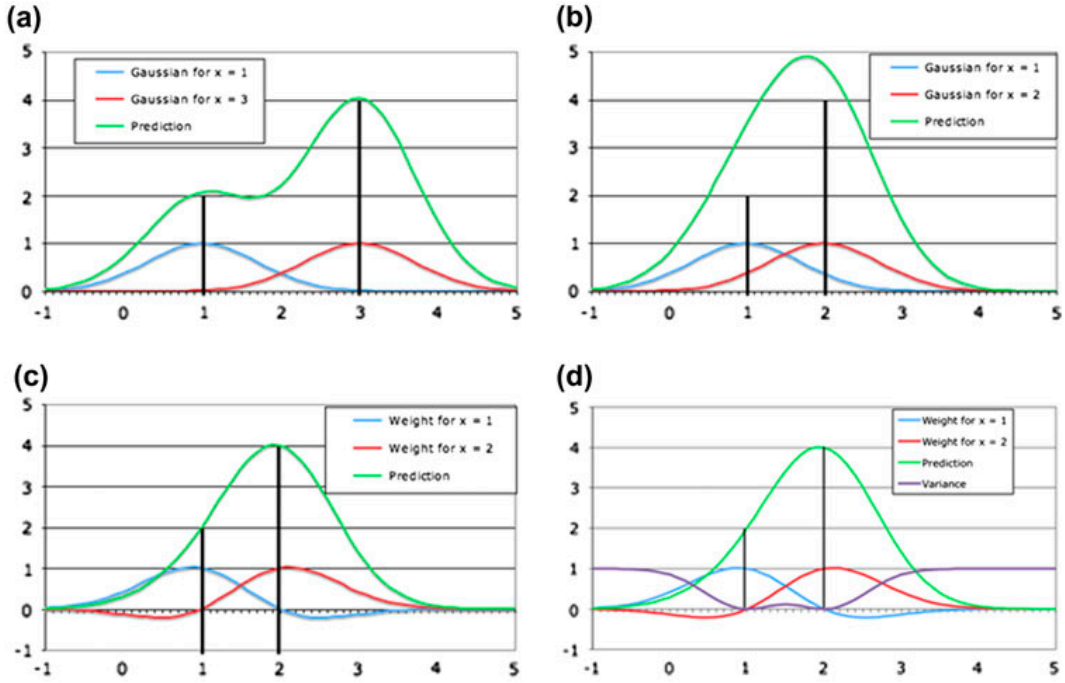


Figure 3. Schematic representations of Gaussian weights: (a) an example of the prediction using simple Gaussian weights; (b) an example of the prediction using Gaussian weights, showing the different values for x and y and how they affect the prediction; (c) an example of a correctly weighted Gaussian process; and (d) an example of a correctly weighted Gaussian process with variance shown.

both data points one of the weights is one and the other is zero. With regard to Equation (1), $k*$ is the vector of values of the Gaussian weights for the new point, $x*$, with each of the known x values. The matrix $(K + \sigma_n^2 I)^{-1}$ is a linear transformation, independent of any new data and thus only computed once; this forces the Gaussian weights to perform interpolation; that is, to be zero for all but one of the known data points. Finally, the transformed weights are used to calculate the prediction as a weighted sum over all the y values.

One further significant feature of the GP regression is that it provides not only a prediction but also the variance associated with the prediction, which can be interpreted as the reciprocal of a confidence value. This has significant application for the real-world modelling of biological functions, as it provides a mean and variance for the prediction – effectively, a very simple method to produce a range which has significance not only for the output but in the context of the inherently variable nature of the inputs used in building such models. The variance is essentially calculated from the size of the vector of weights. If the weights are large then the variance should be low, and vice versa, as high weights mean that the new data point must be near some of the original data and the prediction would, in such a case, have a high level of confidence associated with it. Figure 3(d) adds variance to the examples of Gaussian plots illustrated in Figures 3(a) to 3(c). Figure 3(d) shows that the variance is zero at the actual data points and increases towards one as the x-value moves away from the known data points. It should be noted that the 'variance' is only correctly the variance if the original data has been normalised as Z-scores; nevertheless, in the examples discussed above and shown in Figure 3, it is perfectly reasonable to treat it as a confidence value.

3.3 *Gaussian processes: Applied to pharmaceutical examples*

The major, and valid, criticism of Gaussian processes lies not in its theoretical basis but the inability of non-experts to apply these methods to their fields. While Cronin and Schultz [3] have commented that research groups should feature experts from each discipline, it is often the case that statistical and mathematical methods which become popularised in the biological sciences are those which can be readily used by researchers in those fields. Therefore, this and subsequent sections discuss the practical basis of undertaking a Gaussian process regression, using a small dataset as a guide. The settings of parameters for these models are then discussed.

3.3.1 *Step 1: Get some data*

The issues with the underlying nature and quality of datasets (specifically in the modelling of percutaneous absorption) have been discussed in detail previously [8] and is a broader subject than can be addressed here. The data used in this section are taken from a dataset collated from other literature sources and presented previously [6]. They are shown in Table 3 and Figure 4. Five compounds have been taken from the dataset as training examples and two compounds as test examples. Thus, to run this analysis in MatLab the scripts need to be loaded and the data added to the program.

Loading and opening scripts. The description below relates to MatLab 2014a. To add a new code script, under the Editor tab select 'New' and then 'Script'. This opens a new page into which scripts can be written or pasted; included as supplemental material (available via the multimedia link on the online article webpage) is the full script for performing a Gaussian process analysis,which can be found as a zip file at: https://mega.co.nz/#!jQcBSbiJ!GUkuvTesV-graw1Qw13KjkKdqNTjpo1ttrtWuW-lt4Q. Alternatively, an existing script can be added using the Editor tab to load an existing script. In the MatLab *working directory*, navigate to the directory containing the extracted zip file; right-click the appropriate folder and select 'Add to path' with 'Selected folders and subfolders'. Open the folder and double-click on the file 'Main.m'.

Adding training and test datasets. If the data source is being imported from a source other than Microsoft Excel, the code provided will require minor modification. For example, the code provided states the following instructions:

Table 3. Sample data taken from [6].

	Name	log k_p	log P	*Molecular weight*
Training set	Atropine	–4.12	1.83	289.37
	Cortexolone	–4.13	3.25	346.46
	Ethanolamine	–4.02	–1.31	61.08
	Monomethylhydrazine	–3.75	–1.05	46.07
	Morpholine	–3.86	–0.86	88.12
Test set	Aldosterone	–4.24	1.08	360.44
	Methyl cellosolve	–3.73	–0.77	76.09

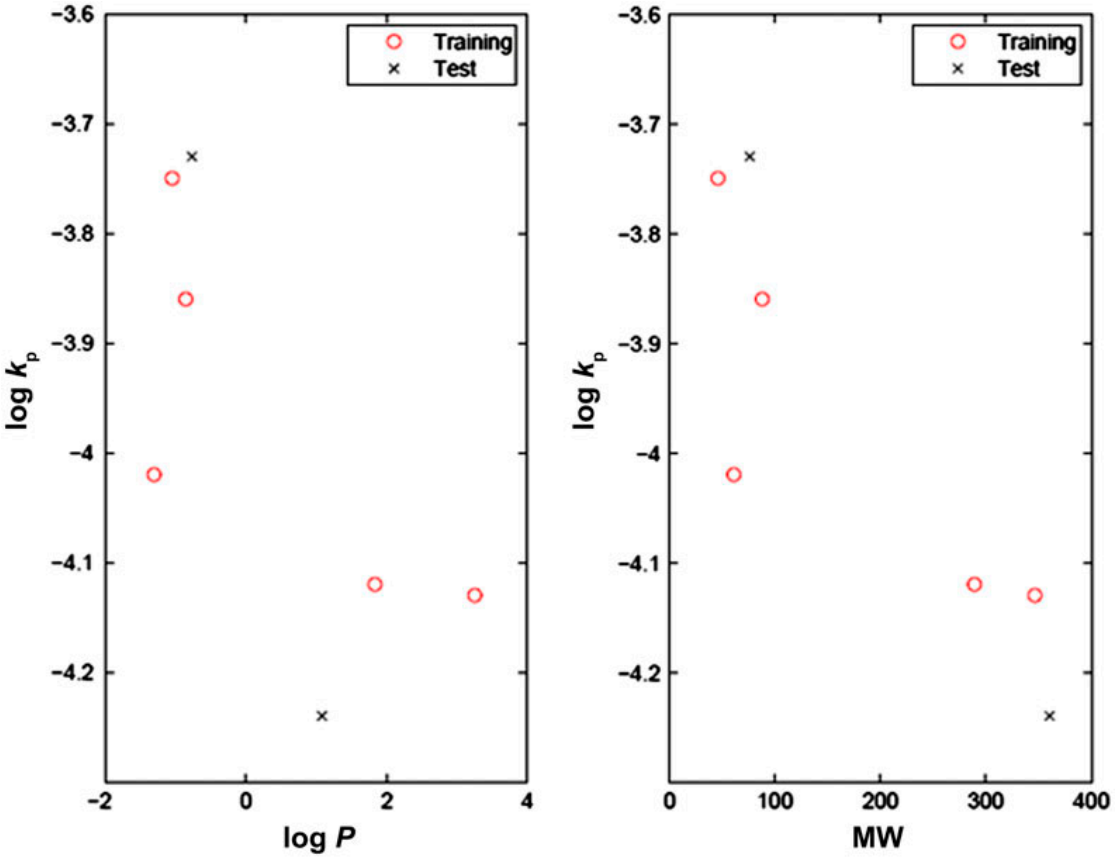


Figure 4. Plot of the log k_p values against molecular weight (MW) and log *P* values for the sample dataset shown in Tables 3 and 4.


```
% Reading train data from excel sheets
Data = xlsread('Train.xlsx');
% Reading test data from excel sheets
Test = xlsread('Test.xlsx');
```

where lines beginning with '%' are comments only. To modify these using specific data, such as that in the given example, the following codes can be added in place of the above lines:

Data = [1.83 289.37 −4.12;3.25 346.46 −4.13;−1.31 61.08 −4.02;−1.05 46.07 −3.75;−0.86 88.12 −3.86];

Test = [1.08 360.44 −4.24;−0.77 76.09 −3.73];

To run a script in Matlab, under the 'Editor' tab select the Run command. It should be noted that the target values (log k_p) are moved to the last columns of the matrix. There are some options in the code to select the covariance function and its hyperparameters which are explained later in this paper, but for now we would recommend keeping them as they are.

3.3.2 *Step 2: Normalise the data*

As can be seen in Table 3, the values of molecular weight differ from the values of log *P* by several orders of magnitude. Thus, the importance of log *P* in determining the outputs may

be diminished by the values of molecular weight. Thus, the data should be normalised to address this issue. One way to resolve this issue is to calculate the mean and standard deviation for each variable in the training dataset, and to then subtract the corresponding mean value from each variable. Then, each variable is re-scaled by dividing it by the standard deviation. In this way, all the variables have zero means and unit standard deviations. For outputs, the mean value is subtracted from target values.

Table 4 shows the data following this normalisation procedure. It should be noted that the test set is normalised using the mean and standard deviation values from the training set since it is assumed that the test data are generated from the same distribution as the training dataset.

In Matlab the *zscore(Data);* command is used to normalise the training set. To normalise the test set to synergise with the training data the following codes are used, which take the mean and standard deviation of the training set and then subtract the mean from the tests set and divide the obtained value by the standard deviation:

```
meanP = repmat(mean(x),R,1);
stdP = repmat(std(x),R,1);
normalZ = (z-meanP)./stdP;
```

where R is the number of rows in the test set. Usually, nothing in the code needs to be changed as, when you run the 'Main' script, the data normalisation is performed automatically.

3.3.3 *Step 3: Select kernel and set parameters*

Different covariance, or kernel, functions may be applied to Gaussian process regression. They include, for example, linear, x^2 or polynomial kernels. In this example, the simple squared exponential covariance function is considered and its parameters are set as below. The use of different covariance functions may be achieved by using the required covariance function in the 'Main' script. The available options are: *covSEiso* (*Squared Exponential* covariance function), *covMaterniso* (*Matérn* covariance function) and *covNNone* (*Neural Network* covariance function). Only one function can be used at the same time and the default function is *covSEiso*.

3.3.4 *Step 4: Calculate the kernel matrix of the training set*

Each element of the matrix M (equal to $(K + \sigma_n^2 I)$ is computed as follows:

$$K(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_i - x_j)^T(x_i - x_j)\right) + \sigma_n^2 \delta_{ij} \tag{6}$$

Table 4. Normalised sample data.

	Chemical name	log k_p	log P	*Molecular weight*
Training set	Atropine	–0.14	0.71	0.88
	Cortexolone	–0.15	1.41	1.28
	Ethanolamine	–0.04	–0.82	–0.75
	Monomethylhydrazine	0.23	–0.69	–0.85
	Morpholine	0.12	–0.60	–0.55
Test set	Aldosterone	–2.64	0.34	1.38
	Methyl cellosolve	0.25	–0.56	–0.64

Original data taken from [6].

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

In this case, $\sigma_f^2 = 0.0079$, $l^2 = 0.9952$ and $\sigma_n^2 = 0.0144$. These values are the best hyperparameters obtained using the maximised marginal likelihood technique. This is, however, just an example of how the kernels are calculated. In practice, readers can use either maximised maginal liklehood (MML) or the cross-validation method to choose the most suitable hyperparameters. This is discussed later in the 'model selection' section. The function is run in MatLab by using the command:

hyp3 = minimize(hyp, @gp, −100, @infExact, [], covfunc, likfunc, normalX, y_final);

The *minimise* function minimises a differentiable multivariate function using the conjugate gradient technique, which is a well-known numerical optimisation technique.

The variance of a training example itself (that is, when $i = j$) is given by:

$$k(x_i, x_j) = \sigma_f^2 + \sigma_n^2 \tag{7}$$

and in this case is equal to 0.0079 + 0.0144 = 0.0223.

Now, it is important to consider the similarity between examples 1 and 2. Since $\delta_{12} = 0$ we have the following expression:

$$k(x_i, x_j) = 0.0079 \exp\left(-\frac{1}{2x0.9552}(x_1 - x_2)^T(x_1 - x_2)\right) \tag{8}$$

The value of $k(x_1, x_2)$ is determined by the Euclidean distance of these two points, which is calculated as follows:

$$\begin{aligned}(x_1 - x_2)^T(x_1 - x_2) &= (x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 = (0.7119 - 1.4053)^2 + 0.8750 - 1.2806)^2 \\ &= 0.6453\end{aligned}$$

And we now have:

$$k(x_1, x_2) = 0.0079 \times exp\left(-\frac{1}{2 \times 0.9552} \times 0.6453\right) = 0.0056$$

Similarly, all elements between two different data points can be calculated in the same way, resulting in a covariance matrix, M:

$$\begin{bmatrix} 0.0223 & 0.0055 & 0.0055 & 0.0055 & 0.0010 \\ 0.0055 & 0.0223 & 0.0001 & 0.0001 & 0.0001 \\ 0.0005 & 0.0001 & 0.0223 & 0.0078 & 0.0075 \\ 0.0005 & 0.0001 & 0.0078 & 0.0223 & 0.0075 \\ 0.0010 & 0.0001 & 0.0075 & 0.0075 & 0.0223 \end{bmatrix}$$

All the mentioned calculations are performed when the 'Main' script is run. As an example, the calculation of the kernel function is given below for the squared exponential covariance function in the *covSEiso* function; thus, to calculate K between training set use the following coding:

K = sq_dist(x'/ell);

Followed by:

*K = sf2*exp(-K/2);*

where *x* is the training set; *ell* is the *length scale*; *sf*2 is the *signal variance*; *exp* is a function built into MatLab which takes the exponential of a value; and *sq_dist* is a function provided in the supplemental material (available via the multimedia link on the online article webpage) which calculates the similarity between points.

3.3.5 *Step 5: Compute the inverse of the kernel matrix*

The inverse of a square matrix A, which is often called a reciprocal matrix, is a matrix A^{-1} such that $A.A^{-1} = I$, the identity matrix. Many high-level computing packages, such as Octave or MatLab, provide an in-built feature that automatically computes the inverse of a square matrix. In MatLab (R2012a) [56,57], the inverse of matrix M can be obtained by using the command:

```
>> inv(M)
```

and the result is shown as follows:

$$\begin{bmatrix} 47.8489 & -11.7902 & -0.2814 & -0.2814 & -1.9036 \\ -11.7902 & 47.8489 & -0.0495 & -0.0495 & 0.3479 \\ -0.2814 & -0.0495 & 54.4407 & -14.5248 & -13.4118 \\ -0.2814 & -0.0495 & -14.5248 & 54.4407 & -13.4118 \\ -1.9036 & 0.3479 & -13.4118 & -13.4118 & 53.9482 \end{bmatrix}$$

In practice, rather than directly inverting the matrix, Cholesky decomposition is recommended as it is faster and numerically more stable [56]. As an example, the inverse function is calculated in MatLab using the coding:

```
sn2 = exp(2*hyp.lik);
L = chol(K/sn2+eye(n));
alpha = solve_chol(L,y-m)/sn2;
```

where: *sn*2 is *signal noise*; *hyp.lik* is the likelihood function of the kernel; *n* is the rows number of *x*; *m* is the mean function which is set to zero in all examples; *Chol* is a Matlab built-in function; and *solve_chol* is a function provided in the supplemental material which implements the functionality of the Cholesky decomposition.

However, it should be noted that the main aspects of the code are contain in the 'Main' script which, to run a rudimentary analysis, does not require modification.

3.3.6 *Step 6: Compute k_**

Using the method described in Step 4, we can obtain k_*^T:

$$\begin{bmatrix} 0.0064 & 0.0042 & 0.0003 & 0.0003 & 0.0006 \\ 0.0009 & 0.0001 & 0.0075 & 0.0076 & 0.0078 \end{bmatrix}$$

In MatLab, to calculate *K* between the training and test sets the following command is used:

```
K = sq_dist(x'/ell,z'/ell);
```

Followed by:

```
K = sf2*exp(-K/2);
```

where: *x* is the training set; *z* is the test set; *ell* is the *length scale*; and *sf*2 is the *signal variance.*

3.3.7 *Step 7: Predict the skin permeability*

The prediction of skin permeability can now be computed by multiplying k_*^T, $(K + \partial_n^2 I)^{-1}$ and y. In the case of the example data used here, this returns values of 0.0201 and 0.0176. The mean value of targets in the training dataset (in this case, –3.9760) is added. Thus, the final predicted log k_p value for aldosterone is –4.0303 and the final predicted log k_p value for methyl cellosolve is –3.9170, which compare well with the values listed in Table 3. To obtain the prediction using MatLab the 'Main' script should be run; the output will depend on the covariance function chosen. The code used to obtain the predictions is:

```
[Prediction var MSE_GP ION_GP corrcoef_GP] = Result_CV (Data, Test);
```

The outputs (predictions) will be displayed in the 'work space' window in MatLab, in which the predictions obtained are displayed, followed by the performance measures *MSE* (*Mean Squared Error*), *ION* (Improve over *Naïve* model) and *Correlation Coefficient* (*CorrCoef*). Higher values of *ION* and *CorrCoef* generally mean better GP performances. However, high values of *MSE* show poor performance of the model.

3.3.8 *Step 8: Compute variance*

The final step is to calculate the variance, σ_*^2. This is calculated as $\sigma_*^2 = var[y_*] + \sigma_n^2$, where $var[y_*]$ is given by Equation (2). The noise variance, σ_n^2, is included as function values themselves are not assessed but they are assessed as values with a noise. In this case the values of $var[y_*]$ are 0.0201 and 0.0176, and σ_n^2 is 0.0144. Therefore, the predictive variances of aldosterone and methyl cellosolve are 0.0345 and 0.0320, respectively. In MatLab the variation is obtained from the code discussed in Step 7 and is saved in 'var' variable.

3.4 *Model selection*

Generally, cross-validation is applied to choose a suitable kernel function and its parameters. Since the key aim of this method is to find the GP model with the best performance on new data points, the simplest appraoch to the comparison of different models is to evaluate the error function using data which is independent of that used for training the model [53]. In the methods described here, and published previously [1,2,39,54], the training set is randomly divided into S distinct segments. A GP regression is then trained from $(S - 1)$ of the segments with its performance being assessed by evaluation of the error function using the remaining segment, the validation dataset. This process is repeated for each of S possible choices. Several different sets of parameters may be pre-set, as required. This process is repeated for each set and an average of S validation dataset errors can be calculated. Thus, a final model with a minimum average of validation set errors may be selected. Since this procedure may potentially lead to over-fitting of the validation set, the performance of the selected model should be confirmed by measuring its performance on a third, independent set of data called the 'test set' [53].

When the training set is small, a special cross-validation method – the leave-one-out technique – of often used. This means that one compound is used for testing and all others are used for training, and this is repeated for each member of the dataset in turn. In the Matlab code provided in the supplemental material which is available via the multimedia link on the online article webpage, this is achieved using the 'Result_CV' function. A simple example of this process is shown below:

3.4.1 *Simple example*

Suppose we have 100 data points, 50 in class A and 50 in B:

(1) Remove a random test set of 20 points with equal numbers of A and B (10 each).
(2) Now remove a further 20 points from the training set, as a validation set (10 A, 10 B).
(3) For each of the following values of *length scale* and s*ignal variance* (for example) l from {0.1, 10}, ∂_f^2 from {0.1, 3.16}. For simplicity, the noise is constantly considered to be log(0.1):
    (a) Train the GP on the 60 remaining points.
    (b) Report the performance (for example *ION, MSE, CORR*) of the trained model on the validation set.
    (c) Choose the values of l and ∂_f^2 that give the best result.
(4) Using the best parameter values train the GP on the 80 points in the full training set.
(5) Finally report the performance on the test set.

There is one final consideration, which is that the value of l and ∂_f^2 found by cross-validation may only work well for the particular training/validation set used and this might lead to poor performance on the test set. So, typically five- (or ten-) fold cross-validation is used. In this case the training set is divided into five equally sized subsets, which means that five different training/validation pairs can be constructed, from which an average accuracy (for a given l/∂_f^2) can be calculated. For example, if the training data is divided into five subsets (A, B, C, D, E), then the five different train/validate pairings are:

- Train on A B C D and validate on E
- Train on A B C E and validate on D
- Train on A B D E and validate on C
- Train on A C D E and validate on B
- Train on B C D E and validate on A

So, for a given parameter setting, an overall performance measure can be found by taking the average of the five individual accuracies. This gives a more reliable assessment of a particular l/∂_f^2 setting. Different covariance functions can then be assessed to obtain the best performing results. It should be noted that this process occurs before the test set has been predicted.

Another way to set parameters is to maximise the marginal likelihood. Parameters obtained at the maximised marginal likelihood are the most suitable for the model. In practice, the initial values for parameters are usually used first and then numerical methods for non-linear partial differential equations may be involved to search for the best parameter values. The 'Main' Matlab script has two options for selecting the hyperparameters: cross validation or maximised maginal liklehood (MML). Finding the best hyperparameters by cross-validation requires an expert to consider proper ranges for the hyperparameters and their initialisation. Using automatic hyperparameters (with MML) makes the process easier, but;it may end up with local minima or maxima. So, automatic hyperparameter selection is generally recommended and there is no need to change the default hyperparameter selection method in MatLab. The supplemental code lets the user choose the method to find the hyperparameters and also select the covariance functions:

```
% Change the way to optimise the hyperparameters
% Please choose 1 for MML hyperparameter optimisation or 2 for Cross
% validation hyperparameter optimisation, default is 1
opt_hyper_parameter =1;
```

% change the covariance function
% Please choose 1 for SE Covariance function, 2 for Matérn Covariance
% function and 3 for Neural Network covariance function, default is 1
Covariance_func =1;

3.5 *Software*

A range of software packages may be used to undertake Gaussian process regression analysis of datasets. The 'Gaussian Process' webpage [56] provides a comprehensive list of available packages. The GP software used in the analysis of skin permeability datasets and highlighted above is based in Matlab (R2012a) [57]. Specific toolboxes and add-on packages are available elsewhere [58]. Figure 5 shows a schematic representation of how the software is used to perform the Gaussian process regression.

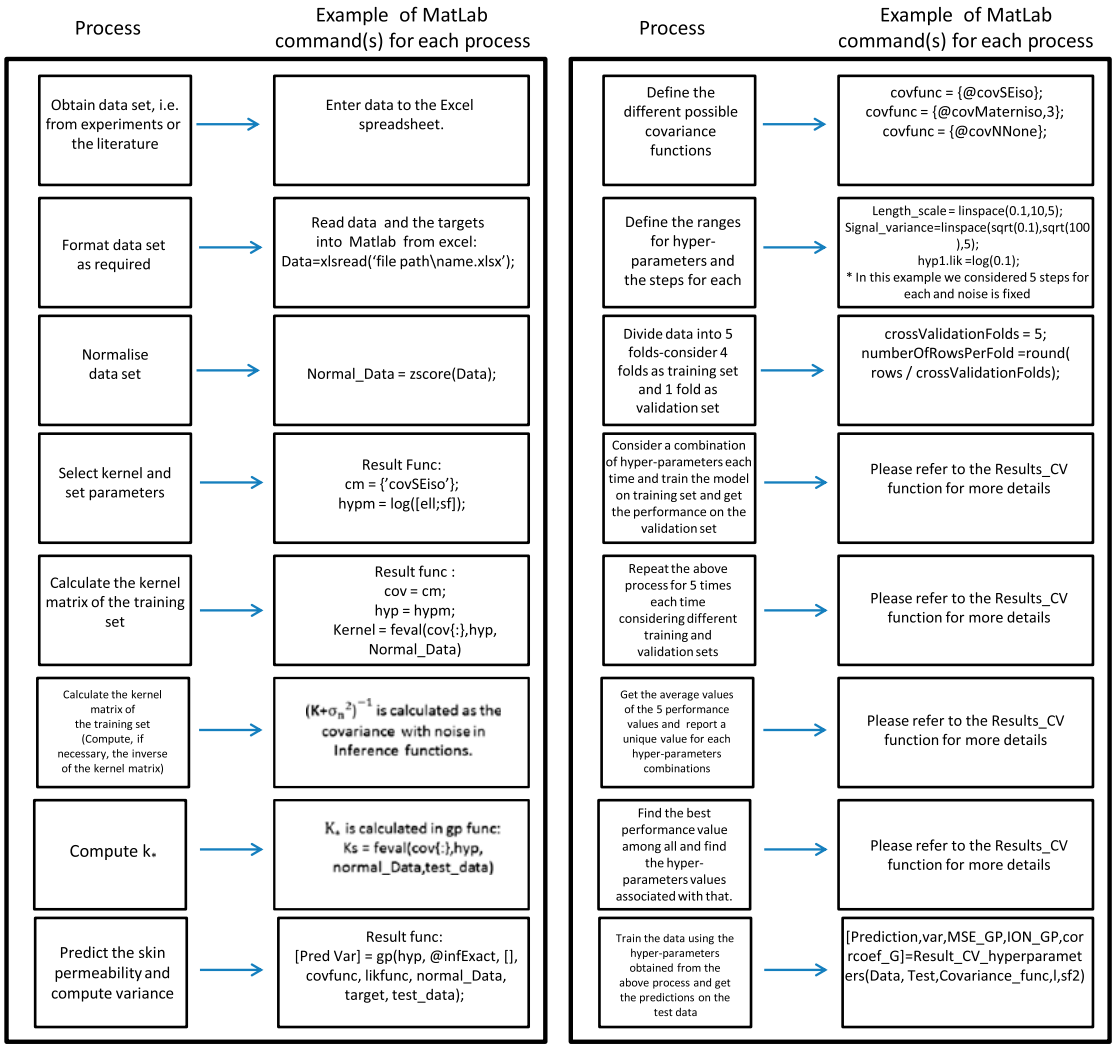


Figure 5. Flow chart for undertaking a Gaussian processs regression in MatLab, with (left) examples of the commands used for each step of the process and (right) examples of how to complete a five-fold cross-validation analysis, including selection of the covariance function and its hyperparameters.

4. Conclusions

Gaussian process regression has been applied to the problem domain of skin permeability, both in terms of predicting permeability for new chemicals and in developing our understanding of the mechanisms of skin permeability. Such applications are not necessarily confined to skin permeability, but clearly have application in other models of membrane permeability. The GP models have consistently shown better predictive ability than quantitative structure–permeability models for skin and, with the use of feature selection and related methods [2], they also demonstrate the ability – in the absence of a descriptive algorithm – to yield significant mechanistic information. This is achieved by a statistical comparison of all possible permutations of descriptors employed in this study, as described in the Introduction. In the case of Lam's study [2], this involved the potential interchangeability of physicochemical descriptors, where a number of different permutations of physicochemical descriptors resulted in models that were, in terms of statistical quality and predictive ability, the same. The implicit understanding from this study is that the use of algorithms featuring discrete terms may not necessarily be the most appropriate models to use.

Therefore, the use of Gaussian processes offers a potentially powerful and widely applicable tool for the analysis of complex biological processes. It has shown an ability to provide higher quality predictive models compared to existing methods and to provide mechanistic insight to the processes associated with permeation.

References

[1] G.P. Moss, Y. Sun, M. Prapopoulou, N. Davey, R. Adams, W.J. Pugh, and M.B. Brown, *The application of Gaussian Processes to the prediction of percutaneous absorption*, J. Pharm. Pharmacol. 61 (2009), pp. 1147–1153.

[2] L.T. Lam, Y. Sun, N. Davey, R.G. Adams, M. Prapopoulou, M.B. Brown, and G.P. Moss, *The application of feature selection to the development of Gaussian Process models for percutaneous absorption*, J. Pharm. Pharmacol. 62 (2010), pp. 738–749.

[3] M.T.D. Cronin and T.W. Schultz, *Pitfalls in QSAR*, J. Mol. Struct. 622 (2003), pp. 39–51.

[4] R.O. Potts and R.H. Guy, *Predicting skin permeability*, Pharm. Res. 9 (1992), pp. 663–669.

[5] B.M. Magnusson, Y.G. Anissimov, S.E. Cross, and M.S. Roberts, *Molecular size as the main determinant of solute maximum flux across the skin*, J. Invest. Dermatol. 122 (2004), pp. 993–999.

[6] G.P. Moss and M.T.D. Cronin, *Quantitative structure–permeability relationships for percutaneous absorption: Re-analysis of steroid data*, Int. J. Pharm. 238 (2002), pp. 105–109.

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

[8] G.P. Moss, J.C. Dearden, H. Patel, and M.T.D. Cronin, *Quantitative structure-permeability relationships (QSPRs) for percutaneous absorption*, Toxicol. In Vitro 16 (2002), pp. 299–317.

[9] A.C. Williams, *Transdermal and Topical Drug Delivery*, Pharmaceutical Press, London, 2003.

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

[11] R.O. Potts and R.H. Guy, *A predictive algorithm for skin permeability: The effects of molecular size and hydrogen bond activity*, Pharm. Res. 12 (1995), pp. 1628–1633.

[12] S. Mitragotri, Y.G. Anissimov, A.L. Bunge, H.F. Frasch, R.H. Guy, J. Hadgraft, G.B. Kasting, M.E. Lane, and M.S. Roberts, *Mathematical models of skin permeability: An overview*, Int. J. Pharm. 418 (2011), pp. 115–129.

[13] A.K. Pannier, R.M. Brand, and D.D. Jones, *Fuzzy modelling of skin permeability coefficients*, Pharm. Res. 20 (2003), pp. 143–148.

[14] D.R. Keshwani, D.D. Jones, and R.M. Brand, *Tagaki–Sugeno fuzzy modelling of skin permeability*, Cutan. Ocul. Toxicol. 24 (2005), pp. 149–163.
[15] S. Agatonovic-Kustrin and R. Beresford, *Basic concepts of artificial neural network (ANN) modelling and its application in pharmaceutical research*, J. Pharm. Biomed. Anal. 22 (2000), pp. 717–727.
[16] S. Agatonovic-Kustrin, R. Beresford, A. Pauzi, and M. Yusof, *ANN modeling of the penetration across a polydimethylsiloxane membrane from theoretically derived molecular descriptors*, J. Pharm. Biomed. Anal. 26 (2001), pp. 241–254.
[17] J. Bourquin, H. Schmidt, P. Van Hoogevest, and H. Leuen-Berger, *Application of artificial neural networks (ANN) in the development of solid dosage forms*, Pharm. Dev. Technol. 2 (1997), pp. 111–121.
[18] J. Bourquin, H. Schmidt, P. Van Hoogevest, and H. Leuen-Berger, *Comparison of artificial neural networks (ANN) with classical modelling technologies using different experimental designs and data from a galenical study on a solid dosage form*, Eur. J. Pharm. Sci. 7 (1998), pp. 1–12.
[19] J. Takahara, K. Takayama, and T. Nagai, *Multi-objective simultaneous optimization technique based on an artificial neural network in sustained release formulations*, J. Cont. Rel. 49 (1998), pp. 11–20.
[20] R.G. Alany, S. Agatonovic-Kustrin, T. Rades, and I.G. Tucker, *Use of artificial neural network to predict quaternary phase systems from limited experimental data*, J. Pharm. Biomed. Anal. 19 (1999), pp. 443–452.
[21] T. Fan, K. Takayama, Y. Hattori, and Y. Maitani, *Formulation optimisation of paclitaxel carried by PEGylated emulsions based on artificial neural network*, Pharm. Res. 21 (2004), pp. 1692–1697.
[22] Y. Hayashi, S. Kikuchi, Y. Onuki, and K. Takayama, *Reliability of nonlinear design space in pharmaceutical product development*, J. Pharm. Sci. 101 (2012), pp. 333–341.
[23] Z. Sun, X. Rao, L. Peng, and D. Xu, *Prediction of protein supersecondary structures based on the artificial neural network method*, Protein Eng. 10 (1997), pp. 763–769.
[24] C.H. Wu, *Artificial neural networks for molecular sequence analysis*, Comput. Chem. 21 (1997), pp. 237–256.
[25] J. Sun, W.Y. Song, L.H. Zhu, and R.S. Chen, *Analysis of tRNA gene sequences by neural network*, J. Comput. Biol. 2 (1995), pp. 409–416.
[26] M. Tacker, P.F. Stadler, E.G. Bornbergbauer, I.L. Hofacker, and P. Schuster, *Algorithm independent properties of RNA secondary structure predictions*, Eur. Biophys. J. 25 (1996), pp. 115–130.
[27] C.M. Reidys, P.F. Stadler, and P. Schuster, *Generic properties of combinatory maps – Neural networks of RNA secondary structures*, Bull. Math. Biol. 59 (1997), pp. 339–397.
[28] P. Willett, *Genetic algorithms in molecular recognition and design*, Trends Biotechnol. 13 (1995), pp. 516–521.
[29] S.S. So and M. Karplus, *Evolutionary optimization in quantitative structure–activity relationship: An application of genetic neural networks*, J. Med. Chem. 39 (1996), pp. 1521–1530.
[30] S.S. So and M. Karplus, *Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 1. Methods and validations*, J. Med. Chem. 40 (1997), pp. 4347–4359.
[31] S.S. So and M. Karplus, *Three-dimensional quantitative structure–activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications*, J. Med. Chem. 40 (1997), pp. 4360–4371.
[32] T. Degim, J. Hadgraft, S. Illbasmis, and Y. Ozkan, *Prediction of skin penetration using artificial neural network (ANN) modelling*, J. Pharm. Sci. 92 (2003), pp. 656–664.
[33] M. Hosseini, D.J. Madalena, and I. Spence, *Using artificial neural networks to classify the activity of capsaicin and its analogues*, J. Chem. Inf. Comput. 37 (1997), pp. 1129–1137.
[34] L.J. Chen, G.P. Lian, and L.J. Han, *Prediction of human skin permeability using artificial neural network (ANN) modelling*, Acta Pharmacol. Sinica. 28 (2007), pp. 591–600.
[35] M.H. Abraham, F. Martins, and R.C. Mitchell, *Algorithms for skin permeability using hydrogen bond descriptors: The problem of steroids*, J. Pharm. Pharmacol. 49 (1997), pp. 858–865.

[36] C.W. Lim, S. Fujiwara, F. Yamashita, and M. Hashida, *Prediction of human skin permeability using a combination of molecular orbital calculations and artificial neural network*, Bio. Pharm. Bull. 25 (2002), pp. 361–366.
[37] S. Siani, S.K. Singh, A. Garg, K. Khanna, A. Shandil, and D.N. Mishra, *Prediction of skin penetration using artificial neural network*, Int. J. Eng. Sci. Tech. 2 (2010), pp. 1526–1531.
[38] Y. Sun, G.P. Moss, N. Davey, R. Adams, and M.B. Brown, *The application of stochastic Machine Learning methods in the prediction of skin penetration*, App. Soft Comput. 11 (2011), pp. 2367–2375.
[39] S. Kia, S. Setayeshi, M. Shamsaei, and M. Kia, *Computer-aided diagnosis (CAD) of the skin disease based an intelligent classification of sonogram using neural network*, Neural Comput. Applic. 22 (2013), pp. 1049–1062.
[40] B. Baert, E. Deconinck, M. van Gele, M. Slodicka, P. Stoppie, S. Bode, G. Slegers, Y. van der Heyden, J. Lambert, J. Beetens, and B. de Spiegeleer, *Transdermal penetration behaviour of drugs: CART-clustering, QSPR and selection of model compounds*, Bioorg. Med. Chem. 15 (2007), pp. 6943–6955.
[41] M.S. Roberts, W.J. Pugh, J. Hadgraft, and A.C. Watkinson, *Epidermal permeability-penetrant structure relationships: 1. An analysis of methods of predicting penetration of monofunctional solutes from aqueous solutions*, Int. J. Pharm. 126 (1995), pp. 219–233.
[42] M.S. Roberts, W.J. Pugh, and J. Hadgraft, *Epidermal permeability–penetrant structure relationships. 2. The effect of H-bonding groups in penetrants on their diffusion through the stratum corneum*, Int. J. Pharm. 132 (1996), pp. 23–32.
[43] W.J. Pugh, M.S. Roberts, and J. Hadgraft, *Epidermal permeability–penetrant structure relationships: 3. The effect of hydrogen bonding interactions and molecular size on diffusion across the stratum corneum*, Int. J. Pharm. 138 (1996), pp. 149–165.
[44] B. Neely, S. Madihally, R.J. Robinson, and K. Gasem, *Nonlinear quantitative structure–property relationship modeling of skin permeation coefficient*, J. Pharm. Sci. 98 (2009), pp. 4069–4084.
[45] G.P. Moss, A.J. Shah, R.G. Adams, N. Davey, S.C. Wilkinson, W.J. Pugh, and Y. Sun, *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. 45 (2012), pp. 116–127.
[46] R.J. Scheuplein, I.H. Blank, G.I. Brauner, and D.J. MacFarlane, *Percutaneous absorption of steroids*, J. Invest. Dermatol. 52 (1969), pp. 63–70.
[47] M.E. Johnson, D. Blankstein, and R. Langer, *Permeation of steroids through human skin*, J. Pharm. Sci. 84 (1995), pp. 1144–1146.
[48] R.P. Chilcott, N. Barai, A.E. Beezer, S.L. Brain, M.B. Brown, A.L. Bunge, S.E. Burgess, S. Cross, C.H. Dalton, M. Dias, A. Farinha, B.C. Finnin, S.J. Gallagher, D.M. Green, H. Gunt, R.L. Gwyther, C.M. Heard, C.A. Jarvis, F. Kamiyama, G.B. Kasting, E.E. Ley, S.T. Lim, G.S. McNaughton, A. Morris, M.H. Nazemi, M.A. Pellett, J. Du Plessis, Y.S Quan, S.L. Raghavan, M. Roberts, W. Romonchuk, C.S. Roper, D. Schenk, L. Simonsen, A. Simpson, B.D. Traversa, L. Trottet, A. Watkinson, S.C. Wilkinson, F.M. Williams, A. Yamamoto, and J. Hadgraft, *Inter- and intra-laboratory variation of in vitro diffusion cell measurements: An international multicenter study using quasi-standardised methods and materials*, J. Pharm. Sci. 94 (2005), pp. 632–638.
[49] J.E. Riviere and J.D. Brooks, *Predicting skin permeability from complex chemical mixtures*, Toxicol. Appl. Pharmacol. 208 (2005), pp. 99–100.
[50] J.E. Riviere and J.D. Brooks, *Prediction of dermal absorption from complex chemical mixtures: Incorporation of vehicle effects and interactions into a QSPR framework*, SAR QSAR Environ. Res. 18 (2007), pp. 31–44.
[51] T. Ghafourian, E.G. Samaras, J.D. Brooks, and J. Riviere, *Modelling the effect of mixture components on permeation through skin*, Int. J. Pharm. 398 (2010), pp. 28–32.
[52] J.E. Riviere and J.D. Brooks, *Predicting skin permeability from complex chemical mixtures: Dependency of quantitative structure permeation relationships on biology of skin models used*, Toxicol. Sci. 119 (2011), pp. 224–232.
[53] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, New York, 1995.

[54] Y. Sun, G.P. Moss, M. Prapodopolou, N. Davey, R. Adams, and M.B. Brown, *Predictions of skin penetration using Machine Learning methods, in ICDM2008 Proceedings of 8th IEEE International Conference on Data Mining*, F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan, X.D. Wu, eds., IEEE Computer Society, Los Alamitos, CA, USA, pp. 1049–1054.
[55] C.E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Boston, 2006.
[56] A. Geiger, *Gaussian Process software*; software available at: http://www.gaussianprocess.org/#code.
[57] *MatLab2012a.* MathWorks Ltd, Nitick, MA, USA, 2012; software available at http://jmlr.org/papers/v11/rasmussen10a.html.
[58] C.E. Rasmussen and H. Nickisch, *Gaussian Processes for Machine Learning (GPML) toolbox*, J. Mach. Learn. Res. 11 (2010), pp. 3011–3015.

Appendix

The supplemental material which is available via the multimedia link on the online article webpage contains full MatLab codes for running Gaussian Process analyses and specific scripts for particular covariance functions. It also contains a short guide on loading the scripts into MatLab and running them.

C.2 Conference paper

The Importance of Hyperparameters Selection within Small Datasets

Parivash Ashrafi[1], Yi Sun[1], Neil Davey[1] , Rod Adams[1] , Marc.B. Brown[2], Maria Prapopoulou[3],Gary Moss[4]

[1]*The Department of Computer Science, The University of Hertfordshire, UK {p.ashrafi2, y.2.sun, n.davey, r.g.adams}@herts.ac.uk*
[2]*MedPharm Ltd., Unit 3, Chancellor Court, 50, Occam Road, Guildford, Surrey, UK*
marc.brown@mdepharm.co.uk
[3]*Department of Pharmacy, King's College London, London, UK*
maria.2.prapopoulou@kcl.ac.uk
[4]*The School of Pharmacy, Keele University, United Kingdom*
g.p.j.moss@keele.ac.uk

***Abstract*—Gaussian Process is a Machine Learning technique that has been applied to the analysis of percutaneous absorption of chemicals through human skin. The normal, automatic method of setting the hyperparameters associated with Gaussian Processes may not be suitable for small datasets. In this paper we investigate whether a handcrafted search method of determining these hyperparameters is better for such datasets.**

***Index Terms*—Gaussian Process, QSAR, covariance function hyperparameters, likelihood maximisation**

I. Introduction

Recent developments in the analysis of percutaneous (through the skin) absorption of exogenous (external) chemicals, in industries including the pharmacy and cosmetics, the industrial use of bulk chemicals and skin toxicology, has become increasingly important [1]. To carry out research on in-vivo or in-vitro experiments, the excised skin of hairless animals (including rats, mice and pigs) or the skin of humans during surgery can be used. However, such studies have limitations due to the time taken and the expense in carrying them out [2]. It is important to understand that the collection of reliable data is difficult: human skin can be quite variable, affected by the site where it has come from and the individual form whom it is taken. Therefore the accuracy of the model is related to the variability of the data. Historically refining mathematical models used to predict percutaneous drug absorption has been thought of as a key factor in this field. Quantitative Structure-Activity Relationships (QSARs) models are used extensively for this purpose [3].

Studies have been performed [4], [5] which illustrate that using advanced Machine Learning techniques such as a Gaussian Process (GP) for the prediction of the percutaneous absorption of the chemicals, gave much better results than the QSAR model. Visualising the datasets reveals that the relationship between absorption and the physico-chemical features of the compound is highly non-linear [4], [6], [7]. The datasets used in this work are relatively small and this raises an interesting issue: does the normal *likelihood maximisation* based selection of hyperparameters still work well with these small datasets. It is known that with very small datasets, *marginal likelihood maximisation* does not work well ([8], sec. 5.4). In order to investigate this question we compare GP regressors in which the hyperparameters are set using *likelihood maximisation* based selection with regressors in which the hyperparameters are found by a manual search through the hyperparameter space. It does not appear to be much work has been done on regression with small datasets. Work done by Steyerberg et al [9] investigates logistic regression with small datasets. The regression estimation methods in their study included standard maximum likelihood, the use of linearlinkage factor, penalised maximum likelihood and the Lasso on univariable regression coefficient. The research indicates that using shrinkage methods in full models including predefined predictors and external information resulted in the best prediction performance in the small datasets. In addition, the work done by Martens et al [10] addresses multivariate calibration and uses Monte Carlo simulation on randomly selecting a small subset from a larger dataset repeatedly. In each of the 100 replications, the dataset was split into a calibration set (training set) and a test set. The result was compared with that obtained with full cross validation. The regression methodology applied in this study includes linear regression models using Partial Least Squares Regression (PLSR). The results indicate that using full cross validation outperforms both the use of an independent validation test set and independent verification test set.

This research focuses on examining the Gaussian Process (GP) and QSAR methods with 6 datasets. The datasets are collected from various sources such as journal papers. So that we have measured absorption rate through human skin under different experimental conditions such as temperature, body site and formulation of the chemicals applied. The sizes of the datasets are various and some datasets overlap. The effects of different sized datasets on the skin permeability predictions are examined. Each dataset includes measurements of both the chemical properties and the absorption rate of the variety of compounds. In fact, there are 5 measured properties such

978-1-4799-1959-8/15/$31.00

as log$P$, molecular weight (MW), solubility parameter (SP) and counts of the hydrogen bond donors (HD) and acceptors (HA) on each molecule and the absorption rate is measured by the permeability coefficient, K_p(cm/h). It should be noted that to experimentally estimate the permeability coefficient (K_p) of a single chemical can take around three days and entail some 6 to 24 experiments (depending on the purpose of the study - research, regulatory, and so on.). Therefore, the cost of experimentally generating only one estimation is as much as £25,000 - £30,000.

II. Problem domain

The difficulties in accurately measuring the percutaneous absorption of the chemicals is an important issue in the pharmaceutical and cosmetics industries. To address this issue, traditionally QSAR methods (a fixed linear regression) are derived to model the prediction of the penetration of various chemical compounds through human skin. The QSARs are widely used because they yield an easy to understand and easy to use algorithm on the mechanism of percutaneous absorption. The physicochemical properties used mainly in this form are: hydrophobicity (P) and molecular weight (MW); however other features may have a significant effect on the results [1]. P is the partition coefficient which is the ratio of the concentration in the lipid layer divided by the concentration in the water layer at equilibrium. The lipid layer is commonly represented by octanol and the water layer by pure water or sometimes a buffer so that the pH is the same and ionisation is controlled. P is also often called the octanol-water partition coefficient. Usually the log of P is reported because the values of P have a wide range between 10^{-7} to 10^{7}. For the percutaneous absorption K_p, the log is normally reported for the same reason [11].

III. Data description

The dataset employed in this study was collated from the literature and has been presented previously [4][12]. This dataset was subdivided into six different test sets, reflecting the diverse range of experimental protocols used in the studies from which this data was abstracted. We have 6 datasets with various sizes from 9 to 86. Table III.1 shows the number of data points in each dataset. There is also a degree of overlap between datasets. For instance, dataset E includes the majority of chemicals in datasets A and B. Similarly, dataset F includes most chemicals in dataset C. The experimental conditions under which the absorption were measured also varied. For example in some cases the temperature at which absorption measured was skin temperature but for some of them the temperature was set to a specific amount. 5 chemical features have been used in this study as mentioned in section I. It is often the case that pharmaceutical data is difficult and expensive to gather, as is the case here. In these circumstances obtaining good predictions from small data sets is important.

A. Data visualisation

Visualising the datasets can reveal more details about the nature of the data and show the distribution of the data. The datasets employed in this study, contain 5 features and it makes the visualisation difficult. In order to be able to demonstrate the data by its important features, Principal Component Analysis (PCA) was implemented. First, all the datasets gathered together and normalised into z-scores and then PCA applied to them. In this section the datasets E and F are visualised as they contain most of the data in other datasets. The first principal component ($PC1$) of datasets E and F against their logK_p are plotted in figure III.1. The first principal component accounts for 40.94% of the total variance. The first PC, $PC1$ is given by the equation $PC1 = 0.53HA + 0.52HD - 0.52\log P + 0.43SP - 0.04MW$. From this equation it could be seen that HA makes the largest contribution, but the most notable feature of $PC1$ is that MW makes very little contribution. This is surprising because in historical use of percutaneous absorption prediction, MW is one of the most important features they used (see section IV-A). Plot III.1 shows there is not a simple and linear relationship between the data features and logK_p in datasets E and F which makes the predictions difficult.

IV. Methodology

The methodology used in this study involves applying Single Layer Networks, QSAR and Gaussian Process with automatic and handcrafted hyperparameter selection to the 6 human skin datasets.

A. Quantitative Structure-Activity Relationship (QSAR) for skin permeability

The Quantitative Structure-Activity Relationship (QSAR) method used in this study to measure the human skin permeability coefficient logK_p, is the one defined by Potts and Guy [13] which is derived from the Flynn [14] dataset. Their model was the first major model for predicting percutaneous absorption which was based on human skin data. It is represented by the below equation:

$$\log K_p(cm/h) = 0.71\ \log P - 0.0061 MW - 2.74, \quad \text{(IV.1)}$$

where logK_p is the target or permeability coefficient of the skin, log$P$ is the partition coefficient (octanol-water partition coefficient) and MW is the molecular weight.

In the research performed by Potts and Guy [13] on the Flynn dataset [14], they concluded that the skin permeability is strongly a function of the partitioning between aqueous and non-aqueous layers (P or log$P$). It could be defined by hydrophobicity in terms of octanol-water partition coefficient. The permeability is also related to molecular weight. This results in defining the equation IV.1 for QSAR and we employ this equation to predict the QSAR skin permeability [1][6].

B. Single Layer Network

As mentioned earlier, visualising the datasets reveal that the relationship between the permeability coefficient of the compounds and their molecular features is highly non-linear.

Table III.1
NUMBER OF DATA-POINTS IN EACH DATASET

Performance/dataset	dataset A	dataset B	dataset C	dataset D	dataset E	dataset F
# Data points	9	25	21	57	51	86

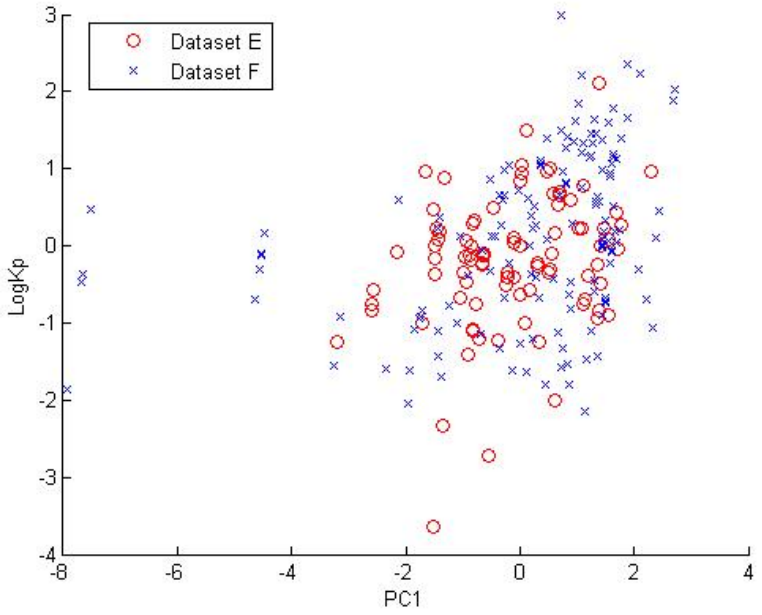


Figure III.1. datasets E, F visualization PC1-logK_p

To confirm this, a Single Layer Network (SLN) was used. An SLN can be considered as a simple generalised linear model. SLNs are known as a statistical technique for linear regression. These models have a linear combination of the input features, the coefficients are the parameters of the model . They also include an activation function tailored to the data being modelled. More information about SLNs can be found in [15].

C. Gaussian Processes

A Gaussian Process (GP) is an increasingly important technique for Machine Learning. A GP performs a non linear regression optimised from the training data. Suppose we have a dataset consisting of a number of input vectors each with a corresponding target value. The input vectors will be denoted as $\mathbf{x}$. The corresponding output values would be denoted by $\mathbf{y}$ and the new data point which we want to make the prediction of y_*, will be denoted as $\mathbf{x}_*$. One of the simplest way to do that is to get the weighted average of the $\mathbf{y}$ values in the training set, weighted by the similarity of the $\mathbf{x}_*$ to the $\mathbf{x}$ vectors in the train set. The idea is that the more similar the higher the weighting. In a GP, similarity is measured using a covariance or kernel function (a function takes two inputs and produces a single real value). Technically the prediction y_* is given by the following equation:

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

Here $\mathbf{k}_*$ is the vector of covariances between the test point and all the $\mathbf{x}$ values in the training set. The term $(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}$ which is completely independent of a new point, gives a linear transformation of this vector such that for the expected value of y_* is always an interpolation: so that at a known data-point its weight is one and the weights of all the other data points are zero. The transformed weights are multiplied by $\mathbf{y}$ values and the prediction achieved. One further significant feature of the GP regression is that it provides not only a prediction but also the variance associated with prediction (Equation IV.3), which can be interpreted as the reciprocal of a confidence value. Based on the Equation IV.3, if the weights are high then the variance should be low and vice-versa.

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

$\mathbf{K}$ denotes the covariance matrix of the training data, $\mathbf{I}$ denotes identity matrix with 1's on the diagonal and zeros elsewhere, σ_n^2 denotes the variance of an independent identically distributed Gaussian noise which means observations are noisy, and y denotes the vector of training targets. In equation IV.3 , $k(\mathbf{x}_*, \mathbf{x}_*)$ is the variance of y_* [8]. The software used for this study is the GP regression source code

which could be found in [16].

- **Covariance function**

A variety of kernels or covariance functions can be used in a GP model. In our initial research, the Matérn, Polynomial and Gaussian covariance functions are applied to the data. However, since the Matérn covariance function resulted in better prediction for our datasets, in this paper we report our results only on the Matérn covariance function. The Matérn covariance function used in this study is a product of linear and exponential function as it is defined by Equation IV.4. There are some other examples of Matérn covariance function that the polynomial order is quadratic, cubic or higher orders.

$$\mathrm{k}(r) = \left(1+\frac{\sqrt{3}r}{l}\right) exp\left(-\frac{\sqrt{3}r}{l}\right), \qquad \text{(IV.4)}$$

Here r $=|\mathbf{x}-\mathbf{x}_*|$, and l denotes the length scale, a positive hyperparameter which defines the scaling of the separation along the x axis. 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 $\mathbf{y}$ values may also be noisy and this is parametrised by the noise variance coefficient σ_n^2 that appears in Equation IV.2. Length scale (l), signal variance σ_f^2 and noise variance σ_n^2 are called hyperparameters of the model as they are not set directly from the training set. However, good values can be inferred from the data using either the *marginal likelihood maximisation* or methods of *cross-validation*. More information is given in [8]. The software used for this study is Gaussian Process modelling for non-linear regression [8], [16].

As mentioned earlier, it is possible that inferring the hyperparameters from the data could be problematic with a small dataset. To investigate this we compare the standard *marginal likelihood maximisation* with a manual search within hyperparameters space. When performing *marginal likelihood maximisation, Monte-Carlo* methods can be used or more commonly a *Gradient Descent* based methods. Two of these are discussed next.

- **Gradient descent for *marginal likelihood maximisation***

Essentially we are trying to find the minimum of a cost function and the first order method is simply to follow the maximum gradient downwards. Usually however second order methods are used such as *Conjugate Gradient* (CG) and and *Quasi-Newton* (QN). See [17], section 10.2 and [18], page 24 for further details on the CG and QN methods respectively.

D. Performance Measures

The *Mean Squared Error (MSE), Pearson Correlation coefficient (CORR) and Improvement Over Naïve (ION)* are used in this study [7]. Probably the simplest prediction (*Naïve*) always predicts the mean of the target value in the training set independently of the input. *ION* measures how much better a predictor is than the *Naïve* predictor. *ION* is given by the below equation:

$$ION = \frac{(MSE_{naïve} - MSE)}{MSE_{naïve}} \times 100\%. \qquad \text{(IV.5)}$$

The *CORR* measures the correlation between the targets and the corresponding prediction for it. In order to have a trustable model, values of both *ION* and *CORR* should be greater in the prediction on the test dataset while lower values of MSE show better predictions [7] .

V. Experiments and the results

To do all the experiments and before any models are trained, the data are normalised into z-scores, by subtracting the mean and dividing by the standard deviation (SD). The test set is normalised by the mean and SD of the training set. Due to the small number of data points in each dataset, we employed a special case of *cross-validation,* namely, *leave-one-out* to do all the experiments. This means that each time, one data point is used for testing and the rest are used as the training set and the test point target is predicted based on the training set features and targets. This process is repeated for each data point in the dataset and the prediction performances are obtained using the average value of all prediction performances.

A. Results using an SLN

The SLN is used with the *leave-one-out* method. After training the model, the prediction value for the test compound is computed [15]. The results are shown in Table V.1. Due to the reason that QSAR uses two features to estimate the skin permeability values, the same molecular features (log$P$ and MW) are also used in linear model and the results are shown also in Table V.1. The first thing to note is that in both datasets A and B the linear model works worse than the *Naïve* model and in particular on dataset A with 5 features the SLN works very poorly. This is almost certainly due to the fact that these datasets are very small with few vectors. It is also apparent that using the SLN on the other datasets brings little if any benefit.

B. GP with automatic hyperparameters optimisation

In this experiment a GP was applied to the datasets using the automatic hyperparameter selection method (using *marginal likelihood maximisation*). As mentioned earlier, to get the best automatic hyperparameters, two parameter optimisation methods: *Conjugate Gradient* (CG) and *Quasi-Newton* (QN) are used to maximise the marginal likelihood. Both CG and QN found similar values. The only difference was seen was in noise (σ_n^2) level which was around 0.02 using CG and 0.00 using QN method in dataset C; however, ION performances obtained from both techniques are the same. Table V.2 Reports the GP prediction performances using *Conjugate Gradient* method. In addition, the table shows the average and standard deviation of the three hyperparameters length-scale *(l)*, signal variance (σ_f^2) and noise (σ_n^2) in each dataset. These hyperparameters are non-negative and represented by their logarithms; thus, initializing them

to zero, corresponds to unit characteristic length-scale, unit signal standard deviation and unit noise values. To investigate the possibility of changing the results using different initial values, the hyperparameters then initialised to 0.5. The results did not show a significant change. Therefore, the results are reported for unit values of hyperparameters.

The first thing to notice is that QSAR consistently underperforms the *Naïve* model. This confirms our earlier work in this area [4], [5]. In fact in all cases QSAR has a negative *ION* indicating a worse than *Naïve* performance. From this and also Table V.1 it is clear that this data is not well suited to a linear predictor. The other thing to notice is that the GP in all but one case makes a minor improvement over the *Naïve* model but overall does not do very well. Surprisingly dataset E does not show a promising prediction performance, although it includes most chemicals in datasets A and B and it is larger in size than datasets A-D.

We also used 2 features in GP application to compare the results when 5 features are employed. Table V.3 demonstrates the results obtained from this investigation which similarly confirms QSAR consistently underperforms the *Naïve* model using 2 features. It also shows that GP works better using 5 features in all datasets.

C. GP with hand-crafted hyperparameters selection

The results obtained from the first experiment are not particularly encouraging. Our hypothesis is that the small size of the datasets may affect the automatic setting of the hyperparameters. To check the validity of this hypothesis, we undertook a systematic search to find effective hyperparameters for these datasets. 5 features are used for this experiment as it results in better prediction performances in the previous experiment. As mentioned earlier the l, σ_f^2 and σ_n^2 are the hyperparameters of the GP functions.

To do this requires a three dimensional grid search, and to keep this computationally tractable we initially searched over 5 values of σ_n^2: 10^{-3},10^{-2}, 10^{-1},10^{0} and 10^{1}. For the other two hyperparameters we searched over 100 values each due to the limitations of computational time. Both l and σ_f^2 take values between 10^{-3} to 10^{3}. The algorithm is as follows:

Part 1:

- For σ_n^2 in $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$, For l from 10^{-3} to 10^{3} step 10, For σ_f^2 from 10^{-3} to 10^{3} step 10
  - Train the GP using leave-one-out
  - Report the mean *ION*
- Report the best hyperparameters

Part 2: Refined search

- If the *ION* was not improved from the first experiment, undertake a finer search for better values of l and σ_f^2.

Part 3: Check the σ_n^2 best value

- Leaving l and σ_f^2 with the values found in part 2, For σ_n^2 from 10^{-1} to 10^{1} step 0.4
  - Train the GP using leave-one-out
  - Report the mean *ION*

Finally:

- Train the model with the best values from parts 1, 2 and 3 and report the final *ION* and *CORR*.

It should be noted that the *leave-one-out* technique was used with each dataset and the GP performances reported are the average of the performances obtained using this method. Table V.4 shows the best handcrafted hyperparameters obtained from this search together with the *ION* and *CORR* performances. Comparing these results to the *ION* performances obtained in the previous experiment (Tables V.2 and V.3) shows an increase in this method's performance in all of the datasets. The contour plot of changing the values l and σ_f^2 in dataset A is shown in Figure V.1. Noise is considered to be 0.1 in this plot as it results in the best *ION* performances with all datasets. It should be noted that for showing the hyperparameters their $\log_{10}$ values are used in this plot. It is clear from Table V.4 that using handcrafted hyperparameters increases the prediction performances in all datasets. The highest increase (78% in GP *ION*) was obtained in dataset A, which is the smallest dataset. The other datasets performances are increased by 7 %, 9%, 13%, 9%, 9% respectively in datasets B, C, D, E and F.

A notable point in V.4 is that hyperparameters l and σ_f^2 in dataset A (n=9) have large values compared to the others. To investigate more about this dataset, the best values of σ_n^2, l and σ_f^2 are considered and for one of the test points the similarity values (covariances) between test and training sets are calculated using Equation IV.4. Using these hyperparameters, similar values obtained for the kernel. This fairly flat kernel shows that using grid search for the small dataset seems to be an unstable solution. What can be be inferred from this section confirms the fact in [8], section 5.4 that the automatic hyperparameter selection using *marginal likelihood maximisation* seems not to work well on small datasets; however, using grid search and getting large values for hyperparameters in small datasets may not be a stable solution either, in the sense that predicting on novel compounds maybe of poor quality. However, for the larger datasets the hyperparameters appear to have more reasonable values, suggesting more robust models.

VI. Discussion and conclusion

We first of all can confirm that using automatic hyperparameters selection for extremely small datasets is problematic. Moreover, a grid search probably produced a brittle predictor which might generalised poorly. With such small dataset (9 items) it is probably impossible to produce a good model of the data, however, it should be taken into consideration that obtaining more data is difficult in this field and skin variability is also one of the main issues that simply cannot be dealt with.

With our other datasets, non of which are very large, automatic selection of hyperparameters also struggle in the sense that we obtained somewhat better models with the handcrafted grid search.

In general, there was a trend for the GP prediction performance to increase with increasing size of the data, although this was not always the case. Obtaining the data that we have used

Table V.1
SLN PERFORMANCES USING 2 AND 5 FEATURES (HIGHER *ION* BETTER)

Performance/dataset	dataset A	dataset B	dataset C	dataset D	dataset E	dataset F
ION* of *SLN (using 5 features)	-6.86	-0.46	0.2	0.01	0.04	0.06
ION* of *SLN (using 2 features)	-1.28	-0.1	-0.14	0.11	-0.03	0.12

Table V.2
USING 5 FEATURES, GP WITH AUTOMATIC HYPERPARAMETERS (USING CONJUGATE GRADIENT METHOD) MEAN AND STD USING MATÉRN COVARIANCE FUNCTION

Performance/dataset	dataset A	dataset B	dataset C	dataset D	dataset E	dataset F
Length scale *(l)* mean±std	0.45±0.14	0.4±0.04	0.85±0.1	1.02±0.06	0.42±0.05	0.56±0.20
Signal variance(σ_f^2) mean±std	0.85±0.09	0.90±0.12	1.14±0.04	0.91±0.05	0.96±0.03	1.23±0.00
Noise (σ_n^2) mean±std	0.20±0.15	0.66±0.19	0.02±0.01	0.61±0.04	0.23±0.04	0.01±0.00
***ION* of QSAR**	-7.43	-3.34	-0.30	-0.50	-2.76	-0.66
***ION* of GP**	**0.19**	**-0.03**	**0.38**	**0.33**	**0.00**	**0.37**
***CORR* of QSAR**	0.13	-0.18	0.30	0.39	-0.10	0.25
***CORR* of GP**	0.19	-0.21	0.56	0.56	0.09	0.61

Table V.3
USING 2 FEATURES, GP WITH AUTOMATIC HYPERPARAMETERS (USING CONJUGATE GRADIENT METHOD) MEAN AND STD USING MATÉRN COVARIANCE FUNCTION

Performance/dataset	dataset A	dataset B	dataset C	dataset D	dataset E	dataset F
Length scale *(l)* mean±std	0.22±0.08	0.16±0.07	0.59±0.09	1.31±0.13	0.29±0.21	0.75±0.06
Signal variance(σ_f^2) mean±std	0.79±0.09	0.14±0.31	0.85±0.09	0.94±0.03	0.34±0.37	1.05±0.02
Noise (σ_n^2) mean±std	0.33±0.21	1.07±0.16	0.73±0.17	0.83±0.11	0.72±0.36	0.84±0.02
***ION* of QSAR**	-7.43	-3.34	-0.30	-0.50	-2.76	-0.66
***ION* of GP**	**0.13**	**-0.08**	**0.13**	**0.3**	**-0.12**	**0.34**
***CORR* of QSAR**	0.13	-0.18	0.30	0.39	-0.10	0.25
***CORR* of GP**	0.13	-0.78	0.27	0.53	-0.44	0.57

Table V.4
USING 5 FEATURES, GP WITH FIX VALUES HANDCRAFTED HYPERPARAMETERS USING MATÉRN COVARIANCE FUNCTION

Performance/dataset	dataset A	dataset B	dataset C	dataset D	dataset E	dataset F
Length scale (l)	60	0.4	1	0.9	6.5	3.7
Signal variance(σ_f^2)	80	0.4	0.5	0.5	0.5	0.5
Noise (σ_n^2)	0.1	0.1	0.1	0.1	0.1	0.1
***ION* of QSAR**	-7.43	-3.34	-0.30	-0.50	-2.76	-0.66
***ION* of GP**	**0.81**	**0.04**	**0.42**	**0.38**	**0.10**	**0.41**
***CORR* of QSAR**	0.13	-0.18	0.30	0.39	-0.10	0.25
***CORR* of GP**	0.88	0.03	0.61	0.59	0.29	0.63

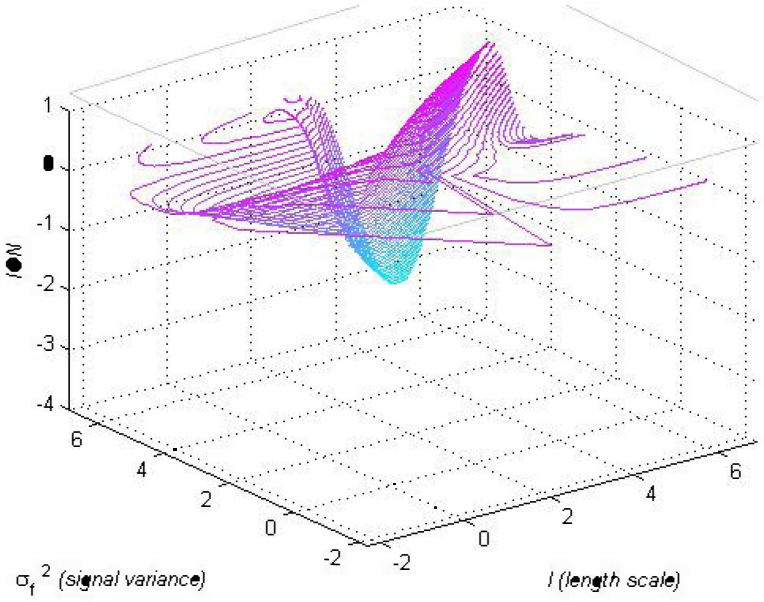


Figure V.1. *ION* values, changing the hyperparameters l and σ_f^2 ($\log_{10}$ values) in range 10^{-3} to 10^3 and fixed noise value (0.1) with dataset A

in this study is extremely difficult and time consuming. For example one researcher spent a year obtaining data for three compounds. We have already began work on using artificial skin as a means to speed up the process of generating data and also to ameliorate the need for animal testing [19], [6]. In future work it may be interesting to see whether taking into account the site where the experimental data is obtained could be useful additional data for our predictive model, as has been suggested in work on congeneric series [1]. Another point of note is that it seems all the models performed worse than or almost the same as the *Naïve* model on datasets B and E, especially with the automatic hyperparameter selection. It could be that the significant difference between these datasets with the other datasets is related to the way the measurements were made. A different type of instrumentations was used for these two datasets. In future work we intend to consider this additional experimental condition as another desciptor to see whether it can help to optimise the prediction performances.

REFERENCES

[1] GP Moss, JC Dearden, H Patel, MT Cronin, et al. Quantitative structure-permeability relationships (qsprs) for percutaneous absorption. *Toxicology in vitro: an international journal published in association with BIBRA*, 16(3):299, 2002.

[2] Eric J Lien and Hua Gaot. Qsar analysis of skin permeability of various drugs in man as compared to in vivo and in vitro studies in rodents. *Pharmaceutical research*, 12(4):583–587, 1995.

[3] Cronin M.T.D. and Schultz T.W. Pitfalls in qsar. *Journal of Molecular Structure: THEOCHEM*, 622(1):39–51, 2003.

[4] Gary P Moss, Yi Sun, Maria Prapopoulou, Neil Davey, Rod Adams, W John Pugh, and Marc B Brown. The application of gaussian processes in the prediction of percutaneous absorption. *Journal of Pharmacy and Pharmacology*, 61(9):1147–1153, 2009.

[5] Y Sun, MB Brown, M Prapopoulou, N Davey, RG Adams, and GP Moss. The application of stochastic machine learning methods in the prediction of skin penetration. *Applied Soft Computing*, 11(2):2367–2375, 2011.

[6] Gary P Moss, Yi Sun, Simon C Wilkinson, Neil Davey, Rod Adams, Gary P Martin, M Prapopopolou, and Marc B Brown. The application and limitations of mathematical modelling in the prediction of permeability across mammalian skin and polydimethylsiloxane membranes. *Journal of Pharmacy and Pharmacology*, 63(11):1411–1427, 2011.

[7] Yi Sun, Marc B Brown, Maria Prapopoulou, Rod Adams, Neil Davey, and Gary P Moss. The application of gaussian processes in the predictions of permeability across mammalian membranes. In *Artificial Neural Networks and Machine Learning–ICANN 2012*, pages 507–514. Springer, 2012.

[8] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[9] Ewout W Steyerberg, Marinus JC Eijkemans, Frank E Harrell, and J Dik F Habbema. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in medicine*, 19(8):1059–1079, 2000.

[10] Harald A Martens and Pierre Dardenne. Validation and verification of regression in small data sets. *Chemometrics and intelligent laboratory systems*, 44(1):99–121, 1998.

[11] Adrian Williams. *Transdermal and topical drug delivery*, volume 17. Pharmaceutical Press London, 2003.

[12] Lun Tak Lam, Yi Sun, Neil Davey, Rod Adams, Maria Prapopoulou, Marc B. Brown, and Gary P. Moss. The application of feature selection to the development of gaussian process models for percutaneous absorption. *Journal of Pharmacy and Pharmacology*, 62(6):738–749, 2010.

[13] Russell O Potts and Richard H Guy. Predicting skin permeability. *Pharmaceutical research*, 9(5):663–669, 1992.

[14] Gordon L Flynn. Physicochemical determinants of skin absorption, 1990.

[15] Ian Nabney. *NETLAB: algorithms for pattern recognition*. Springer, 2002.

[16] K. I. Williams Rasmussen. Gaussian processes for machine learning @ONLINE, 2006.

[17] Gene H Golub and Charles F van Van Loan. Matrix computations (johns hopkins studies in mathematical sciences). 1996.

[18] Jorge Nocedal and Stephen J Wright. Numerical optimization 2nd. 2006.

[19] Yi Sun, Gary P Moss, Maria Prapopoulou, Rod Adams, Marc B Brown, and Neil Davey. The application of gaussian processes in the prediction of percutaneous absorption for mammalian and synthetic membranes. In *ESANN*, 2010.

C.3 Conference poster abstracts

C.3.1 Poster 1

THE EFFECT OF QUALITY AND CONSISTENCY OF DATA ON THE DEVELOPMENT OF PREDICTIVE MACHINE LEARNING MODELS FOR PERCUTANEOUS ABSORPTION.

P. ASHRAFI[1], Y. SUN[1], N. DAVEY[1], R.G. ADAMS[1], M.B. BROWN[2], S.C WILKINSON[3], G.P. MOSS[4].

[1]School of Computer Science, University of Hertfordshire, Hatfield, UK; [2]Department of Pharmacy, University of Hertfordshire and MedPharm Ltd, Surrey, UK; [3]School of Clinical & Laboratory Sciences, Medical School, University of Newcastle-upon-Tyne, UK; [4]School of Pharmacy, Keele University, Keele, UK.

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

Introduction

It is well established that changes to the input of a system will significantly affect the output. For example, a review of the development of quantitative structure-permeability relationships (QSPRs) for percutaneous absorption[1] indicates that a wide range of different models have been developed in the last twenty years. These models have mostly been produced using the same, or very similar methods of analysis and yet by changing the dataset the output (the resulting QSPR algorithm) is also changed significantly – with this, the implications for predicting skin permeation and developing a mechanistic understanding of the skin permeability process are notably affected. It follows that the inference made by each model is therefore also different; this is seen best in the two landmark studies by Potts and Guy[2,3]. Machine Learning techniques have recently shown that they can outperform classical QSPR-based approaches, in terms of prediction accuracy and statistical robustness, to predict skin permeability[4,5,6,7]. These studies have also shown that the relationship between permeability and the physicochemical descriptors of a molecule are inherently non-linear. The datasets used in these studies are comparatively small and raises the concern that normal likelihood maximisation-based selection of hyperparameters may not necessarily work with such small datasets as are available to those researching percutaneous absorption. Thus, the current study has systematically examined this problem by dividing a large dataset discussed previously[4,5] into smaller subsets which reflect particular aspects of experimental design, removing possible bias in the data and potentially providing a clearer picture on whether the underlying trends in the data are linear or non-linear. In order to investigate this question GP regressors in which the hyperparameters are set using *likelihood maximisation*-based selection were compared with regressors in which the hyperparameters are found by a manual search through the hyperparameter space.

Methods

The dataset employed in this study was collated from the literature and has been presented previously[4,5]. This dataset was subdivided into six different test sets, reflecting the diverse range of experimental protocols used in the studies from which this data was abstracted. Subset A (n=9) – flow-through cells, abdominal skin 0.2-0.5mm thickness, 32°C skin surface temperature, 24h duration, aqueous donor vehicle, PBS receptor; subset B (n=25) – as subset A but contains all skin thicknesses listed; subset C (n=21) – as subset B but using static, Franz-type cells rather than flow-through cells; subset D (n=57) – as above but with increasingly non-aqueous components added to the donor vehicle and deploying flow-through or static diffusion cells; subset E (n=51) - as subset D but using flow-through cells only; subset F (n=86) – as subset E except deploying only static diffusion cells. There is also a degree of overlap between datasets. For instance, dataset E includes the majority of chemicals in datasets A and B. Similarly, dataset F includes most chemicals in dataset C. For each data point in each dataset five physicochemical descriptors were used (log P, molecular weight, solubility parameter and counts of the hydrogen bond donors and acceptors on each molecule). The Potts and Guy algorithm[2] was applied to each dataset. Further, Gaussian Process modelling was used, as reported previously, for non-linear regression[4,5,6,7]. As reported previously the selection of the covariance function is central to model quality. Thus, training points that are near to a test point could be helpful in better predicting the new test point. In this study the Matérn Polynomial and Squared Exponential covariance functions were applied to the data. The Matérn covariance function, which is the product of an exponential and a polynomial of order m (where m = 0, 1 or 2, with m = 1 yielding the best performance in initial studies), resulted in better prediction for the datasets examined and its outputs only are presented herein. The details of the covariance function are specified by its hyperparameters which in this experiment are the length scale (which defines the scaling of the separation along the x-axis), signal variance and noise variance. These parameters can be inferred or learned from the data, based on using either the *marginal likelihood maximisation* or methods of *cross-validation*[8]. One of the key concerns of this study is that inferring the hyperparameters from data could be problematic given the small sizes of some of the datasets. Therefore, a manual search within the space of the hyperparameters was also performed. Performance measures were used as reported previously[4,5,6,7] – correlation coefficient (CORR), improvement over the naïve model (ION) and mean squared error (MSE).

Results and Discussion

In the initial experiments, the GP was applied to the datasets using the automatic hyperparameter selection method (using *marginal likelihood maximisation*) for the covariance function. Due to the small number of data points in each dataset, leave-one-out cross validation was also employed. Thus, the average GP prediction performances are shown in Table 1. In addition, the table shows the average and standard deviation of the three hyperparameters in each dataset. It can be seen that the QSPR model significantly underperforms compared to the naïve model – in all cases the QSPR model has a negative value for ION, indicating a worse performance than the naïve model. This suggests that these datasets are not well suited to linear predictors. Similarly, GP models in all but one case make a minor improvement over the naïve model but, overall do not do perform significantly or consistently better. Given the poor outcome of the initial experiment the hypothesis subsequently proposed was that the small size of the datasets may affect the automatic setting of the hyperparameters. Thus, a systematic search to find effective hyperparameters for these data was undertaken. This required a three-dimensional grid search to keep it computationally tractable; therefore four orders of magnitude of signal variance and over 100 values each of the other hyperparameters were used. The GP model was trained using a leave-one-out methodology and the ION and best hyperparameters were reported; if no improvement in the ION was observed the search was refined by repeating it in a more detailed fashion. Finally, the model with the best outcome for these parameters and correlation was proposed. Table 2 shows the best handcrafted hyperparameters obtained from this search and the ION and CORR performances. Comparison with the results in Table 1 shows a significant improvement in the quality of the model's performance in most of the datasets. Figure 1 shows how the changes in the values of the hyperparameters affect the ION – it should be noted that the figure contains a log-scale for the signal variance). It can be clearly seen from Table 2 that using handcrafted hyperparameters increases the prediction performance in all datasets. In one case an increase of 78% over the naïve model is observed for dataset A which is, incidentally, the smallest dataset.

Therefore, it is clear that the quality of the model produced is affected significanty by the dataset size. This has significant implications for the use of single-dataset analyses of skin absorption. It is also clear that the use of smaller subsets – although logical in terms of removing any variance introduced by different experimental protocols – has limitations, given the overall quality (as determined, for example, by the CORR values obtained). This is significant in the wider understanding of error and variance associated with quantitative models of skin absorption. However, despite such limitations the application of Machine Learning methods has shown that model quality can be substantially improved and that the use of – in this case – handcrafted hyperparameters can significantly improve the predictive ability and statistical quality of the models.

Table 1. Automatic hyperparameters (mean and standard deviation): application of GP with *Matérn* Covariance function (m = 1) and automatic hyperparameter generation.

Performance/dataset	dataset A	dataset B	dataset C	dataset D	dataset E	dataset F
Length scale (l) mean±std	0.45±0.14	0.4±0.04	0.85±0.1	1.02±0.06	0.42±0.05	0.82±0.05
Signal variance(σ_f^2) mean±std	0.72±0.14	0.81±0.21	1.3±0.09	0.91±0.05	0.93±0.04	1.29±0.04
Noise (σ_n^2) mean±std	0.006±0.004	0.006±0.001	0.000±0.000	0.009±0.000	0.008±0.000	0.009±0.001
ION of QSAR	-7.43	-3.34	-0.30	-0.50	-2.76	-0.66
ION of GP	**0.18**	**-0.03**	**0.38**	**0.33**	**0.00**	**0.37**
CORR of QSAR	0.13	-0.18	0.30	0.39	-0.10	0.25
CORR of GP	0.18	-0.21	0.56	0.56	0.09	0.61

Table 2. Handcrafted hyperparameters fixed values: application of GP with *Matérn* Covariance function (m = 1) and handcrafted hyperparameter selection.

Performance/dataset	dataset A	dataset B	dataset C	dataset D	dataset E	dataset F
Length scale (l)	60	0.4	1	0.9	6.5	3.7
Signal variance(σ_f^2)	110	0.4	0.5	0.5	0.5	0.5
Noise (σ_n^2)	0.1	0.1	0.1	0.1	0.1	0.1
ION of QSAR	-7.43	-3.34	-0.30	-0.50	-2.76	-0.66
ION of GP	**0.81**	**0.04**	**0.42**	**0.38**	**0.10**	**0.41**
CORR of QSAR	0.13	-0.18	0.30	0.39	-0.10	0.25
CORR of GP	0.88	0.03	0.61	0.59	0.29	0.63

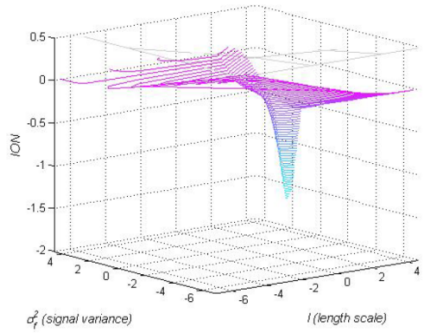


Figure 1. Effect of changing hyperparameters on the ION values. Hyperparameters used are length-scale (l), and signal variance ($\sigma^2 f$) in dataset E with fixed noise value (0.1) over the range 10^{-3} – 10^3.

References

1. Moss, G.P., Dearden, J.C., Patel, H, Cronin, M.T. D. Quantitative Structure-Permeability Relationships (QSPRs) for percutaneous absorption. *Tox. In Vitro*, **16**, 299 – 317, 2002.
2. Potts, R.O., Guy, R.H. Predicting skin permeability. *Pharm. Res.*, **9**, 663–669, 1992.
3. Potts, R.O., Guy, R.H. A predictive algorithm for skin permeability: the effects of molecular size and hydrogen bond activity. *Pharm. Res.*, 12, 1628–1633, 1995.
4. Moss, G.P., Sun, Y., Davey, N., Adams, R., Pugh, W.J., Brown, M.B. The application of Gaussian Processes to the prediction of percutaneous absorption. *J. Pharm. Pharmacol.*, **61**, 1147 – 1153, 2009.
5. Lam, L.T., Sun, Y., Davey, N., Adams, R., Prapopoulou, M., Brown, M. B., Moss, G.P. The application of feature selection to the development of Gaussian Process models for percutaneous absorption. *J. Pharm. Pharmacol.*, **62**, 738 – 749, 2010.
6. Moss, G.P., Sun, Y., Wilkinson, S.C., Davey, N., Adams, R., Martin, G.P., Prapopoulou, M., Brown, M.B. (2011). The application and limitations of mathematical modelling in the prediction of permeability across mammalian skin and polydimethylsiloxane membranes. *Journal of Pharmacy and Pharmacology*, **63**, 1411 – 1427.
7. Sun, Y., Moss, G.P., Davey, N., Adams, R., Brown, M.B. (2011) The application of stochastic machine learning methods in the prediction of skin penetration. *Applied Soft Computing*, **11**, 2367 – 2375.
8. Rasmussen, C.E., Williams, C.K.I. *Gaussian processes for machine learning*. Cambridge: MIT Press, 2006.

C.3.2 Poster 2

INVESTIGATION OF INCONSISTENCY IN A SKIN PERMEABILITY DATASET USING THE MONTE CARLO METHOD

P ASHRAFI[1], Y SUN[1], RG ADAMS[1], N DAVEY[1], SC WILKINSON[2], GP MOSS[3]

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

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

Introduction

Skin permeability datasets are often criticised for exhibiting a significant degree of variation in their source data. This is generally due to the nature of data collation which, in the case of the skin permeability datasets used, has been the result of accumulated studies from a range of different researchers and laboratories. Inherent in such a strategy for generating databases of sufficient size to be viable is the inter-laboratory variation which adds to the biological variation associated with such studies[1]. Such inherent variation is often cited as a major limitation in models of percutaneous absorption and a reason for their broader lack of application by those investigating skin absorption[2,3,4]. The aim of a Monte Carlo simulation is to generate values for uncertain values in a model through random sampling. Thus, the Monte Carlo techniques apply random sampling methods to complex datasets which often exhibit inconsistent target values, such as the simulation of risk or, in this case, the analysis of inherently variable biological data. It is therefore the aim of this study to use the Monte Carlo method to examine a skin permeability data set, with the aim of determining whether this technique can improve the model quality and address issues of data variability.

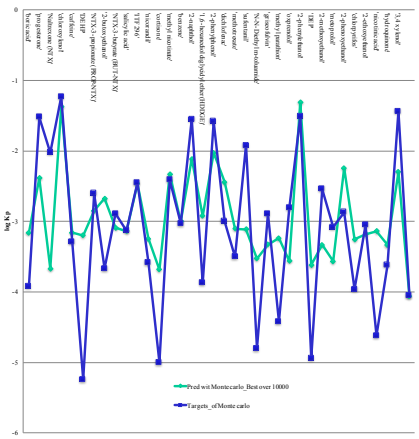
Methods

The data set used in this study has been reported previously[4,5]. Briefly, this is a data set of *in* vitro human skin permeability studies whose core is the Flynn data set as modified by Moss and Cronin[6,7] and added to by Prapopoulou[4] (n=642). Each experiment involves the random selection of one of the target values for the same compounds; this process was repeated 10,000 times and therefore 10,000 data sets were generated. For each data set the performance was assessed by comparing the quality of predictive performance, where the data was randomly split into training (two-thirds) and test (one-third) data sets. Performance measures were the mean squared error (MSE; estimates vs. laboratory results), improvement over the naïve model (ION) and the correlation coefficient (r, or CORR). Monte Carlo experiments were benchmarked against normal methods of analysis[5], including QSARs, where the mean values of reported experimental results were used to construct a model.

Results and Discussion

By applying the Monte Carlo method the ION improved from 0.34 to 0.41, and the correlation coefficient improved from 0.55 to 0.64, in cases were the Monte Carlo method was applied to a data set constructed from mean permeability values for each relevant chemical; these are the values obtained from the best Monte Carlo simulation of all 10,000 variants that were run. For clarity of comparison (estimates vs. experimental values) Figures 1 and 2 show a summary of the total results – they highlight the estimates for 35 chemicals taken from the "best" Monte Carlo data set. From these results it is apparent that, in most cases, the application of the Monte Carlo method works better than using a data set containing mean values of chemicals in that it produces better estimates. This analysis also yields an "optimal" data set; starting with the complete data set repetitive data points were removed and the best target values of the statistical metrics were produced. This results in a smaller data set which yielded the best-performing model. This data set is discussed below.

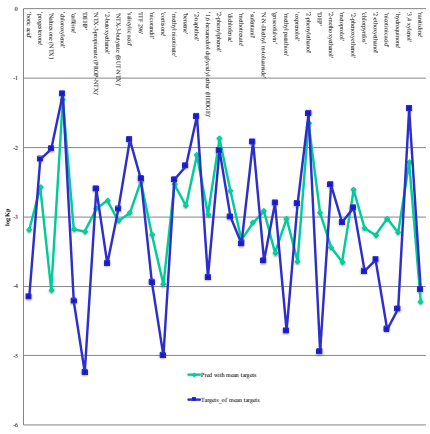
Figure 1. Comparison of estimates with targets (Monte Carlo method).



Using the optimal dataset a QSAR equation was developed using multiple linear regression analysis (SPSS v21). This yielded an algorithm of poor statistical quality: log k_p = -0.003MW – 2.681 (r^2 = 0.2). This suggests the inherent lack of linearity in this dataset, which is in agreement with previous findings[3,5].

In examining the data returned as producing the most effective model it is clear that, in terms of experimental conditions, few specific trends are identified. The types of membranes used are evenly distributed between full thickness, dermatomed and heat-separated epidermal sheets, with a small number (6 out of 129) using different methods of membrane pre-treatment or preparation, particularly the use of isolated *stratum corneum*. While diffusion experiments were performed using mostly abdominal skin the use of back, breast, dorsal and leg skin was also common. No anatomical site was listed in fourteen of the 129 experiments. Experiments were divided between static, Franz-type diffusion cells (80) and flow-through cells (49). While the majority of donor phase vehicles were aqueous in nature the optimal data set did feature a significant subset of experiments which were conducted using non-aqueous solvents, mostly various alcohols, isopropyl myristate or acetone. In terms of qualitative examination of the data no clear trends were apparent between estimates and experimental data when investigated based on the degree of error (estimates vs. experiments) on a chemical-by-chemical basis.

Figure 2. Comparison of estimates with targets (conventional GP methods).



It may therefore be suggested that, given the overall variation in the dataset, the effects of specific conditions, such as receptor compartment composition, vehicle, skin thickness and diffusion cell design, on the outcome of experiments, are either not significant or are difficult to detect due to the inherent biological variation associated with the data.

References

1. Chilcott RP, Barai N, Beezer AE, Brain SL, Brown MB, Bunge AL, Burgess SE, Cross S, Dalton CH, Dias M, Farinha A, Finnin BC, Gallagher SJ, Green DM, Gunt H, Gwyther RL, Heard CM, Jarvis CA, Kamiyama F, Kasting GB, Ley EE, Lim ST, McNaughton GS, Morris A, Nazemi MH, Pellett MA, Du Plessis J, Quan YS, Raghavan SL, Roberts M, Romonchuk W, Roper CS, Schenk D, Simonsen L, Simpson A, Traversa BD, Trottet L, Watkinson A, Wilkinson SC, Williams FM, Yamamoto A, Hadgraft J. Inter- and intra-laboratory variation of in vitro diffusion cell measurements: An international multicenter study using quasi-standardised methods and materials. *J. Pharm. Sci*, **94**, 632 – 638, 2005.
2. Moss GP, Dearden JC, Patel H, Cronin MTD. Quantitative Structure-Permeability Relationships (QSPRs) for percutaneous absorption. *Tox. In Vitro,* **16,** 299 – 317, 2002.
3. Moss GP, Gullick DR, Wilkinson SC. *Predictive Methods In Percutaneous Absorption.* Springer, Berlin, 2015, pp159-176.
4. Prapopoulou, M. *The development of a mathematical / computational model to predict drug absorption across the skin.* PhD Thesis, King's College London, 2012.
5. 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.*, **61**, 1147 – 1153, 2009.
6. 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.
7. Moss GP, Cronin MTD. Quantitative structure-permeability relationships for percutaneous absorption: re-analysis of steroid data. *Int. J. Pharm*. **238**, 105 – 109, 2002.

C.3.3 Poster 3

THE EFFECT OF EXPERIMENTAL CONDITIONS ON THE DEVELOPMENT OF QUANTITATIVE MODELS OF SKIN PERMEATION

P ASHRAFI[1], Y SUN[1], N DAVEY[1], RG ADAMS[1], GP MOSS[2] , SC WILKINSON[3],

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

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

Introduction

There are numerous studies which describe the many and varied mathematical models of percutaneous absorption developed from experimental data[1,2]. These studies are reviewed extensively and they describe a series of models, mostly quantitative structure-permeability relationships (QSPRs), which simply collate existing literature data and then develop relationships between molecular physicochemical descriptors and a biological process (i.e. skin permeability)[1,2,3]. However, in recent years, models are being developed with more consideration given to the input data used, utilising data from as few sources as possible and considering in the construction of models a greater range of input factors[4,5,6]. This is partially due to new methods which may allow more accurate estimates to be produced but also due to an increase in data availability which allows data sets to be constructed more appropriately. More recently, Machine Learning techniques have recently shown that they can outperform classical QSPR-based approaches, in terms of prediction accuracy and statistical robustness, to predict skin permeability[5,6,7,8]. These studies have also shown that the relationship between permeability and the physicochemical descriptors of a molecule are inherently non-linear. Thus, the aim of this study is to determine the effect of diffusion cell type (i.e. 'static', Franz-type cells and 'flow-through' Bronaugh-type cells) on the construction of models of skin permeability. This work is complimentary to our other poster (on Monte Carlo analysis of skin permeability) and should be considered in the context of that study.

Methods

The Gaussian Process Regression methods reported previously were employed[5,6,7,8]. All the human skin data was collated into two categories; experiments conducted with static, Franz-type diffusion cells or with flow-through cells. Five molecular descriptors were used: log P, MW, HA, HD and melting point. The target values used in the model development were the measured permeability coefficient values taken from the source literature. The data set used, after removing inconsistent data, has 143 members (85 static cell experiments and 53 flow-through cell experiments). Analyses on separate flow-through and static cell experiments were carried out, followed by experiments examining mixtures of both, using the leave-one-out method. To analyse the data 10 datasets were constructed. Each of them, in the second type of experiments, includes the mixed static and flow-through cell data[4,5] randomly selected from 143 members in order to remove possible bias in analysis. Two-thirds of the data was used as the training set. Each pair – of training set and test set – was checked to ensure that there was no common data between them (hence, repetitions between training and test sets were removed). The ten training sets were trained separately and the predictions obtained for the ten test. Model quality was assessed as described before[4,5] by using the ION (improvement of the naïve model) and CORR (correlation coefficient).

Results and Discussion

GP prediction performances were obtained for flow-through and static cell experimental data in order to assess the models. The results are shown in Table 1, where it can be seen that the performance of the flow-through model is poor whereas the model using only static cell data was substantially better (ION = 0.04_{FT} vs. 0.42_{STAT}; correlation coefficient = 0.20_{FT} vs. 0.66_{STAT}).

The data from flow-through and static cell experiments was then collated and analysed together as a single data set. The results are shown in Table 2 and the average performance indicates that the best predictive models are always obtained when data from static diffusion cell experiments is used (i.e. Table 1: 42% and 66% for ION and CORR, respectively) compared to models constructed from flow-through cell experiments (i.e. Table 2: -7% and 19% for ION and CORR, respectively). These results were obtained regardless of whether data from static or flow-through cells, or mixtures of both, were used to train models. Further, training models based on flow-through cell data only, and predicting the permeability of 'unseen' test data, resulted in poor models with a performance of 4% and 20% for ION and CORR, respectively (Table 1). Using data from Franz-type cell experiments to train a predictive model for flow-through diffusion cells, and vice versa (Table 3) resulted in poorly predictive models, suggesting a lack of comparability between both types of cells.

It is apparent that the quality of the model is directly affected by the nature of the input data, and that the inclusion of data from flow-through experiments may reduce overall model quality and predictive power, while models based solely on such data offer poor predictions of skin permeability. This is in significant contrast to models developed from static diffusion cell experiments, which resulted in comparatively highly predictive models. It may therefore be suggested that, in order to optimise the model quality, data from only static, Franz-type, experiments should be used to construct the model. Interestingly however, this should be taken in the context of our Monte Carlo simulation study[9] which demonstrated that models constructed from mixtures of diffusion cell experiments were not necessarily reduced in their predictive qualities by combining such data together – this may indicate that the influence of variable data does, in some cases, make it difficult to fully discern trends such as those observed in this study.

Table 1. Prediction performances for static and flow-though diffusion cells used to assess the models.

	Flow-through cell		Static cell	
	Mean	STD	Mean	STD
MSE_{GP}	0.84	0.12	0.98	0.13
ION_{GP}	0.04	0.04	0.42	0.09
$MSE_{NaiveGP}$	0.87	0.13	1.68	0.09
$CORR_{GP}$	0.20	0.13	0.66	0.06

Table 2. Prediction performances for a single dataset (with collated data from flow-through and static diffusion cell experiments) used to assess the models.

	Flow-through cell		Static cell	
	Mean	STD	Mean	STD
MSE_{GP}	0.93	0.23	0.96	0.09
ION_{GP}	-0.07	0.09	0.43	0.05
$MSE_{NaiveGP}$	0.86	0.14	1.70	0.09
$CORR_{GP}$	0.19	0.16	0.67	0.05

Table 3. Performance measures for training flow-through or static cell models with data from the other experiments.

	Static cell data to train a predictive model for a flow-through cell model		Flow-through cell data to train a predictive model for a static cell model	
	Mean	STD	Mean	STD
MSE_{GP}	1.17	0.32	1.64	0.06
ION_{GP}	-0.35	0.22	0.05	0.04
$MSE_{NaiveGP}$	0.07	0.24	0.19	0.16
$CORR_{GP}$	0.86	0.14	1.73	0.10

References

1. Moss GP, Dearden JC, Patel H, Cronin MTD. Quantitative Structure-Permeability Relationships (QSPRs) for percutaneous absorption. *Tox In Vitro*, **16**, 299-317, 2002.
2. Moss GP, Gullick DR, Wilkinson SC. Predictive *Methods In Percutaneous Absorption*. Springer, Berlin, 2015.
3. Moss GP, Wilkinson SC, Sun Y. Mathematical modelling of percutaneous absorption. *Curr. Op. Coll. Interface Sci.*, **17**, 166–172, 2012.
4. Magnusson BM, Anissimov YG, Cross SE, Roberts MS. Molecular size as the main determinant of solute maximum flux across the skin. *J. Invest. Dermatol*., **122**, 993-999, 2004.
5. 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*, **61**, 1147-1153, 2009.
6. Lam LT, Sun Y, Davey N, Adams R, Prapopoulou M, Brown MB, Moss GP. The application of feature selection to the development of Gaussian Process models for percutaneous absorption. *J. Pharm. Pharmacol*. **62**, 738-749, 2010.
7. Moss GP, Sun Y, Wilkinson SC, Davey N, Adams R, Martin GP, Prapopoulou M., Brown MB. The application and limitations of mathematical modelling in the prediction of permeability across mammalian skin and polydimethylsiloxane membranes. *J. Pharm. Pharmacol*.,**63**, 1411 – 1427, 2011.
8. Sun Y, Moss GP, Davey N, Adams R, Brown MB. The application of stochastic machine learning methods in the prediction of skin penetration. *App. Soft Comp.*, **11**, 2367 – 2375, 2001.
9. Ashrafi P, Sun Y, Adams RG, Davey N, Wilkinson SC, Moss GP. Investigation of inconsistency in a skin permeability dataset using the Monte Carlo method. Poster presented at PPP2016.

C.3.4 Poster 4

ASSESSMENT OF CHEMICAL ENHANCERS OF TRANSDERMAL DRUG DELIERY BY SUPPORT VECTOR REGRESSION

A SHAH[1], P ASHRAFI[2], Y SUN[2], RG ADAMS[2], N DAVEY[2], SC WILKINSON[3], GP MOSS[4]

[1]Department of Software Engineering & 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, University of Newcastle-upon-Tyne, UK

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

Introduction

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[1]. 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[1]. 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[2]. 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 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[3]. 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.

Machine Learning methods have shown considerable promise in providing accurate estimates of skin permeability[4,5]. The support vector regression (SVR) method has not previously been applied to a pharmaceutically relevant endpoint. The support vector classification (SVC) method previously reported does offer significant improvements in model quality compared with discriminant analysis[6]. SVC is limited as it is a classification method and was able to provide class membership only, as defined by the degree of enhancement benefit (the ER ratio), rather than estimates of performance Improvement[7]. 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' to optimise predictive power and subsequent significance of the pharmaceutically relevant endpoint. Thus, the aims of the current study were 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.

Methods

The dataset used in this study has been published previously[7] and consists of 71 chemical used to enhance the permeation of hydrocortisone across mouse skin. Data visualisation was conducted via Principal Component Analysis. Development of models was accomplished by using a range of simple regression models, Gaussian Process Regression (GPR) and Support Vector Regression (SVR)[7,8] – both the ε-SVR and ν-SVR methods – with up to five physicochemical descriptors[6] and a range of kernels, including the linear and Radial Basis Function (RBF) kernels. Errors in prediction for the GPR and SVR methods were compared with a two-tailed binomial test.

Results and Discussion

Data description

A quantile – quantile plot of the dataset used (Figure 1) suggests that the data has two defined subsets ('good' or 'poor' enhancers, defined by MW) with 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 quantile-quantile plot suggested that 'good' and 'poor' enhancers could not be discriminated from each other based on hydrogen bonding, even though PC2 is strongly correlated to hydrogen bonding. Removal of hydrogen bonding from the model reduces model quality only slightly – this may be a possible synergistic effect similar to that associated with the use of models of skin permeability featuring, for example, only log P or a combination of log P and MW. Principal component analysis indicates that the first principal component (PC1 = 0.05HB – 0.47CC – 0.49 MW – 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% of the variance.

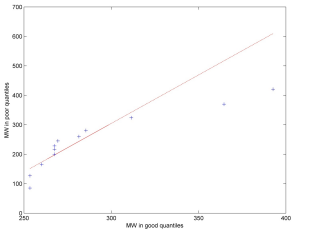


Figure 1. A quantile-quantile plot of MW for those enhancers classified as "good" and "poor".

Different experimental conditions

Comparison of models with **three descriptors** (carbon chain length, CC, molecular weight, MW and the number of hydrogen bonding groups per molecule, HB) indicate no significant differences in statistical measured used to assess model quality (MSE or r^2) for GP or different SVR methods (ε-SVR and ν-SVR) . **Five descriptor models** (MW, CC, HB and adding logP, logS) indicate that the ν–SVR method is significantly better than ε-SVR method. While models are better differences are small, this is possibly due to variance in the source (experimental) data. Fitted linear regression was shown to be unsuitable for analysis as there are 11 chemicals where log ER > 1.0, and the estimates provided are all < 1.0. This is possibly due to the nature of the data, where 59 (out of 71) chemicals have ER < 1.0. Application of the RBF and *Matern* kernels (with 3 and 5 molecular features) indicates that the GPR method gives the best results. The ε-SVR method has the same performance on both linear and RBF kernels with three molecular features. The ν-SVR method gives a slight improvement on the r^2 measurement using the RBF kernel, while the MSE measurement remains the same for both kernels with three features. The ν–SVR method resulted in the best model when five descriptors were used – MSE and r^2 improved by 23% and 7%, respectively, compared to ε-SVR (Figure 2). Errors in prediction (MSE) between GP and SVR methods were not significantly different (Spearman test; p=0.12, suggesting a degree of correlation between SVR and GP methods; Wilcoxon signed rank; P=0.34 suggesting that the GP and SVM predictions may be from a distribution with the same median).

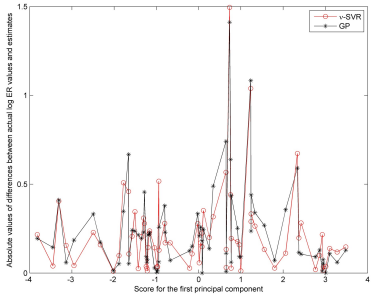


Figure 2. Plot of the relative values of differences of actual logER values and estimates against the first principal component.

Discussion

Overall, the SVR methods produce models of similar statistical quality to the GP methods. The ν–SVR method is better in estimating ER for 40 out of 71 chemicals. However, it is important to note that GP and SVR methods perform better at certain points of the 'chemical space' of the data set. For example, GP has a better performance where PC1 ≥ 3 and GPR gives a more reliable estimate on 8 of 12 chemicals that have a PC1 ≥ 2.3. In addition, GP has a better performance at PC1 [1.17, 0.94] on all six chemicals (Figures 3 and 4). This indicates that, while it is important for model quality to consider the physicochemical descriptors used, it is also important to consider that different models may produce different predictions and that the concept of using a single, 'global' model may not always be appropriate. Thus, combining classification and regression methods to provide model optimisation at different parts of the chemical space may provide a significant new approach to improving model quality and relevance.

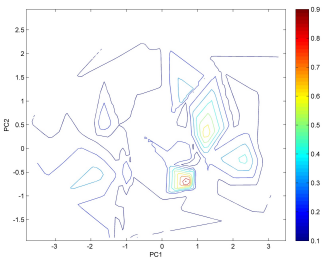


Figure 3. Contour plot of the principal components for the Gaussian Process Regression (GPR) method.

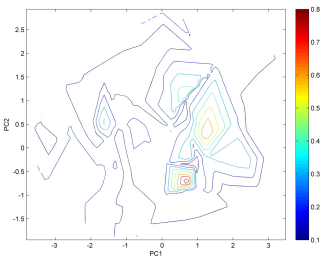


Figure 4. Contour plot of the principal components for the Support Vector Regression (SVR) method.

References

1. Moss GP, Dearden JC, Patel H, Cronin MTD. Quantitative Structure-Permeability Relationships (QSPRs) for percutaneous absorption. *Tox. In Vitro,* **16,** 299 – 317, 2002.
2. Riviere JE, Brooks JD. Prediction of dermal absorption from complex chemical mixtures: incorporation of vehicle effects and interactions into a QSAR framework. *SAR QSAR Environ. Res.*, **18**, 31 – 44, 2007.
3. Ghafourian T, Samaras EG, Brooks JD, Riviere JE. Validated models for predicting skin penetration from different vehicles. *Eur. J. Pharm. Sci.*, **41**, 612 – 616, 2010.
4. 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.*, **61**, 1147 – 1153, 2009.
5. Lam LT, Sun Y, Davey N, Adams R, Prapopoulou M, Brown MB, Moss GP. The application of feature selection to the development of Gaussian Process models for percutaneous absorption. *J. Pharm. Pharmacol.*, **62**, 738 – 749, 2010.
6. Pugh WJ, Wong R, Falson F, Michniak BB, Moss GP. Discriminant analysis as a tool to identify compounds with potential as transdermal enhancers. *J. Pharm. Pharmacol.*, **57**, 1389 – 1396, 2005.
7. 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.*, 45, 116 – 127, 2012.
8. Shah AJ, Sun Y, Adams RG, Davey N, Wilkinson SC, Moss GP. Support vector regression to estimate the permeability enhancement of potential transdermal enhancers. *J. Pharm. Pharmacol.*, **68**, 170 – 184, 2016.

Bibliography

M. H. Abraham, H. S. Chadha, and R. C. Mitchell. The factors that influence skin penetration of solutes*. *Journal of pharmacy and pharmacology*, 47(1):8–16, 1995.

M. H. Abraham, F. Martins, and R. C. Mitchell. Algorithms for skin permeability using hydrogen bond descriptors: the problem of steroids*. *Journal of pharmacy and pharmacology*, 49(9):858–865, 1997.

M. H. Abraham, H. S. Chadha, F. Martins, R. C. Mitchell, M. W. Bradbury, and J. A. Gratton. Hydrogen bonding part 46: A review of the correlation and prediction of transport properties by an lfer method: Physicochemical properties, brain penetration and skin permeability. *Pesticide Science*, 55(1):78–88, 1999.

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions: With Formulars, Graphs, and Mathematical Tables*, volume 55. DoverPublications. com, 1964.

H. L. Anderson. Metropolis, monte carlo, and the maniac. *Los Alamos Science*, 14:96–108, 1986.

P. Ashrafi, G. Moss, S. Wilkinson, N. Davey, and Y. Sun. The application of machine learning to the modelling of percutaneous absorption: an overview and guide. *SAR and QSAR in Environmental Research*, 26(3):181–204, 2015.

R. Bardenet and B. Kégl. Surrogating the surrogate: accelerating gaussian-process-based global optimization with a mixture cross-entropy algorithm. In *27th International Conference on Machine Learning (ICML 2010)*, pages 55–62. Omnipress, 2010.

M. Barratt. Quantitative structure-activity relationships for skin permeability. *Toxicology in Vitro*, 9(1):27–37, 1995.

J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pages 2546–2554, 2011.

R. L. Bronaugh and H. I. Maibach. *Percutaneous absorption: mechanisms–methodology–drug delivery*, volume 8. Marcel Dekker Inc, 1989.

S. L. Brown and J. E. Rossi. A simple method for estimating dermal absorption of chemicals in water. *Chemosphere*, 19(12), 1989.

D. Büche, N. N. Schraudolph, and P. Koumoutsakos. Accelerating evolutionary algorithms with gaussian process fitness function models. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 35(2):183–194, 2005.

F. R. Burden. Quantitative structure-activity relationship studies using gaussian processes. *Journal of chemical information and computer sciences*, 41(3):830–835, 2001.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

O. Chapelle and V. Vapnik. Model selection for support vector machines. In *NIPS*, pages 230–236, 1999.

L.-j. Chen, G.-p. Lian, and L.-j. Han. Prediction of human skin permeability using artificial neural network (ann) modeling. *Acta Pharmacologica Sinica*, 28(4):591, 2007.

V. Cherkassky and Y. Ma. Selecting of the loss function for robust linear regression. *Neural computation*, 2002.

V. Cherkassky and F. M. Mulier. *Learning from data: concepts, theory, and methods*. John Wiley & Sons, 2007.

R. Chilcott, C. Dalton, I. Hill, C. Davison, K. Blohm, E. Clarkson, and M. Hamilton. In vivo skin absorption and distribution of the nerve agent vx (o–ethyl–s–[2 (diisopropylamino) ethyl] methylphosphonothioate) in the domestic white pig. *Human & experimental toxicology*, 24(7):347–352, 2005.

R. L. Cleek and A. L. Bunge. A new method for estimating dermal absorption from chemical exposure. 1. general approach. *Pharmaceutical research*, 10(4):497–506, 1993.

I. Cortes-Ciriano, A. Bender, and T. E. Malliavin. Comparing the influence of simulated experimental errors on 12 machine learning algorithms in bioactivity modeling using 12 diverse data sets. *Journal of chemical information and modeling*, 2015.

J. Couto. Kernel k-means for categorical data. In A. Famili, J. Kok, J. Pena, A. Siebes, and A. Feelders, editors, *Advances in Intelligent Data Analysis VI*, volume 3646 of *Lecture Notes in Computer Science*, pages 46–56. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-28795-7. doi: 10.1007/11552253_5. URL `http://dx.doi.org/10.1007/11552253_5`.

Cronin and Schultz. Pitfalls in qsar. *Journal of Molecular Structure: THEOCHEM*, 622(1):39–51, 2003. doi: doi:10.1016/S0166-1280(02)00616-4. URL `http://www.ingentaconnect.com/content/els/01661280/2003/00000622/00000001/art00616`.

M. Cronin, J. Dearden, G. Moss, and G. Murray-Dickson. Investigation of the mechanism of flux across human skin in vitro by quantitative structure–permeability relationships. *European Journal of Pharmaceutical Sciences*, 7(4):325–330, 1999.

S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5):352–359, 2002.

N. El Tayar, R.-S. Tsai, P. Vallat, C. Altomare, and B. Testa. Measurement of partition coefficients by various centrifugal partition chromatographic techniques: a comparative evaluation. *Journal of Chromatography A*, 556(1):181–194, 1991.

P. M. Elias, E. R. Cooper, A. Korc, and B. E. Brown. Percutaneous transport in relation to stratum corneum structure and lipid composition. *Journal of Investigative Dermatology*, 76(4):297–301, 1981.

A. Fick. V. on liquid diffusion. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 10(63):30–39, 1855.

G. L. Flynn. Physicochemical determinants of skin absorption, 1990.

B. Fritzke et al. A growing neural gas network learns topologies. *Advances in neural information processing systems*, 7:625–632, 1995.

N. Gibson, S. Aigrain, S. Roberts, T. Evans, M. Osborne, and F. Pont. A gaussian process framework for modelling instrumental systematics: application to transmission spectroscopy. *Monthly Notices of the Royal Astronomical Society*, 419(3):2683–2694, 2012.

N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001.

T. Hastie, R. Tibshirani, and J. J. H. Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001.

C. R. Houck, J. Joines, and M. G. Kay. A genetic algorithm for function optimization: a matlab implementation. *NCSU-IE TR*, 95(09), 1995.

I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

D. P. Kroese, T. Brereton, T. Taimre, and Z. I. Botev. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392, 2014. ISSN 1939-0068. doi: 10.1002/wics.1314. URL `http://dx.doi.org/10.1002/wics.1314`.

L. T. Lam, Y. Sun, N. Davey, R. Adams, M. Prapopoulou, M. B. Brown, and G. P. Moss. The application of feature selection to the development of gaussian process models for percutaneous absorption. *Journal of Pharmacy and Pharmacology*, 62(6):738–749, 2010.

E. J. Lien and H. Gaot. Qsar analysis of skin permeability of various drugs in man as compared to in vivo and in vitro studies in rodents. *Pharmaceutical research*, 12(4): 583–587, 1995.

C. W. Lim, S. ichi Fujiwara, F. Yamashita, and M. Hashida. Prediction of human skin permeability using a combination of molecular orbital calculations and artificial neural network. *Biological and Pharmaceutical Bulletin*, 25(3):361–366, 2002. doi: 10.1248/bpb.25.361.

M. Linting and A. van der Kooij. Nonlinear principal components analysis with catpca: a tutorial. *Journal of personality assessment*, 94(1):12–25, 2012.

H. S. Loos and B. Fritzke. Demogng v1. 5, 1998.

D. J. MacKay. Gaussian processes-a replacement for supervised neural networks? 1997.

B. M. Magnusson, Y. G. Anissimov, S. E. Cross, and M. S. Roberts. Molecular size as the main determinant of solute maximum flux across the skin. *Journal of Investigative Dermatology*, 122(4):993–999, 2004.

A. N. Martin, J. Swarbrick, A. Cammarata, and A. Chun. Physical pharmacy: physical chemical principles in the pharmaceutical sciences. 1993.

L. Michielan and S. Moro. Pharmaceutical perspectives of nonlinear qsar strategies. *Journal of chemical information and modeling*, 50(6):961–978, 2010.

T. P. Minka and R. W. Picard. Learning how to learn is learning with point sets. *Web. Revised*, 1999.

M. Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

G. Moss, J. Dearden, H. Patel, and M. Cronin. Quantitative structure permeability relationships (qsprs) for percutaneous absorption. *Toxicology in Vitro*, 16(3):299 – 317, 2002. ISSN 0887-2333. doi: http://dx.doi.org/10.1016/S0887-2333(02)00003-6. URL `http://www.sciencedirect.com/science/article/pii/S0887233302000036`.

G. P. Moss and M. T. Cronin. Quantitative structure–permeability relationships for percutaneous absorption: re-analysis of steroid data. *International journal of pharmaceutics*, 238(1):105–109, 2002.

G. P. Moss, Y. Sun, M. Prapopoulou, N. Davey, R. Adams, W. J. Pugh, and M. B. Brown. The application of gaussian processes in the prediction of percutaneous absorption. *Journal of Pharmacy and Pharmacology*, 61(9):1147–1153, 2009.

G. P. Moss, Y. Sun, S. C. Wilkinson, N. Davey, R. Adams, G. P. Martin, M. Prapopopolou, and M. B. Brown. The application and limitations of mathematical modelling in the prediction of permeability across mammalian skin and polydimethylsiloxane membranes. *Journal of Pharmacy and Pharmacology*, 63(11):1411–1427, 2011.

G. P. Moss, S. C. Wilkinson, and Y. Sun. Mathematical modelling of percutaneous absorption. *Current Opinion in Colloid & Interface Science*, 17(3):166–172, 2012.

G. P. Moss, D. R. Gullick, and S. C. Wilkinson. Predictive methods in percutaneous absorption. 2015.

I. Nabney. *NETLAB: algorithms for pattern recognition*. Springer, 2002.

D. Neumann, O. Kohlbacher, C. Merkwirth, and T. Lengauer. A fully computational model for predicting percutaneous drug absorption. *Journal of chemical information and modeling*, 46(1):424–429, 2006.

U. Norinder. Support vector machine models in drug design: applications to drug transport processes and qsar using simplex optimisations and variable selection. *Neurocomputing*, 55(1):337–346, 2003.

O. Obrezanova and M. D. Segall. Gaussian processes for classification: Qsar modeling of admet and target activity. *Journal of chemical information and modeling*, 50(6):1053–1061, 2010.

O. Obrezanova, G. Csányi, J. M. Gola, and M. D. Segall. Gaussian processes: a method for automatic qsar modeling of adme properties. *Journal of chemical information and modeling*, 47(5):1847–1857, 2007.

W. Pedrycz and S. Chen. Time series analysis, modeling and applications. *A Computational Intelligence Perspective (e-book Google)*, 2013.

D. Petelin, A. Grancharova, and J. Kocijan. Evolving gaussian process models for prediction of ozone concentration in the air. *Simulation modelling practice and theory*, 33: 68–80, 2013.

R. O. Potts and R. H. Guy. Predicting skin permeability. *Pharmaceutical research*, 9(5): 663–669, 1992.

R. O. Potts and R. H. Guy. A predictive algorithm for skin permeability: the effects of molecular size and hydrogen bond activity. *Pharmaceutical research*, 12(11):1628–1633, 1995.

M. Prapopoulou. *The development of a mathematical / computational model to predict drug absorption across the skin*. PhD thesis, PhD thesis, King's College London, 2012.

W. Pugh, R. Wong, F. Falson, B. Michniak, and G. Moss. Discriminant analysis as a tool to identify compounds with potential as transdermal enhancers. *Journal of pharmacy and pharmacology*, 57(11):1389–1396, 2005.

C. E. Rasmussen. *Gaussian processes in machine learning*. Springer, 2004.

C. E. Rasmussen. Christopher ki williams gaussian processes for machine learning, 2006a.

C. E. Rasmussen. Gaussian processes for machine learning. 2006b.

C. E. Rasmussen and H. Nickisch. The gpml toolbox version 3.5. 2015.

K. I. W. Rasmussen. Gaussian processes for machine learning @ONLINE, 2006c. URL `http://www.gaussianprocess.org/gpml/`.

M. Roberts, R. Anderson, and J. Swarbrick. Permeability of human epidermis to phenolic compounds. *Journal of pharmacy and pharmacology*, 29(1):677–683, 1977.

R. J. Scheuplein. Mechanism of percutaneous absorption: Ii. transient diffusion and the relative importance of various routes of skin penetration** from the research laboratories of the department of dermatology of the harvard medical school at the massachusetts general hospital, boston, massachusetts 02114. *Journal of Investigative Dermatology*, 48 (1):79–88, 1967.

R. J. Scheuplein. Permeability of the skin. *Comprehensive Physiology*, 2011.

A. Shah, G. P. Moss, Y. Sun, R. Adams, N. Davey, and S. Wilkinson. Using a support vector machine and sampling to classify compounds as potential transdermal enhancers. In *Artificial Neural Networks and Machine Learning–ICANN 2012*, pages 499–506. Springer, 2012.

A. Shah, Y. Sun, R. G. Adams, N. Davey, S. C. Wilkinson, and G. P. Moss. Support vector regression to estimate the permeability enhancement of potential transdermal enhancers. *Journal of Pharmacy and Pharmacology*, 2016.

J. R. Shewchuk. An introduction to the conjugate gradient method without the agonizing pain, 1994.

J. Shlens. A Tutorial on Principal Component Analysis. *ArXiv e-prints*, Apr. 2014.

A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

E. L. Snelson. *Flexible and efficient Gaussian process models for machine learning*. PhD thesis, Citeseer, 2007.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

E. W. Steyerberg, M. J. Eijkemans, F. E. Harrell, and J. D. F. Habbema. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in medicine*, 19(8):1059–1079, 2000.

Y. Sun, G. P. Moss, M. Prapopoulou, R. Adams, M. B. Brown, and N. Davey. Prediction of skin penetration using machine learning methods. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 1049–1054. IEEE, 2008.

Y. Sun, G. P. Moss, M. Prapopoulou, R. Adams, M. B. Brown, and N. Davey. The application of gaussian processes in the prediction of percutaneous absorption for mammalian and synthetic membranes. In *ESANN*, 2010.

Y. Sun, M. Brown, M. Prapopoulou, N. Davey, R. Adams, and G. Moss. The application of stochastic machine learning methods in the prediction of skin penetration. *Applied Soft Computing*, 11(2):2367–2375, 2011.

Y. Sun, M. B. Brown, M. Prapopoulou, R. Adams, N. Davey, and G. P. Moss. The application of gaussian processes in the predictions of permeability across mammalian membranes. In *Artificial Neural Networks and Machine Learning–ICANN 2012*, pages 507–514. Springer, 2012.

V. N. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

E. Wakelam, N. Davey, Y. Sun, A. Jefferies, P. Alva, and A. Hocking. The mining and analysis of data with mixed attribute types. In *Proceedings: IMMM 2016: Sixth International Conference on Advances in Information Mining and Management*, pages 55–62, 2016.

A. Williams. *Transdermal and topical drug delivery*, volume 17. Pharmaceutical Press London, 2003.

A. Wilschut, F. Wil, P. J. Robinson, and T. E. McKone. Estimating skin permeation. the validation of five mathematical skin permeation models. *Chemosphere*, 30(7):1275–1296, 1995.

G. Winter, J. Periaux, M. Galan, and P. Cuesta. *Genetic algorithms in engineering and computer science*. John Wiley & Sons, Inc., 1996.

D. Woolfson and D. McCafferty. *Percutaneous local anaesthesia*. CRC Press, 1993.

C. Xue, R. Zhang, H. Liu, X. Yao, M. Liu, Z. Hu, and B. T. Fan. Qsar models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine. *Journal of chemical information and computer sciences*, 44(5): 1693–1700, 2004.