CHEMOMETRICS AND MULTIBLOCK METHODS FOR QUANTITATIVE STRUCTURE-ACTIVITY STUDIES OF ARTEMISININ ANALOGUES AND POLYCHLORINATED DIPHENYLETHERS

ROSMAHAIDA JAMALUDIN

UNIVERSITI TEKNOLOGI MALAYSIA

CHEMOMETRICS AND MULTIBLOCK METHODS FOR QUANTITATIVE STRUCTURE-ACTIVITY STUDIES OF ARTEMISININ ANALOGUES AND POLYCHLORINATED DIPHENYLETHERS

ROSMAHAIDA JAMALUDIN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Chemistry)

Faculty of Science
Universiti Teknologi Malaysia

SEPTEMBER 2015

To Allah (SWT) and my beloved family

ACKNOWLEDGEMENT

The very first words that came through my mind was Syukur Alhamdulillah. It is of great honor for me to take this opportunity to thank many people whose contributions, helps and encouragement are so valuable throughout the journey.

First and foremost, I wish to extend my greatest appreciation to my supervisor, Professor Dr Mohamed Noor Hasan and my co-supervisor Dr Mohd Zuli Jaafar who have been instrumental in providing me beneficial guidance, valuable advices, kind supervision, patience and confidence on me throughout the course of the study. Sincere gratitude also goes to Dr Barry Lavine from Oklahoma State University, USA and Dr Neni Frimayanti from University Malaya for their professional assistance as well as interest towards the research. All the knowledge transfer and kindness will always be cherished in my heart.

Special thanks are due to all my colleagues, Alvin, Zalikha, Fatin and Bishir for their kind co-operation, undivided assistance as well as willingness to share knowledge and experiences. Wishing you good luck in your future.

Last but not least, I wish to forward my deepest gratitude to my husband, Hairil Anuar, my late father, Jamaludin Jaafar, my mother, Rosnah Juleh and my children, Amirul Hafidz, Akmal Darwisy, Alia Nurmaisarah, Adam Nurluqman and Ayman Nurqayyum for their understanding and strong encouragement. I love all of you. May Allah bless us.

ABSTRACT

Three major aspects of chemometrics have been investigated in this study namely Quantitative Structure-Activity Relationship (QSAR) and database mining, classification and multiblock methods. In the first analysis, 197 artemisinin compounds were divided into training set and test set together with structural descriptors generated by DRAGON 6.0 software had been used to develop three QSAR models. Statistics of the models were (r^2/ r_{test}^2) 0.790/0.853 for Forward Stepwise-Multiple Linear Regression (MLR), 0.807/0.789 for Genetic Algorithm (GA)-MLR and 0.795/0.811 for GA-Partial Least Square (PLS). The rigorously validated QSAR models were then applied to mine a chemical database which resulted in four potential new anti-malarial agents. The same artemisinin data set was then classified into active and less active compounds to develop reliable predictive classification models and to investigate the consequences of using various data splitting and data pre-processing methods on classification. Principal Component Analysis (PCA) and boundary plot had been utilized to visualize the four classifiers namely Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Linear Vector Quantization (LVQ) and Quadratic Discriminant Analysis (QDA). Kennard-Stone data splitting and standardization had produced better results in terms of percent correctly classified (% CC) compared to Duplex data-splitting and mean-centering. Moreover, LDA was found to be superior as compared to the other three classifiers with lower risk of over-fitting. Lastly, multiblock analysis methods such as Multiblock PLS and Consensus PCA have been implemented on polychlorinated diphenyl ethers (PCDEs) data set together with their respective descriptors blocked into three groups labelled as X_{1D}, X_{2D}, X_{3D} and a property block, Y which consists of log P_L (Pa, 25°C), log K_{OW} (25°C) and log S_{WL} (mol/L, 25°C). Their performance were then compared to single block methods that is PLS and PCA. The PLS models of each descriptor block with respect to each property were statistically best-fitted and well predicted with r_{train}^2 values greater than 0.96 while the r_{test}^2 values range from 0.86 to 0.98. It is interesting to note that the combination of the three descriptor blocks into a single block to produce Multiblock PLS super-scores (MBSS) model which was superior than Multiblock PLS block-scores (MBBS) yielded slightly better r_{train}^2 value and significantly better prediction with higher r_{test}^2 as compared to PLS model of individual descriptor block. In addition, three measures of block similarity such as Mantel Test, R_v coefficient and Procrustes analysis were used to investigate similarity and correlation between the blocks along with Monte Carlo simulations to determine their significance. Based on the similarity index between two blocks, X_{1D} descriptors resembled Y block better while X_{2D} was more correlated to X_{1D} block. In short, the chemometric methods had been applied successfully on both data sets using various descriptors generated by DRAGON software and yielded promising results beneficial not only in chemometrics area but also in drug design.

ABSTRAK

Tiga aspek utama bidang kimometrik telah disiasat dalam kajian ini iaitu kaedah Hubungan Kuantitatif Struktur-Aktiviti (QSAR) dan pangkalan data, klasifikasi dan multiblok. Dalam analisis yang pertama, 197 sebatian artemisinin telah dibahagikan kepada set latihan dan set ujian beserta deskriptor struktur yang dijanakan oleh perisian DRAGON 6.0 telah diguna untuk menghasilkan tiga model QSAR. Statistik model ialah (r^2/ r_{test}^2) 0.790/0.853 bagi kaedah Langkah Maju-Regresi Linear Berganda (MLR), 0.807/0.789 bagi Algoritma Genetik (GA)-MLR dan 0.795/0.811 bagi GA-Regresi Linear Separa (PLS). Model QSAR yang sah digunakan untuk mencari dalam pangkalan data kimia lalu menghasilkan empat bahan kimia baharu yang berpotensi sebagai agen anti malaria. Set data artemisinin yang sama kemudian dikelaskan kepada aktif dan kurang aktif untuk membina model klasifikasi, di samping menyiasat kesan penggunaan pelbagai teknik pemisahan dan pra-prosesan data terhadap klasifikasi. Analisis Komponen Prinsipal (PCA) dan plot sempadan telah digunakan untuk menggambarkan empat jenis model klasifikasi iaitu Mesin Vektor Sokongan (SVM), Analisis Pembezalayan Linear (LDA), Pengkuantuman Vektor Linear (LVQ) dan Analisis Pembezalayan Kuadratik (QDA). Kaedah Kennard-Stone dan pra-prosesan piawai telah menghasilkan keputusan yang lebih baik dari segi peratus pengkelasan yang betul (% CC) berbanding Duplex dan pra-prosesan purata-tengah. Di samping itu, LDA didapati lebih baik dengan risiko suaian lampau yang lebih rendah. Akhir sekali, analisis multiblok seperti Multiblok PLS dan konsensus PCA telah dijalankan ke atas set data poliklorin difenil eter (PCDEs) beserta dengan tiga kumpulan blok deskriptor masing-masing iaitu X_{1D}, X_{2D}, X_{3D} dan blok sifat, Y yang terdiri daripada log P_L (*Pa*, 25°C), log K_{OW} (25°C) and log S_{WL} (*mol/L*, 25°C). Prestasi kaedah ini seterusnya dibandingkan dengan kaedah blok tunggal iaitu PLS dan PCA. Model PLS setiap blok deskriptor terhadap setiap sifat secara statistiknya best-fitted dan ramalan baik dengan nilai r_{train}^2 lebih besar daripada 0.96 manakala nilai r_{test}^2 adalah dalam julat 0.86 hingga 0.98. Sesuatu yang menarik untuk diperhatikan bahawa gabungan tiga blok deskriptor ke dalam blok tunggal menghasilkan model Multiblok PLS Super-Skor (MBSS) yang lebih baik daripada Multiblok PLS Blok-Skor (MBBS) menghasilkan nilai r_{train}^2 dan r_{test}^2 yang lebih tinggi berbanding model PLS blok deskriptor individu. Sebagai tambahan, tiga pengukuran keserupaan blok seperti ujian *Mantel*, pekali R_v dan analisis *Procrustes* telah digunakan untuk menyiasat keserupaan dan korelasi antara blok diikuti simulasi *Monte Carlo* untuk menentukan kepentingannya. Berdasarkan indeks keserupaan antara dua blok, deskriptor X_{1D} lebih menyerupai blok Y manakala deskriptor X_{2D} mempunyai korelasi lebih kepada blok X_{1D}. Ringkasnya, kaedah kimometrik telah berjaya digunakan ke atas kedua-dua set data menggunakan pelbagai deskriptor yang dijanakan oleh perisian DRAGON dan menghasilkan keputusan bermanfaat bukan sahaja dalam bidang kimometrik tetapi juga bidang rekabentuk ubatan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

%CC	-	Percentage correctly classified
ALL-QSAR	-	Automated lazy learning QSAR
BS	-	Block-scores
CPCA	-	Consensus Principal Component Analysis
CNN	-	Computational Neural Network
CoMFA	-	Comparative Molecular Field Analysis
CV	-	Cross validation
EDC	-	Euclidean Distance to Centroids
EDDFA	-	Electron Density-Derived Field Analysis
ETA	-	Extended Topochemical Atom
FRED	-	Fast random elimination of descriptors
GA	-	Genetic Algorithm
GAPLS	-	Genetic Algorithm Partial Least Squares
GA-VSS	-	Genetic Algorithm-Variable Subset Selection
GFA	-	Genetic Function Approximation
GUI	-	Graphical user interface
HCA	-	Hierarchical Cluster Analysis
HQSAR	-	Hologram QSAR
ICS	-	International Chemometrics Society
IC_{50}.	-	50% inhibitory concentration and reported in ng/ml
IMDDI-	-	Individual molecular data set diversity index
K-ANN	-	Kohonen Artificial Neural Network
k-NN	-	*k*-Neural Network
kNN	-	*k*-Nearest Neighbour
LDA	-	Linear Discriminant Analysis
LF	-	Linear Function

LFER	-	Linear Free Energy Relationships
LMO	-	Leave-many-out cross validation
LOOCV	-	Leave-one-out cross-validation
log RA	-	Logarithms of relative activity
log K_{ow}	-	*n*-octanol/water partition coefficient
log P_L	-	298K supercooled liquid vapour pressures
log $S_{w,L}$	-	Aqueous solubilities
log ED_{50}	-	Immunotoxicity values
LSO	-	Leave-several-outcross validation
LV	-	Latent variables
LVQ	-	Linear Vector Quantization
MBBS	-	Multiblock PLS block-scores
MBPLS	-	Multiblock Partial Least Squares
MBSS	-	Multiblock PLS super-scores
MCI	-	Molecular connectivity indices
MIR	-	Mid-infrared
MLR	-	Multiple Linear Regression
MM2	-	Molecular mechanics 2
MMDDI	-	Molecular data set diversity index,
MOPAC	-	Molecular Orbital Package
Mp	-	Modelling power
MW	-	Molecular weight
NCI	-	National Cancer Institute
NIPALS	-	Nonlinear Iterative Partial Least Squares
NIR	-	Near-infrared
NMR	-	Nuclear Magnetic Resonance
PC	-	Principal Component
PCA	-	Principal Component Analysis
PCR	-	Principal Component Regression
PCDE	-	Polychlorinated diphenyl ethers
PF	-	Polynomial Function
PLS	-	Partial Least Squares
PLSDA	-	Partial Least Squares Discriminant Analysis
POPs	-	Persistent Organic Pollutants (POPs)

QDA	-	Quadratic Discriminant Analysis
QHS	-	Qinghaosu
QSAR	-	Quantitative-Structure Activity Relationship
QSPR	-	Quantitative-Structure Property Relationship
RA	-	Relative activity
RBF	-	Radial Basis Function
RDA	-	Regularised Discriminant Analysis
RMSE	-	Root-mean-square-error
RMSEC	-	Root-mean-square-error of calibration
RMSECV	-	Root-mean-square error of cross-validation
RMSEP	-	Root-mean-square error of prediction
RRT	-	Relative retention time
SA-PLS	-	Simulated annealing-partial least square
SAR	-	Structure Activity Relationship
SF	-	Sigmoid Function
SIMCA	-	Soft Independent Modelling of Class Analogy
SOMs	-	Self Organising Maps
SS	-	Super-scores'
SVD	-	Singular Value Decomposition
SVM	-	Support Vector Machine
TAE	-	Transferable Atom Equivalent
UFS	-	Usupervised forward selection
VR	-	Volume ratio
VSA	-	Van der Waals surface area

LIST OF SYMBOLS

r^2	-	Correlation coefficient for training set
r_{cv}^2	-	Correlation coefficient for cross-validation
r_{test}^2	-	Correlation coefficient for test set
X_{1D}	-	0-dimensional and 1-dimensional descriptor block
X_{2D}	-	2-dimensional descriptor block
X_{3D}	-	3-dimensional descriptor block
P_a	-	Atmospheric pressure
mol/L	-	Mol per liter
d_n	-	n^{th} structural descriptors of QSAR model
a_n	-	n^{th} coefficients of QSAR model
$\bar{y}$	-	Average value of the dependent variable
y_i,	-	Measured value of the dependent variable
$\hat{y}_i$	-	Predicted value of the dependent variable
X	-	Scores matrix
P	-	Loadings matrix
E	-	Residual matrix
a	-	Slope of the line
b	-	The intercept of the line on the y-axis
y	-	Dependent variables
t_i	-	Latent variables or LVs
σ	-	The standard deviation of these Euclidean distances
Z	-	An arbitrary parameter to control the significance level
D_T	-	Applicability domain threshold
r	-	Pairwise correlation coefficient
C	-	Capacity or penalty parameter
K	-	Kernel type

σ	-	Width of the RBF
d	-	Mahalanobis distance
x_i	-	Measurement obtained for the *i*th sample
$\bar{\mathrm{x}}_{\mathrm{g}}$	-	Centroid of class *g*
$\mathrm{C}_{\mathrm{P}}^{-1}$	-	Inverse of the pooled variance-covariance matrix
y_i	-	Class membership of the sample
w	-	A vector of SVM weights and
b	-	A bias term used
K	-	Slope of the regression lines
Xx	-	Information matrix
X	-	Data matrix
$\bar{\mathrm{y}}_{\mathrm{tr}}$	-	Average value of the dependent variable
x_{im}	-	Values of m^{th} descriptor for compounds *i*
x_{jm}	-	Values of m^{th} descriptor for compounds *j*
$\bar{\mathrm{x}}_{\mathrm{j}}$	-	Mean for variable *j*
I_g	-	Correspond to number of samples
C_g	-	Variance–covariance matrix of class *g*
ψ_0	-	Initial learning rate
$X_{\boldsymbol{R}}$	-	Reference block
$X_{\boldsymbol{c}}$	-	Comparison block
${}^{scale}T_R$	-	Scaled scores of reference matrix in procrustes
${}^{cen}T_R$	-	Mean centered scores of reference matrix in procrustes
${}^{rot}T_C$	-	Rotation of the scores of comparison matrix in procrustes
X_s	-	Super matrix
T_{SS}	-	PLS super-scores matrix
T_{BS}	-	PLS block scores matrix
q	-	*y* loading
${}^{resid}X_b$	-	Residual of block *b*
$\mathrm{t}_{\mathrm{X_b}}$	-	PLS block scores vector for block *b*
$\mathrm{h}_{\mathrm{X_b}}$	-	PLS block weight for block *b*
h_T	-	PLS super-weights
t_T	-	PLS super-scores vector from two or more *X* block

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of Research

Chemometrics and cheminformatics is a multi-disciplinary information science (Kowalski, 1981) that integrates subject areas like chemistry, mathematics, statistics, biology and computer science. These fields are expanding rapidly and considerably in recent years due to the increasing computational power together with advances in computer technology and data analysis as well where huge amount of data is readily available with increasing efficiency of chemical information storage and retrieval capabilities. Furthermore, the data can be analysed rapidly and efficiently with improved analytical measurements, modern analytical instrumentation for data acquisition, storage, display and processing as well as detecting and correcting analytical instruments problems that can then easily be converted to useful information and knowledge especially in pharmaceutical and environmental areas. Hence, computer revolution leads to a new branch of analytical chemistry, namely Chemometrics.

Chemometrics was first introduced in 1971 by Svante Wold who three years later collaborated with Bruce Kowalski to form The International Chemometrics Society (ICS). Chemometrics involves the implementation of statistical and mathematical methods analogously similar to biometrics, econometrics and psychometrics but concentrated only on chemical data and practices (Wold, 1995).

According to Massart (1997), chemometrics can be defined as "chemical discipline that uses mathematics, statistics and formal logic (a) to design or select optimal experimental procedures; (b) to provide maximum relevant chemical information by analyzing chemical data; and (c) to obtain knowledge about chemical systems".

Basically, chemometrics is an application-driven discipline where the focus of chemometrician is to develop solutions to chemical problems such as in multivariate calibration and pattern recognition. Major applications of chemometrics include exploratory data analysis, multivariate regression or calibration, clustering and classification as well as variable selection. Besides application, chemometrics also covers fundamentals and methodology. However, Wold (1995) strongly believes that chemometrics should focus on chemical problem-solving rather than method development. Several computer programs or software packages have been developed for specific instruments or for general use in chemometrics such as PLS Toolbox (Eigenvector_Research_Inc., 2010) and Unscrambler (Camo_Software_AS, 2010). Nevertheless, some of the expert algorithm developers in chemometrics prefer to use programming language like MATLAB (The_Mathworks_Inc., 2008) as platform for method development due to its flexibility. Moreover, modification of chemical method and development of new data analysis techniques may be required to handle complex data analysis problems (Lavine, 2000).

At first, limited number of chemometrics articles on methods and applications were published in a separate section in Analytica Chimica Acta and Analytical Chemistry under the series entitled ‘Computer Techniques and Optimization’ and 'Statistical and Mathematical Methods in Analytical Chemistry' which later changed to 'Chemometrics' respectively. When this area became increasingly and widely accepted, journal publication dedicated to chemometrics, namely Chemometrics and Intelligent Laboratory Systems was first published in 1986 followed by Journal of Chemometrics which currently covers mostly methodology and fundamentals of chemometrics (Hopke, 2003). However, various applications of chemometrics would be presented in the broader analytical or application-oriented journals such as Applied Spectroscopy and SAR and QSAR in Environmental Research.

Interestingly, application areas of chemometrics have spread and contributed to other disciplines that involve chemical instrumentation such as process engineering and environmental science as well as represented as new domain like cheminformatics, process modelling, genomics and proteomics. Cheminformatics also known as chemical informatics is a subfield of chemometrics that was introduced in the late 1990s where it integrated other disciplines like computational chemistry, molecular modelling and chemical information to solve problems in chemistry (Gasteiger, 2006). As defined by Brown (2005), this interesting new field is "the mixing of those information resources to transform data into information and information to knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization". Chonde (2014) discusses the progress of three stages of cheminformatic research area that includes capturing, storing and mining data. Thus, the application of cheminformatics includes storage and retrieval of large amount of data or information relating to compounds, virtual screening and QSAR or QSPR especially in drug discovery and development (Leach and Gillet, 2003). There are altogether thirteen main journals in cheminformatics research area such as The Journal of Chemical Information and Modelling, Journal of Chemical Theory and Computation, Journal of Cheminformatics, and Drug Discovery Today where 40% of them are dedicated to biological research and drug design (Chonde and Kumara, 2014).

The development of new compounds with specialized properties particularly drugs is becoming more interesting due to rapid advancement in technology and increasing demand for new drugs. In medicinal chemistry, the traditional process of producing new chemical compounds with novel properties requires laborious screening and testing which involves lengthy, very time consuming and costly process. As an alternative, computer has been used as tool to facilitate the design and discovery of new drugs. The computing devices able to handle huge amount of data in a relatively short period of time, visualize molecules and gain better insight into the chemical and biological impacts of the problem at hand with least efforts yet yielding maximum information. Moreover, significant advances in information technology and widespread availability of public databases further support the development and enhancement of established computational methodologies

(Agrafiotis *et al.*, 2007). Such technique should narrow down the number of potential molecules to be tested for their biological, physical and chemical properties. This consequently minimizes the costs, time and efforts involved in drug research.

In addition, the increase in resistance to older drugs and newly discovered types of infections such as mutated bacterial and viral infection have created an urgent and continuous need for discovery and development of new drugs (Gozalbes *et al.*, 2002). Quantitative Structure Activity Relationship (QSAR) that offers valuable information about biological predictivity represents one of the best computationally inexpensive methodology in the design of potential bioactive drugs.

1.2 Quantitative Structure-Activity Relationships

Quantitative-Structure Activity Relationship or commonly known in abbreviated form as QSAR is an important area in chemometrics and chemistry in general. It is a statistical analysis which directly calculates physical and biological properties of molecules from their physical structure. Based on the definition of QSAR above, the objective of a QSAR model is to develop inductively relationship between structure and property using information extracted from a set of numerical descriptors characterizing molecular structures.

Figure 1.1 illustrates that the molecular structure of a compound is somehow related to its property. Since the exact relationship is not known, an indirect approach is used which consists of two main parts (Gasteiger, 2006). The first part is calculation of structural descriptors that represent molecular structure of each compound. Next, selection of subsets of descriptors to develop model that predict the desired property.

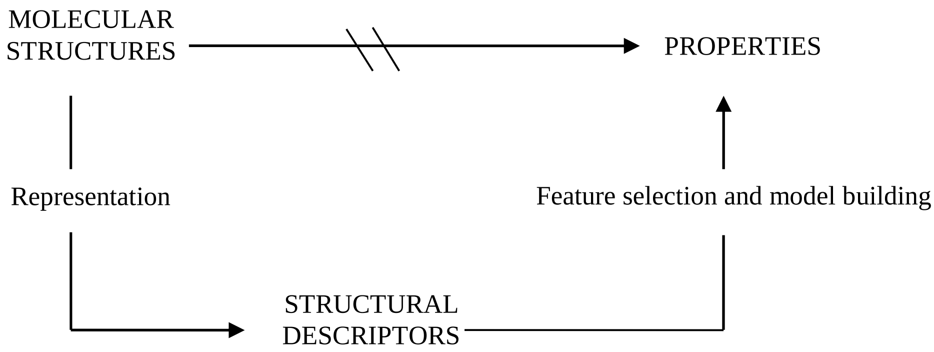

Figure 1.1 The general QSAR problem

Theoretically, QSAR is a modelling technique in which the observed activities or properties of chemical compounds are correlated with structural descriptors derived from the molecular structures and can be represented by mathematical equation as shown below:

$$\text{Molecular activity} = f\,(\text{descriptor}) = a_1d_1 + a_2d_2 + a_3d_3 + \ldots + a_nd_n \qquad (1.1)$$

where $d_1, d_2, d_3, \ldots d_n$ are structural descriptors and $a_1, a_2, a_3 \ldots a_n$ are coefficients. The correlation model developed can be utilized to predict activities of compounds not included in the model development process, to form the basis for understanding factors affecting their activity or to get better understanding of interactions between molecules (Parvu, 2003). Hence, a medicinal chemist has to focus making inferences from molecular properties and structural descriptors because the interaction mechanism of how drugs exert their biological effects is complex and mostly unknown.

In this study, structure-activity relationship approach as discussed above will be implemented to develop models that can correlate structural features of the compounds obtained from literature with their anti malarial-activity. Good models developed using this method will be applied to screen large chemical databases.

Results of the screening probes can be used to postulate structure of lead molecules that can be synthesized in the production of new drugs in the pharmaceutical industries.

1.3 Chemometrics

Regression and classification methods are commonly used in chemometrics and employed extensively in this research. Regression or calibration relates samples to one or more continuous numerical properties and consequently can be used to predict actual value of the response. Traditional chemical or physical relationship usually consider one or few variables at the same time. Univariate regression deals with only one variable while multivariate regression such as PLS involve more than one variable and take into account joint effect of all variables. Typically, multivariate regression is used to predict chemical activity of interest based on the relationship between specific response and corresponding data generated by instruments such as Mass Spectrometer, Gas Chromatogram and Nuclear Magnetic Resonance (NMR) Spectrometer as well as the information extracted from molecular structure in the area of drug design as employed in this research. Examples of regression methods commonly used in chemometrics are MLR and PLS. In addition, new regression techniques have been introduced for instance Support Vector Regression (SVR) that based on Vapnik's concept of support vectors (Brereton and Lloyd, 2010; Smola and Schölkopf, 2004).

Regression methods discussed previously are known as single block approaches that simply relates two blocks of data *i.e.* response and variable blocks. Interestingly, these techniques can be further expanded involving more than two blocks of data using multiblock methods such as MBPLS. In this study, typical single block regression methods such as PLS has been utilized to measure the correlation of any two blocks together with the more advanced multiblock methods to find the correlation of more than two blocks of data with extra information on common trends and possible connection among the blocks. In addition, numerical techniques or indicators were used to investigate or measure the trends or relative fit between two blocks of data include Mantel Test, *Rv* coefficient, Procrustes analysis and Monte Carlo simulation.

Pattern recognition involves finding similarities and differences between chemical samples based on measurements made on the samples and can be divided into two parts that are supervised pattern recognition and unsupervised pattern recognition or cluster analysis. The latter is used to discover patterns in complex data sets or group similar objects together. Classification which is one of the main focus in this research falls under the category of supervised pattern recognition (Dunn III and Wold, 1980) that determines whether the samples can be related to groupings with the aim to classify the unknowns (Brereton, 2009). This can be achieved by using models developed from training set. Basically, there are two types of classification which are linear that use statistical methods such as Linear Discriminant Analysis (LDA), Regularised Discriminant Analysis (RDA) and non-linear or machine learning methods like *k*-Nearest Neighbour (kNN) and Support Vector Machine (SVM). In short, multivariate analysis was performed in this study to extract meaningful information efficiently from the data.

1.4 Problem Statement

Generally, QSAR methodologies are only effective for QSAR development when applied to structurally similar analogues data set. The larger structural variation of QSAR training set, constructing good QSAR model becomes harder. As a results, further application of QSAR models in screening very large chemical databases can probably be troublesome in any QSAR studies. Single QSAR model in high dimensional descriptor space cannot describe structure-activity correlations within a large database as well as unsuitable to represent large diverse data set of compounds. Instead, multiple QSAR models that consists of different combination of variable selection and model building should potentially be taken into consideration. Moreover, several models with different combination of descriptors could also help to understand the anti-malarial characteristic of artemisinin compounds.

The choice of appropriate classification methods for certain data set is usually highlighted in chemometric research since different type of data structure may require different type of classifier and no classification method is superior and applicable to all data set. Even nonlinear approaches that are capable of producing complex boundaries especially for complex data set unfortunately not a direct indication of its superiority since it has higher tendency of over-fitting. Therefore, comparison between several classification techniques is critical to determine the best method for a particular data set. The challenge in attempting to find the best classifiers for the data set become more complicated as other variables should also be taken into consideration. Several types of data splitting and data pre-processing methods have been selected and compared simultaneously besides changing the number of Principal Components (PCs) accordingly. Hence, this research on classification should produce not only the best method of classification but also the best pre-processing and data splitting techniques for artemisinin data set as well as reasonable number of PCs with minimum risk of over-fitting.

Since the number of descriptors generated by DRAGON software are significantly large and variety, it is needed to block them into several meaningful groups based on the types of structures they represent prior to building model. According to Zarzo (2004), PLS models with variable reduction often removes information, but splitting up the variables into a number of blocks and employing multiblock methods like Multiblock PLS and Consensus PCA not only analyze several blocks simultaneously, but also provide more information on the correlation and common trends between blocks. Hence, the influence of each group of descriptor variables on each property can be studied separately. Moreover, extra information can be obtained from specific parts of the block. In addition, the relationship between blocks can also be analysed as well resulting in easier interpretation of the data where all the indicators used should be consistent and parallel with the significance test. The reliability of multiblock approach compared to single block approach will be determined in this research. The multiblock analysis is expected to provide the overall picture of the model and data involved comparable to ordinary PLS method which has been widely used in QSAR research in terms of time and complexity.

1.5 Research Objectives

The main goal of this research is to study the structure-activity relationship of selected data sets and develop several models based on various combinations and advanced methods in chemometrics and hence subjected to relevant applications in chemistry. Therefore, the thesis can be divided into three main parts with the following major objectives corresponding to each category. The first objective of this research is to develop robust QSAR regression models applicable to high dimensional data (artemisinin data set) that are stable and predictive both internally and externally so as to correlate biological activity of chemical compounds in natural products with their structural characteristics. Consequently, these multiple computer models will be used to predict activity of new compounds and screen a large library of compounds in large database to discover or identify new compounds with specialized properties (anti-malarial agents).

The second objective is to develop and compare the performance of four types of classification models on artemisinin data set using different data pre-processing and data splitting methods at different number of PCs. Consequently, the most suitable method of data pre-processing, data splitting and efficient classification model for artemisinin data set can be determined. Thus, the selected binary classification model could predict accurately the anti-malarial activity of artemisinin directly from their molecular structures.

The last objective is to develop Multiblock PLS models and Consensus PCA on Polychlorinated Diphenyl Ethers (PCDEs) data set using three descriptor blocks and one property block. The performance of these multiblock methods will be compared to frequently used single block method that is PCA and PLS. At the same time, the similarity and correlation between the blocks will be investigated using three different similarity measures in order to determine common trends in these data and their level of influence on activity block.

1.6 Scope of Study

This research is based on 2 types of data sets. The first data set consists of 197 artemisinin compounds with anti-malarial activity measured as log *RA* (relative activity) (Avery *et al.*, 2002). The data set has been used in development of QSAR models and database mining. The same set of compounds was employed in the study of data pre-processing and data splitting for classification. The second data set was subjected to building multiblock models. It consists of 107 PCDEs compounds with three properties that are log P_L (Pa, 25°C), log K_{OW} (25°C) and log S_{WL} (mol/L, 25°C) (Yang *et al.*, 2003).

The descriptors used in this study should represent the molecular structure accordingly and should be relevant to describe the activity being studied as well as can be processed rapidly. Therefore, various types of descriptors generated by DRAGON descriptor generator (Todeschini *et al*., 2006) were used and can be categorized according to their dimensional property ranging from 0-dimensional to 3-dimensional descriptors. The study on development of QSAR models was split into two parts where the first part includes 3D descriptors while the other part exclude 3D descriptors for database mining purpose. Similarly, the study on classification only utilized 2D descriptors. On the other hand, all types of DRAGON descriptors have been used in the multiblock study where the descriptors were classified into three groups based on their dimensional property. The first block consisted of combination of 0D and 1D block while the other two blocks consisted of 2D descriptors and 3D descriptors.

This first part of structure-activity study focused on the development of QSAR models that correlate biological activity which is anti-malaria and chemical structures of artemisinin compounds. Genetic algorithm (GA) and forward stepwise were incorporated into the feature selection routine combined with regression tools namely Partial Least Square (PLS) and Multiple Linear Regression (MLR) in QSAR modelling. Mathworks Matlab 7.5 (2007) was used as the platform to build the QSAR models together with the latest version of PLS Toolbox 5.2. The resulting QSAR models were applied to mining chemicals in large database ("National Cancer Institute (NCI) Database ") for potentially active compounds.

In the classification research, four types of linear and nonlinear methods of classification have been used and compared with respect to artemisinin data set. The selected techniques were Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Linear Vector Quantization (LVQ) and Support Vector Machine (SVM). Their performances were evaluated in terms of percentage values using percent correctly classified of both training and test set as well as illustrated graphically using PCA and boundary plot. They were measured and recorded at different principal component (PC) number ranging from 1PC to 20PC. At the same time, data pre-processing that involved row scaling, standardization and mean centring had been investigated and the most suitable one has been identified for the data set. Besides that, Duplex and Kennard-Stone methods of data splitting had been employed and compared in this study.

Basically, the scope of multiblock study can be categorized to three different segments. The first part includes finding the correlations between the three descriptor blocks and a property block. Three types of indicators have been utilized to evaluate the similarities that were Mantel test, *Rv* coefficient and Procrustes analysis together with Monte Carlo simulation as significance test. The subsequent work incorporated both single block method and multiblock method and their results were analysed and compared. The second segment dealt with the visualization of the data set using PCA and CPCA methods. Next, the third part of the work involved building PLS model of each block and consequently, the multiblock PLS of all the blocks. Two types of multiblock model *i.e.* MBSS and MBBS had been employed in this work and their performance were also compared.

1.7 Significance of Study

The main significance of this research is to develop an improved method to discover new potentially active compounds with efficient QSAR modelling along with significant improvement in prediction of QSAR models. The approach in QSAR modelling should be applicable to diverse data set and other large database. In this study, data mining method has been implemented in the QSAR studies and the outcome is new potential anti-malarial compounds.

Another potential significance is in the utilization of natural products to develop anti-malarial agents and thus, successful development of new agents will increase the value of natural resources. As discussed earlier, there is an urgent need to develop effective agents against malaria and findings from this study can be fanned out in the production of new drugs particularly anti-malarial agents in pharmaceutical industries.

The significance of this QSAR study will be applicable to several industries especially pharmaceutical and biotechnological industries. The method can be extended and utilized in a wide variety of available experimental data sets with different biological activity or for a wider class of application. As a result, the cost and time in the development of new drugs will be minimized once the method can be proven to be able to select higher percentages of bioactive compounds as compared to conventional methods. The outputs expected from this research include methodology for building QSAR models and discovery of new compounds with anti-malaria activity. A successful implementation of this methodology would lead to an alternative way to generate and screen potential drug candidates.

The performance of data splitting methods and classification models has been evaluated using the percentage of correct classification. It enhances the significance of this study as the best combination of data pre-processing, data splitting and classification model for artemisinin data set can be identified. Hence, the simple classification scheme that categorized the compounds as active and inactive could be employed to prioritize compounds to be tested with in vivo and in vitro assays and to determine the possible activity in newly produced chemicals or in other words could also be used as a practical tool for the rapid screening of potential anti-malarial agents. At the same time, the consequences of increasing number of PCs on the classification models will be observed and this pattern can be used to determine the reliability of the model and potential risk of over-fitting. Thus, the same framework can be applied to other data sets and subsequently produce better classification results.

The novelty that can be found in the study on PCDEs is the implementation of multiblock methods. Based on previous literatures, the study on PCDEs are limited to single block method analysing only single property at a time. In this study, several properties of PCDEs can be modelled simultaneously and the importance of each category of descriptors can be assessed. Thus, the overall picture of properties and descriptors relationship can be illustrated and compared in a single analysis. Furthermore, time taken to analyze the data can be reduced significantly.

1.8 Layout of the Thesis

In general, the thesis is organized into seven main chapters. The introductory chapter begins with the discussion on the background of three main areas included in this research that are QSAR and data mining, classification and multiblock QSAR followed by their respective problem statements and research objectives. In addition, the scope and significance of the study have been presented as well.

The next chapter, namely literature review discusses and analyzes specific areas or issues through research, summary, classification and comparison of prior research studies and literatures. This section focuses on important topics pertinent to this study which include QSAR, descriptors, feature selection, model development, validation, data mining, classification and multiblock QSAR along with the overview of artemisinin and PCDEs data sets.

Chapter 3 presents detailed description of the chemometric methods used throughout the study. The development of QSAR models and data mining performed using genetic algorithm and forward stepwise combined with PLS and MLR methods have been explained in detail. Besides that, techniques on model validation and data mining have been discussed as well. The subsequent study utilized four types of classification (*i.e.* LDA, QDA, SVM and LVQ) and two data splitting methods (*i.e.* Duplex and Kennard-Stone). The last part of this research adopted four types of similarity measures to measure the similarity correlation between blocks of descriptors together with the development of multiblock PLS models.

Results and discussion are divided into three chapters. Chapter 4 presents the results of development QSAR models from artemisinin data set followed by the application of the models to search for new compounds in database mining. Then, chapter 5 discusses the results of classification of artemisinin data set using four classifiers and two data splitting methods as well as four conditions of data pre-processing. The results were evaluated and compared in terms of percent correctly classified of training and test set.

The application of multiblock methods on PCDEs data set is described in Chapter 6. The data set consisted of four blocks and their correlations were assessed using Mantel test, Procrustes analysis and R_v coefficient. The single block method like PCA and PLS were then compared with the multiblock method such as Consensus PCA (CPCA) and Multiblock PLS.

Finally, chapter 7 concludes the thesis with the brief discussion and summary of the results from each topic or analysis of the research. It highlights the novelty of the research findings, achievement and contribution of this study. In addition, the limitations and some suggestions for future research are also discussed.

REFERENCES

Abbasitabar, F., and Zare-Shahabadi, V. (2011). Development predictive QSAR models for artemisinin analogues by various feature selection methods: A comparative study. [doi: 10.1080/1062936X.2011.623316]. *SAR and QSAR in Environmental Research*, *23*(1-2), 1-15.

Abdi, H. (2007). RV Coefficient and Congruence Coefficient. In Salkin, N. (Ed.), *Encyclopedia of Measurement ans Statistics*. Thousand Oaks, CA: Sage

Abdi, H. (2010). Conguence: Congruence coefficient, RV -coefficient, and Mantel coefficient. In Salkind, N. (Ed.), *Encyclopedia of Research Design.* Thousand Oaks, CA: Sage.

Agrafiotis, D. K., Bandyopadhyay, D., Wegner, J. K., and Vlijmen, H. v. (2007). Recent Advances in Chemoinformatics. *J. Chem. Inf. Model. , 47*, 1279-1293.

AlGhazzawi, A., and Lennox, B. (2008). Monitoring a complex refining process using multivariate statistics. *Control Engineering Practice*, *16*(3), 294-307.

Andersen, C. M., and Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics, 24*(11-12), 728-737.

Avery, M. A., Alvim-Gaston, M., Rodrigues, C. R., Barreiro, E. J., Cohen, F. E., Sabnis, Y. A., et al. (2002). Structure-Activity Relationships of the Antimalarial Agent Artemisinin. 6. The Development of Predictive In Vitro Potency Models Using CoMFA and HQSAR Methodologies. *J. Med. Chem.* , *45*, 292-303.

Axelson, D. E. (2010). *Data Preprocessing for Chemometric and Metabolomic Analysis* (first ed.). Kingston, ON: MRi Consulting.

Bajorath, J. (2001). Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci., 41*(2), 233-245.

Balakrishnama, S., and Ganapathiraju, A. (1998). *Linear discriminant analysis: A brief tutorial.*: Institute for Signal and information processing,.

Bartlett, M. S. (1963). The spectral analysis of point patterns (with discussion). *J. Roy. Stat. Soc., Ser. B, 25*, 264–296.

Bashyal, S., and Venayagamoorthy, G. K. (2008). Recognition of facial expressions using Gabor wavelets and learning vector quantization. *Engineering Applications of Artificial Intelligence, 21*(7), 1056-1064.

Baumann, K. (1999). Uniform-length molecular descriptors for quantitative structure–property relationships (QSPR) and quantitative structure–activity relationships (QSAR): classification studies and similarity searching. *TrAC Trends in Analytical Chemistry, 18*(1), 36-46.

Baurin, N., Mozziconacci, J.-C., Arnoult, E., Chavatte, P., Marot, C., and Morin-Allory, L. (2003). 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. [doi: 10.1021/ci0341565]. *Journal of Chemical Information and Computer Sciences, 44*(1), 276-285.

Beebe, K. R., Pell, R. J., and Seasholtz, M. B. (1998). *Chemometrics: A practical Guide*. New York: John Wiley & Sons, Inc,.

Bishop, C. (1997). *Neural Networks for Pattern Classification*. Oxford: Clarendon Press.

Blasius, J., Eilers, P. H. C., and Gower, J. (2009). Better biplots. *Computational Statistics & Data Analysis, 53*(8), 3145-3158.

Bras, L. P., Bernardino, S. A., Lopes, J. A., and Menezes, J. C. (2005). Multiblock PLS as an approach to compare and combine NIR and MIR spectra in calibrations of soybean flour. *Chemometrics and Intelligent Laboratory Systems, 75*(1), 91-99.

Breneman, C. M., Thompson, T. R., Rhem, M., and Dung, M. (1995). Electron Density Modeling of Large Systems Using the Transferable Atom Equivalent Method. *Computers & Chemistry, 19*(3), 161.

Brereton, R. G. (2000). Introduction to multivariate calibration in analytical chemistry. *Analyst, 125*, 2125-2154.

Brereton, R. G. (2003). *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. Chichester, UK: John Wiley & Sons, Ltd.

Brereton, R. G. (2006). Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. *TrAC Trends in Analytical Chemistry, 25*(11), 1103-1111.

Brereton, R. G. (2009). *Chemometrics for Pattern Recognition*. Chichester, UK: John Wiley & Sons, Ltd.

Brereton, R. G., and Lloyd, G. R. (2010). Support Vector Machines for classification and regression. *Analyst, 135*, 230–267.

Bro, R., and Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics, 17*(1), 16-33.

Brown, F. (2005). Editorial opinion: chemoinformatics - a ten year update. *Current opinion in drug discovery & development, 8*(3), 298-302.

Brown, R. D., and Martin, Y. C. (1996). Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. [doi: 10.1021/ci9501047]. *Journal of Chemical Information and Computer Sciences, 36*(3), 572-584.

Brown, R. D., and Martin, Y. C. (1997). The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. [doi: 10.1021/ci960373c]. *Journal of Chemical Information and Computer Sciences, 37*(1), 1-9.

Burbidge, R., Trotter, M., Buxton, B., and Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & Chemistry, 26*(1), 5-14.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery 2*, 121-167.

CambridgeSoft. (2006). ChemOffice (Version 2006). Massachusetts, USA: PerkinElmer, Inc.

Camo_Software_AS. (2010). The Unscrambler X (Version 10.0). Oslo, Norway.

Capron, X., Walczak, B., de Noord, O. E., and Massart, D. L. (2005). Selection and weighting of samples in multivariate regression model updating. *Chemometrics and Intelligent Laboratory Systems 76*, 205– 214.

Capron X., Walczak B., Noord O.E. de, and D.L., M. (2005). Selection and weighting of samples in multivariate regression model updating. *Chemometrics and Intelligent Laboratory Systems 76*, 205– 214.

Cazelles, J., Robert, A., and Meunier, B. (2001). Alkylation of heme by artemisinin, an antimalarial drug. *C. R. Acad. Sci. Paris, Chimie / Chemistry () 4* 85–89.

Chang, J., Lei, B., Li, J., Li, S., Shen, Y., and Yao, X. (2008). Accurate and Validated Quantitative Structure–Activity Relationship Model of Caspase-mediated Apoptosis-inducing Activity of Phenolic Compounds Using Density Functional Theory Calculation and Genetic Algorithm–Multiple Linear Regression. *QSAR & Combinatorial Science, 27*(11-12), 1318-1325.

Chemical_Computing_Group_Inc. (2009). Molecular Operating Environment (MOE) (Version 2009.10). Montreal, Canada.

Chen, S.-D., Zeng, X.-L., Wang, Z.-Y., and Liu, H.-X. (2007). QSPR modeling of n-octanol/water partition coefficients and water solubility of PCDEs by the method of Cl substitution position. *Science of The Total Environment, 382*(1), 59-69.

Chiung-Sheue, K., Liu, C., Yang, S.-L., Roberts, M. F., Elford, B. C., and Philipson, J. D. (1992). Antimalarial activity of *Artemisia annua* flavanoids from whole plants and cell cultures. *Plant Cell Reports, 11*, 637-640.

Cho, S. J., and Hermsmeier, M. A. (2002). Genetic Algorithm Guided Selection: Variable Selection and Subset Selection. *J. Chem. Inf. Comput. Sci., 42*, 927-936.

Chonde, S., and Kumara, S. (2014). *Cheminformatics: An Introductory Review.* Paper presented at the Industrial and Systems Engineering Research Conference.

Cleydson Breno R. Santos, Vieira, J. B., Lobato, C. C., Hage-Melim, L. I. S., Souto, R. N. P., Lima, C. S., et al. (2014). A SAR and QSAR Study of New Artemisinin Compounds with Antimalarial Activity. *Molecules 19*, 367-399.

Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. [doi: 10.1021/ja00226a005]. *Journal of the American Chemical Society, 110*(18), 5959-5967.

Czermiński, R., Yasri, A., and Hartsough, D. (2001). Use of Support Vector Machine in Pattern Classification: Application to QSAR Studies. *Quantitative Structure-Activity Relationships, 20*(3), 227-240.

Daren, Z. (2001). QSPR studies of PCBs by the combination of genetic algorithms and PLS analysis. *Computers & Chemistry, 25*(2), 197-204.

Darnag, R., Minaoui, B., and Fakir, M. (2012). QSAR models for prediction study of HIV protease inhibitors Using Support Vector Machines, Neural Networks and Multiple Linear Regression. *Arabian Journal of Chemistry*(0).

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems, 50*(1), 1-18.

Debnath, A. K. (2001). Quantitative Structure-Activity Relationship (QSAR) Paradigm – Hansch Era to New Millennium. *Mini Reviews in Medicinal Chemistry, 1*, 187-195.

Derks, E. P. P. A., Westerhuis, J. A., Smilde, A. K., and King, B. M. (2003). An introduction to Multi-block Component Analysis by means of a flavor language case study. *Food Quality and Preference, 14*(5–6), 497-506.

Dhingra, V., Rao, V. K., and Narasu, L. M. (2000). Minireview: Current status of artemisinin and its derivatives as antimalarial drugs. *Life Sciences. , 66*(4), 279-300.

Dijksterhuis, G., Bom F., Michael, B., and Derek, V. (2002). Selection of a subset of variables: minimisation of Procrustes loss between a subset and the full set. *Food Quality and Preference, 13*(2), 89-97.

Dijksterhuis, G. B., and Gower, J. C. (1992). The interpretation of Generalized Procrustes Analysis and allied methods. *Food Quality and Preference, 3*(2), 67-87.

Dimitrov, S., Dimitrova, G., Pavlov, T., Dimitrova, N., Patlewicz, G., Niemela, J., et al. (2005). A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. [doi: 10.1021/ci0500381]. *Journal of Chemical Information and Modeling, 45*(4), 839-849.

Dixon, S. J., and Brereton, R. G. (2009). Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure. *Chemometrics and Intelligent Laboratory Systems, 95*(1), 1-17.

Domingo, J. L. (2006). Polychlorinated diphenyl ethers (PCDEs): Environmental levels, toxicity and human exposure: A review of the published literature. *Environment International, 32*(1), 121-127.

Doweyko, A. M. (2008). QSAR: dead or alive? *J Comput Aided Mol Des, 22*, 81–89.

Dunn III, W. J., and Wold, S. (1980). Relationships between chemical structure and biological activity modeled by SIMCA pattern recognition. *Bioorganic Chemistry, 9*(4), 505-523.

Dutilleul, P., Jason, D. S., Frigon, D., and Legendre, P. (2000). The Mantel Test versus Pearson's Correlation Analysis: Assessment of the Differences for Biological and Environmental Studies. *Journal of Agricultural, Biological, and Environmental Statistics, 5*(2), 131-150.

Eigenvector_Research_Inc. (2010). PLS_Toolbox. Washington, USA.

Eriksson, L., Andersson, P., Johansson, E., and Tysklind, M. (2006a). Megavariate analysis of environmental QSAR data. Part I – A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). *Molecular Diversity, 10*(2), 169-186.

Eriksson, L., Andersson, P., Johansson, E., and Tysklind, M. (2006b). Megavariate Analysis of Environmental QSAR Data. Part II – Investigating Very Complex Problem Formulations Using Hierarchical, Non-Linear and Batch-Wise Extensions of PCA and PLS. *Molecular Diversity*, *10*(2), 187-205.

Fatemi, M. H., and Gharaghani, S. (2007). A novel QSAR model for prediction of apoptosis-inducing activity of 4-aryl-4-H-chromenes based on support vector machine. *Bioorganic & Medicinal Chemistry, 15*(24), 7746-7754.

Fernandez Pierna, J. A., Volery, P., Besson, R., Baeten, V., and Dardenne, P. (2005). Classification of Modified Starches by Fourier Transform Infrared Spectroscopy Using Support Vector Machines. *Journal of Agricultural and Food Chemistry, 53*(17), 6581-6585.

Ferreira, J. E. V., Figueiredo, A. F., Barbosa, J. P., Cristino, M. G. G., Macedo, W. J. C., Silva, O. P. P., et al. (2010). A study of new antimalarial artemisinins through molecular modeling and multivariate analysis. *Journal of the Serbian Chemical Society, 75*(11), 1533–1548

Frank, I. E., and Kowalski, B. R. (1985). A multivariate method for relating groups of measurements connected by a causal pathway. *Analytica Chimica Acta, 167*(0), 51-63.

Fujita, T., and Ban, T. (1971). Structure-activity relation. 3. Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. [doi: 10.1021/jm00284a016]. *Journal of Medicinal Chemistry, 14*(2), 148-152.

Gallego-Álvarez, I., Rodríguez-Domínguez, L., and García-Rubio, R. (2013). Analysis of environmental issues worldwide: a study from the biplot perspective. *Journal of Cleaner Production, 42*(0), 19-30.

Gasteiger, J. (2006). The central role of chemoinformatics. *Chemometrics and Intelligent Laboratory Systems 82*, 200 – 209.

Gasteiger, J., and Engel, T. (2006). *Chemoinformatics: a textbook*. Germany: John Wiley & Sons.

Geladi, P., and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta, 185*(0), 1-17.

Golbraikh, A. (2000). Molecular Dataset Diversity Indices and Their Applications to Comparison of Chemical Databases and QSAR Analysis. *J. Chem. Inf. Comput. Sci., 40*, 414-425.

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H., and Tropsha, A. (2003). Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design 17*, 241–253.

Golbraikh, A., and Tropsha, A. (2002a). Beware of q^2! *Journal of Molecular Graphics and Modelling, 20*, 269–276.

Golbraikh, A., and Tropsha, A. (2002b). Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design, 16*, 357–369.

González Martín, Y., Luis Pérez Pavón, J., Moreno Cordero, B., and García Pinto, C. (1999). Classification of vegetable oils by linear discriminant analysis of Electronic Nose data. *Analytica Chimica Acta, 384*(1), 83-94.

Gonzalez, M. P., Teran, C., Saiz-Urra, L., and Teijeira, M. (2008). Variable Selection Methods in QSAR: An Overview. [doi:10.2174/156802608786786552]. *Current Topics in Medicinal Chemistry, 8*(18), 1606-1627.

Goodarzi, M., Saeys, W., de Araujo, M. C. U., Galvão, R. K. H., and Vander Heyden, Y. (2014). Binary classification of chalcone derivatives with LDA or KNN based on their antileishmanial activity and molecular descriptors selected using the Successive Projections Algorithm feature-selection technique. *European Journal of Pharmaceutical Sciences, 51*(0), 189-195.

Gower, J. C. (1975). Generalized Procrustes Analysis. *Psychometrika, 40*(1).

Gozalbes, R., Doucet, J. P., and Derouin, F. (2002). Application of Topological Descriptors in QSAR and Drug Design: History and New trends. *Current Drug Targets:Infect. Disord., 2,* 93-102.

Gramatica, P. (2007). Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science, 26*(5), 694-701.

Gramatica, P., Pilutti, P., and Papa, E. (2004). Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training−Test Sets and Consensus Modeling. *J. Chem. Inf. Comput. Sci., 44*(5), 1794-1802.

Guha, R., and Jurs, P. C. (2004a). Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *J. Chem. Inf. Comput. Sci., 44*(6), 2179-2189.

Guha, R., and Jurs, P. C. (2004b). Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues. *J. Chem. Inf. Comput. Sci., 44*(4), 1440-1449.

Guha, R., and Jurs, P. C. (2005). Determining the Validity of a QSAR Model − A Classification Approach. *J. Chem. Inf. Model., 45*(1), 65-73.

Gunn, S. R. (1998). Support Vector Machines for Classification and Regression. *in Online reference manual*.

Gupta, A. K., and Saxena, A. K. (2012). Triple-layered QSAR studies on substituted 1,2,4-trioxanes as potential antimalarial agents: superiority of the quantitative pharmacophore-based alignment over common substructure-based alignment. [doi: 10.1080/1062936X.2012.742136]. *SAR and QSAR in Environmental Research, 24*(2), 119-134.

Gupta, M. K., and Prabhakar, Y. S. (2006). Topological Descriptors in Modeling the Antimalarial Activity of 4-(3‘,5‘-Disubstituted anilino)quinolines†. [doi: 10.1021/ci0501140]. *Journal of Chemical Information and Modeling, 46*(1), 93-102.

Hansch, C., and Fujita, T. (1964). p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. [doi: 10.1021/ja01062a035]. *Journal of the American Chemical Society, 86*(8), 1616-1626.

Hasegawa, K., and Funatsu, K. (1998). GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model. *Journal of Molecular Structure (Theochem), 425,* 255-262.

Hasegawa, K., Miyashita, Y., and Funatsu, K. (1997). *J. Chem. Inf. Comput. Sci, 37*, 306.

Hasegawa, K., Miyashita, Y., and Funatsu, K. (1997). GA Strategy for Variable Selection in QSAR Studies: GA-Based PLS Analysis of Calcium Channel Antagonists⊥. [doi: 10.1021/ci960047x]. *Journal of Chemical Information and Computer Sciences, 37*(2), 306-310.

Helguera, A. M., Combes, R. D., Gonzalez, M. P., and Cordeiro, M. N. D. S. (2008). Applications of 2D Descriptors in Drug Design: A DRAGON Tale. [doi:10.2174/156802608786786598]. *Current Topics in Medicinal Chemistry, 8*(18), 1628-1655.

Hemmateenejad, B., Miri, R., Akhond, M., and Shamsipur, M. (2002). QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. An application of genetic algorithm for variable selection in MLR and PLS methods. *Chemometrics and Intelligent Laboratory Systems*, *64*, 91– 99.

Hewitt, M., Cronin, M. T. D., Madden, J. C., Rowe, P. H., Johnson, C., Obi, A., et al. (2007). Consensus QSAR Models: Do the Benefits Outweigh the Complexity? *J. Chem. Inf. Model., 47*(4), 1460-1468.

Hope, A. C. A. (1968). A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society. Series B (Methodological), 30*(3), 582-598.

Hopke, P. K. (2003). The evolution of chemometrics. *Analytica Chimica Acta, 500*(1–2), 365-377.

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics, 2*(3), 211-228.

Höskuldsson, A., and Svinning, K. (2006). Modelling of multi-block data. *Journal of Chemometrics, 20*(8-10), 376-385.

Hsu, C. W., Chang, C. C., and Lin, C. J. (2003). A Practical Guide to Support Vector Classication. *in Online reference manual*.

Huang, J., Yu, G., Yang, X., and Zhang, Z.-l. (2004). Predicting physico-chemical properties of polychlorinated diphenyl ethers (PCDEs): potential persistent organic pollutants (POPs). *Journal of Environmental Sciences, 16*(2), 204-207.

Hui-Ying, X., Jian-Wei, Z., Gui-Xiang, H., and Wei, W. (2010). QSPR/QSAR models for prediction of the physico-chemical properties and biological activity of polychlorinated diphenyl ethers (PCDEs). *Chemosphere, 80*(6), 665-670.

Jaafar, M. Z. (2011). *Chemometrics and Pattern Recognition Methods with Applications to Environmental and Quantitative Structure-Activity Relationship Studies.* University of Bristol.

Jaafar, M. Z., Khan, A. H., Adnan, S., Markwitz, A., Siddique, N., Waheed, S., et al. (2011). Multiblock analysis of environmental measurements: A case study of using Proton Induced X-ray Emission and meteorology dataset obtained from Islamabad Pakistan. *Chemometrics and Intelligent Laboratory Systems, 107*(1), 31-43.

Jagiello, K., Sosnowska, A., Walker, S., Haranczyk, M., Gajewicz, A., Kawai, T., et al. (2014). Direct QSPR: the most efficient way of predicting organic carbon/water partition coefficient (log K OC) for polyhalogenated POPs. *Structural Chemistry*, *25*(3), 997-1004.

Ji, L., Wang, X., Qin, L., Luo, S., and Wang, L. (2009). Predicting the Androgenicity of Structurally Diverse Compounds from Molecular Structure Using Different Classifiers. *QSAR & Combinatorial Science*, *28*(5), 542-550.

Johnson, M. A., and Maggiora, G. M. (1990). *Concepts and Applications of Molecular Similarity*. New York: Wiley.

Johnson, S. R. (2008). The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model. ,, , 48*, 25-26.

Josse, J., PagÃ¨s, J., and Husson, F. (2008). Testing the significance of the RV coefficient. *Computational Statistics & Data Analysis*, *53*(1), 82-91.

Jurs, P. C. (2003). Quantitative Structure-Property Relationships. In Gasteiger, J. (Ed.), *Handbook of Chemoinformatics* (Vol. 3, pp. 1321-1324). Weinheim: WILEY-VCH Verlag GmbH & Co.

Jurs, P. C., Dixon, S. L., and Egolf, L. M. (1995). Molecular Concepts. In van de Waterbeemd, H. (Ed.), *Chemometric Methods in Molecular Design.* . Germany: VCH.

Kamchonwongpaisan, S., and Meshnick, S. R. (1996). The Mode of Action of the Antimalarial Artemisinin and its Derivatives. *Gen. Pharmac. , 27*(4), 587-592.

Kandel, D., Raychaudhury, C., and Pal, D. (2014). Two new atom centered fragment descriptors and scoring function enhance classification of antibacterial activity. *Journal of Molecular Modeling, 20*(4), 1-13.

Katritzky, A. R., Karelson, M., and Petrukhin, R. (2005). Codessa. University of Florida: Codessa Pro.

Kaur, K., Jain, M., Kaur, T., and Jain, R. (2009). Antimalarials from nature. *Bioorganic & Medicinal Chemistry, 17*, 3229–3256.

Kennard, R. W., and Stone, L. A. (1969). Computer Aided Design of Experiments. [doi: 10.1080/00401706.1969.10490666]. *Technometrics, 11*(1), 137-148.

Kher, A., Mulholland, M., Green, E., and Reedy, B. (2006). Forensic classification of ballpoint pen inks using high performance liquid chromatography and infrared spectroscopy with principal components analysis and linear discriminant analysis. *Vibrational Spectroscopy, 40*(2), 270-277.

Kohonen, T. (2001). *Self-organizing maps* (3 ed.). Berlin: Springer-Verlag.

Kourti, T. (2003). Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *Journal of Chemometrics, 17*(1), 93-109.

Kowalski, B. R. (1981). Analytical chemistry as an information science. *TrAC Trends in Analytical Chemistry, 1*(3), 71-74.

Kurz, J., and Ballschmiter, K. (1994). Relationship between structure and retention of polychlorinated diphenyl ethers (PCDE) in HRGC in comparison with other groups of halogenated aromatic compounds *Fresenius J Anal Chem :, 349*, 533-537.

Kurz, J., and Ballschmiter, K. (1999). Vapour pressures, aqueous solubilities, Henry's law constants, partition coefficients between gas/water (Kgw), N-octanol/water (Kow) and gas/N-octanol (Kgo) of 106 polychlorinated diphenyl ethers (PCDE). *Chemosphere, 38*(3), 573-586.

Kutner, M. H., Nachtsheim, C. J., and Neter, J. (2004). *Applied Linear Regression Models* (4 ed.). New York: Mc-Graw Hill Irwin.

Lachenbruch, P. A. (2004). Discriminant Analysis *Encyclopedia of Statistical Sciences*: John Wiley & Sons, Inc.

Lavine, B. K. (2000). Chemometrics. *Analytical Chemistry, 72*, 91R-97R.

Lavine, B. K., and Davidson, C. E. (2003). Electronic van der Waals Surface Property Descriptors and Genetic Algorithms for Developing Structure-Activity Correlations in Olfactory Databases. *J. Chem. Inf. Comput. Sci., 43*, 1890-1905.

Leach, A. R. (2001). *Molecular Modelling: Principles and Applications*. England: Prentice Hall.

Leach, A. R., and Gillet, V. J. (2003). *An Introduction to Chemometrics*. Dordrecht: Kluwer Academic Publishers.

Leardi, R. (2001). Genetic algorithms in chemometrics and chemistry: a review. *Journal of Chemometrics, 15*(7), 559-569.

Leardi, R. (2007). Genetic algorithms in chemistry. *Journal of Chromatography A, 1158*, 226–233.

Leardi, R., and Gonzalez, A. L. (1998). Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemometrics and Intelligent Laboratory Systems, 41*, 195-207.

Leonard, J. T., and Roy, K. (2006). On Selection of Training and Test Sets for the Development of Predictive QSAR models. *QSAR & Combinatorial Science, 25*(3), 235-251.

Livingstone, D. J. (2000). The Characterization of Chemical Structures Using Molecular Properties. A Survey. *J. Chem. Inf. Comput. Sci., 40*(2), 195-209.

Lloyd, G. R., Brereton, R. G., Faria, R., and Duncan, J. C. (2007). Learning Vector Quantization for Multiclass Classification: Application to Characterization of Plastics. *J. Chem. Inf. Model. , 47*(4), 1553-1563.

Lopes, J. A., Menezes, J. C., Westerhuis, J. A., and Smilde, A. K. (2002). Multiblock PLS analysis of an industrial pharmaceutical process. *Biotechnology and Bioengineering, 80*(4), 419-427.

MacGregor, J. F., Jaeckle, C., Kiparissides, C., and Koutoudi, M. (1994). Process monitoring and diagnosis by multiblock PLS methods. *AIChE Journal, 40*(5), 826-838.

Maggiora, G., Vogt, M., Stumpfe, D., and Bajorath, J. (2014). Molecular Similarity in Medicinal Chemistry. [doi: 10.1021/jm401411z]. *Journal of Medicinal Chemistry, 57*(8), 3186-3204.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta), 2*, 49-55.

Mahmoudi, N., de Julián-Ortiz, J.-V., Ciceron, L., Gálvez, J., Mazier, D., Danis, M., et al. (2006). Identification of new antimalarial drugs by linear discriminant analysis and topological virtual screening. *Journal of Antimicrobial Chemotherapy, 57*(3), 489-497.

Marriott, F. H. C. (1979). Barnard's Monte Carlo Tests: How Many Simulations? *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 75-77.

Martens, H., and Naes, T. (1989). *Multivariate Calibration*. Chichester, UK: John Wiley & Sons, Ltd.

Martín, M. J., Pablos, F., and González, A. G. (1996). Application of pattern recognition to the discrimination of roasted coffees. *Analytica Chimica Acta, 320*(2–3), 191-197.

Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002). Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem., 45*(19), 4350-4358.

Massart, D. L., Vandeginste, B. G. M., Buysden, L., De Jong S., Lewi, P. J., and Smeyers-Verbeke, J. (1997). *Handbook of Chemometrics and Qualimetrics: Part A*. Amsterdam: Elsevier Science.

May, R. J., Maier, H. R., and Dandy, G. C. (2010). Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Networks, 23*(2), 283-294.

Meshnick, S. R. (2002). Artemisinin: mechanisms of action, resistance and toxicity. *International Journal for Parasitology 32*, 1655–1660.

Milanez, K. D. T. M., and Pontes, M. J. C. (2014). Classification of edible vegetable oil using digital image and pattern recognition techniques. *Microchemical Journal, 113*(0), 10-16.

Moonseong, H., and Ruben Gabriel, K. (1998). A permutation test of association between configurations by means of the rv coefficient. *Communications in Statistics - Simulation and Computation, 27*(3), 843-856.

Morley, C. (2006). OpenBabelGUI (Version 2.3.0): GNU General Public License.

National Cancer Institute (NCI) Database Retrieved 20 June 2011, from http://cactus.nci.nih.gov/ncidb2.1/

Nettles, J. H., Jenkins, J. L., Bender, A., Deng, Z., Davies, J. W., and Glick, M. (2006). Bridging Chemical and Biological Space: "Target Fishing" Using 2D and 3D Molecular Descriptors. [doi: 10.1021/jm060902w]. *Journal of Medicinal Chemistry, 49*(23), 6802-6810.

Nevalainen, T., and Kolehmainen, E. (1994). New qsar models for polyhalogenated aromatics. *Environmental Toxicology and Chemistry, 13*(10), 1699-1706.

Nikolova, N., and Jaworska, J. (2003). Approaches to Measure Chemical Similarity - a Review. *QSAR Comb. Sci., 22*, 1006-1026.

Norinder, U. (2003). Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimisations and variable selection. *Neurocomputing, 55*(1–2), 337-346.

Olliaro, P. L., Haynes, R. K., Meunier, B., and Yuthavong, Y. (2001). Possible modes of action of the artemisinin-type compounds. *TRENDS in Parasitology 17*(3), 122-126.

Oloff, S., Mailman, R. B., and Tropsha, A. (2005). Application of Validated QSAR Models of D1 Dopaminergic Antagonists for Database Mining. *J. Med. Chem., 48*(23), 7322-7332.

Oprea, T. I. (2002). On the information content of 2D and 3D descriptors for QSAR. *Journal of the Brazilian Chemical Society, 13*, 811-815.

Osuna, E., Freund, R., and Girosi, F. (1997, 17-19 Jun 1997). *Training support vector machines: an application to face detection.* Paper presented at the Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on.

Parvu, L. (2003). QSAR-a piece of drug design. *J. Cell. Mol. Med., 7*(3), 333-335.

Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., and Weinberger, L. E. (1996). Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. [doi: 10.1021/jm960290n]. *Journal of Medicinal Chemistry, 39*(16), 3049-3059.

Peres-Neto, P., and Jackson, D. (2001). How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia, 129*(2), 169-178.

Pinheiro, J. C., Ferreira, M. M. C., and Romero, O. A. S. (2001). Antimalarial activity of dihydroartemisinin derivatives against P. falciparum resistant to mefloquine: a quantum chemical and multivariate study. *Journal of Molecular Structure: THEOCHEM, 572*(1–3), 35-44.

Ploypradith, P. (2004). Development of artemisinin and its structurally simplified trioxane derivatives as antimalarial drugs. *Acta Tropica, 89*(3), 329-342.

Posner, G. H., Cumming, J. N., Ploypradith, P., and Oh, C. H. (1995). Evidence for Fe(IV)=O in the Molecular Mechanism of Action of the Trioxane Antimalarial Artemisinin. *J. Am. Chem. Soc, 117*, 5885-5886.

Posner, G. H., Park, S. B., Gonza´lez, L. s., Wang, D., Cumming, J. N., Klinedinst, D., et al. (1996). Evidence for the Importance of High-Valent FedO and of a Diketone in the Molecular Mechanism of Action of Antimalarial Trioxane Analogs of Artemisinin. *J. Am. Chem. Soc. , 118*, 3537-3538.

Puzyn, T., Mostrag-Szlichtyng, A., Gajewicz, A., Skrzyn, M., and Worth, A. P. (2011). Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. *Struct Chem :, 22*, 795–804.

Rajkhowa, S., Hussain, I., K. Hazarika, K., Sarmah, P., and Chandra Deka, R. (2013). Quantitative Structure-Activity Relationships of the Antimalarial Agent Artemisinin and Some of its Derivatives – A DFT Approach. *Combinatorial Chemistry & High Throughput Screening, 16*(8), 590-602.

Ramadan, Z., Mulholland, M., and Hibbert, D. B. (1998). Classification of detectors for ion chromatography using principal components regression and linear discriminant analysis. *Chemometrics and Intelligent Laboratory Systems, 40*(2), 165-174.

Randic, M. (2001). Novel Shape Descriptors for Molecular Graphs. *J. Chem. Inf. Comput. Sci. , 41*, 607-613.

Ren, Y., Liu, H., Xue, C., Yao, X., Liu, M., and Fan, B. (2006). Classification study of skin sensitizers based on support vector machine and linear discriminant analysis. *Analytica Chimica Acta, 572*(2), 272-282.

Ren, Y., Liu, H., Yao, X., and Liu, M. (2007). Prediction of ozone tropospheric degradation rate constants by projection pursuit regression. *Analytica Chimica Acta, 589*(1), 150-158.

Robert, P., and Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 25*(3), 257-265.

Sagrado, S., and Cronin, M. T. D. (2008). Application of the modelling power approach to variable subset selection for GA-PLS QSAR models. *Analytica Chimica Acta 609* 169–174.

Schlich, P., and Guichard, E. (1989). Selection and classification of volatile compounds of apricot using the RV coefficient. *Journal of Agricultural and Food Chemistry, 37*(1), 142-150.

Schmidbauer, O., and Tebelskis, J. (1992). *An LVQ based reference model for speaker-adaptive speech recognition.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing. .

Seasholtz, M. B., and Kowalski, B. R. (1992). The effect of mean centering on prediction in multivariate calibration. *Journal of Chemometrics, 6*(2), 103-111.

Seber, G. A. F., and Lee, A., J. . (2003). *Linear Regression Analysis* (2 ed.). New Jersey: John Wiley & Sons, Inc.

Servera-Francés, D., Arteaga-Moreno, F., Gil-Saura, I., and Gallarza, M. G. (2013). A multiblock PLS-based algorithm applied to a causal model in marketing. *Applied Stochastic Models in Business and Industry, 29*(3), 241-253.

Shen, M., Bguin, C., Golbraikh, A., Stables, J. P., Kohn, H., and Tropsha, A. (2004). Application of Predictive QSAR Models to Database Mining: Identification and Experimental Validation of Novel Anticonvulsant Compounds. *J. Med. Chem., 47*(9), 2356-2364.

Sheridan, R. P., and Kearsley, S. K. (2002). Why do we need so many chemical similarity search methods? *Drug Discovery Today.* , *7*(17), 903-911.

Shukla, K. L., Gund, T. M., and Meshnick, S. R. (1995). Molecular modeling studies of the artemisinin (qinghaosu)-hemin interaction: Docking between the antimalarial agent and its putative receptor. *Journal of Molecular Graphics, 13*, 215-222.

Singh, K. P., Malik, A., Sinha, S., and K., S. (2007). Multi-Block Data Modeling for Characterization of Soil Contamination: A Case Study. *Water, Air and Soil Pollution*(185), 79–93.

Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M., and van Erk, M. J. (2009). Matrix correlations for high-dimensional data: the modified RV-coefficient. *Bioinformatics, 25*(3), 401-405.

Smilde, A. K., Westerhuis, J. A., and de Jong, S. (2003). A framework for sequential multiblock component methods. *Journal of Chemometrics, 17*(6), 323-337.

Smola, A., and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199-222.

Snee, R. D. (1977). Validation of Regression Models: Methods and Examples. *Technometrics, 19*(4), 415-428.

Snelder, T. H., Dey, K. L., and Leathwick, J. R. (2007). A Procedure for Making Optimal Selection of Input Variables for Multivariate Environmental Classifications

Un Procedimiento para la Selección Óptima de Variables para Clasificaciones Ambientales Multivariadas. *Conservation Biology, 21*(2), 365-375.

Sosnowska, A., Barycki, M., Jagiello, K., Haranczyk, M., Gajewicz, A., Kawai, T., et al. (2014). Predicting enthalpy of vaporization for Persistent Organic Pollutants with Quantitative Structure–Property Relationship (QSPR) incorporating the influence of temperature on volatility. *Atmospheric Environment, 87*(0), 10-18.

Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. New York: John Wiley & Sons, Inc.

Srivastava, M. S., Singh, H., and Naik, P. K. (2009). Quantitative structure–activity relationship (QSAR) of artemisinin: the development of predictive in vivo antimalarial activity models. *Journal of Chemometrics, 23*, 618–635.

Stanimirova, I., Kubik, A., Walczak, B., and Einax, J. (2008). Discrimination of biofilm samples using pattern recognition techniques. *Analytical and Bioanalytical Chemistry, 390*(5), 1273-1282.

Struck, W., Siluk, D., Yumba-Mpanga, A., Markuszewski, M., Kaliszan, R., and Markuszewski, M. J. (2013). Liquid chromatography tandem mass spectrometry study of urinary nucleosides as potential cancer markers. *Journal of Chromatography A, 1283*, 122-131.

Stuper, A. J., and Jurs, P. C. (1990). ADAPT Program. Pennsylvania State University, USA: Jurs Research Group.

Sun, L., Zhou, L., Yu, Y., Lan, Y., and Li, Z. (2007). QSPR study of polychlorinated diphenyl ethers by molecular electronegativity distance vector (MEDV-4). *Chemosphere, 66*(6), 1039-1051.

Sutherland, J. J., and Weaver, D. F. (2003). Development of Quantitative Structure−Activity Relationships and Classification Models for Anticonvulsant Activity of Hydantoin Analogues. *J. Chem. Inf. Comput. Sci., 43*(3), 1028-1036.

Tang, H., Wang, X. S., Huang, X.-P., Roth, B. L., Butler, K. V., Kozikowski, A. P., et al. (2009). Novel Inhibitors of Human Histone Deacetylase (HDAC)

Identified by QSAR Modeling of Known Inhibitors, Virtual Screening, and Experimental Validation. *J. Chem. Inf. Model.*

The_Mathworks_Inc. (2008). MATLAB (Version 7.6.0.324 (R2008a)). Massachusetts, USA.

Todeschini, R., and Consonni, V. (2000). *Handbook of Molecular Descriptors*. Weinheim (Germany): Wiley-VCH.

Todeschini, R., Consonni, V., Mauri, A., and Pavan, M. (2006). DRAGON - Software for Molecular Descriptor Calculations (Version 5.4 for Windows). Milan, Italy: Talete srl.

Todeschini, R., Consonni, V., Mauri, A., and Pavan, M. (2010). DRAGON - Software for Molecular Descriptor Calculations (Version 6.0 for Windows). Milan, Italy: Talete srl.

Tong, S., and Koller, D. (2002). Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, *2*, 45-66.

Torrecilla, J. S., Rojo, E., Oliet, M., DomiÌ•nguez, J. C., and RodriÌ•guez, F. (2009). Self-Organizing Maps and Learning Vector Quantization Networks As Tools to Identify Vegetable Oils. *Journal of Agricultural and Food Chemistry, 57*(7), 2763-2769.

Tropsha, A. (2004). Application of Predictive QSAR Models to Database Mining. In Oprea, T. I. (Ed.), *Chemoinformatics in Drug Discovery* (pp. 437-455). Weinheim: WILEY-VCH Verlag GmbH &Co. KGaA.

Tropsha, A., Gramatica, P., and Gombar, V. K. (2003). The Importance of Being Earnest: Validation is the absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.*, *22*, 69-77.

van den Berg, R., Hoefsloot, H., Westerhuis, J., Smilde, A., and van der Werf, M. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics, 7*(1), 142.

van Velzen, E. J. J., Westerhuis, J. A., van Duynhoven, J. P. M., van Dorsten, F. A., Hoefsloot, H. C. J., Jacobs, D. M., et al. (2008). Multilevel Data Analysis of a Crossover Designed Human Nutritional Intervention Study. [doi: 10.1021/pr800145j]. *Journal of Proteome Research, 7*(10), 4483-4491.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Vilar, S., Santana, L., and Uriarte, E. (2006). Probabilistic Neural Network Model for the In Silico Evaluation of Anti-HIV Activity and Mechanism of Action. [doi: 10.1021/jm050932j]. *Journal of Medicinal Chemistry, 49*(3), 1118-1124.

Waller, C. L., and Bradley, M. P. (1999). Development and Validation of a Novel Variable Selection Technique with Application to Multidimensional Quantitative Structure−Activity Relationship Studies. [doi: 10.1021/ci980405r]. *Journal of Chemical Information and Computer Sciences, 39*(2), 345-355.

Waller, L. A., Smith, D., Childs, J. E., and Real, L. A. (2003). Monte Carlo assessments of goodness-of-fit for ecological simulation models.
. *Ecological Modelling 164*, 49–63.

Wangen, L. E., and Kowalski, B. R. (1988). A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics, 3*(1), 3-20.

Weaver, S., and Gleeson, M. P. (2008). The importance of the domain of applicability in QSAR modeling. *Journal of Molecular Graphics and Modelling 26*, 1315–1326.

Westerhuis, J. A., and Coenegracht, P. M. J. (1997). Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics, 11*(5), 379-392.

Westerhuis, J. A., Kourti, T., and Macgregor, J. F. (1998). Analysis of Multiblock and Hierarchical PCA and PLS Models. *Journal of Chemometrics 12*, 301–321.

Westerhuis, J. A., and Smilde, A. K. (2001). Deflation in multiblock PLS. *Journal of Chemometrics, 15*(5), 485-493.

Whitley, D. C., Ford, M. G., and Livingstone, D. J. (2000). Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. [doi: 10.1021/ci000384c]. *Journal of Chemical Information and Computer Sciences, 40*(5), 1160-1168.

Willett, P., Barnard, J. M., and Downs, G. M. (1998). Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci. , 38*(6), 983-996.

Wold, S. (1995). Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems, 30*(1), 109-115.

Wold, S., Eriksson, L., and Clementi, S. (2008). Statistical Validation of QSAR Results *Chemometric Methods in Molecular Design* (pp. 309-338): Wiley-VCH Verlag GmbH.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1â€“3), 37-52.

Wold, S., Kettaneh-Wold, N., and Skagerberg, B. (1989). Nonlinear PLS modeling. *Chemometrics and Intelligent Laboratory Systems, 7*(1–2), 53-65.

Wold, S., Kettaneh, N., and Tjessem, K. (1996). Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics, 10*(5-6), 463-482.

Wold, S., Sjostrom, M., and Eriksson, L. (2001a). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems 58*, 109–130.

Wold, S., Trygg, J., Berglund, A., and Antti, H. (2001b). Some recent developments in PLS modeling. *Chemometrics and Intelligent Laboratory Systems, 58*, 131–150.

Woolfrey, J., Avery, M., and Doweyko, A. (1998). Comparison of 3D quantitative structure-activity relationship methods: Analysis of the in vitro antimalarial activity of 154 artemisinin analogues by hypothetical active-site lattice and comparative molecular field analysis. *Journal of Computer-Aided Molecular Design, 12*(2), 165-181.

Wu, W., Mallet, Y., Walczak, B., Penninckx, W., Massart, D. L., Heuerding, S., et al. (1996). Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. *Analytica Chimica Acta, 329*(3), 257-265.

Xu, Y., Dixon, S., Brereton, R., Soini, H., Novotny, M., Trebesius, K., et al. (2007). Comparison of human axillary odour profiles obtained by gas chromatography/mass spectrometry and skin microbial profiles obtained by denaturing gradient gel electrophoresis using multivariate pattern recognition. *Metabolomics, 3*(4), 427-437.

Xu, Y., Zomer, S., and Brereton, R. G. (2006). Support Vector Machines: A Recent Method for Classification in Chemometrics. *Critical Reviews in Analytical Chemistry, 36*(3-4), 177-188.

Yang, P., Chen, J., Chen, S., Yuan, X., Schramm, K. W., and Kettrup, A. (2003). QSPR models for physicochemical properties of polychlorinated diphenyl ethers. *Science of The Total Environment, 305*(1â€“3), 65-76.

Yao, X. J., Panaye, A., Doucet, J. P., Zhang, R. S., Chen, H. F., Liu, M. C., et al. (2004). Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression. [doi: 10.1021/ci049965i]. *Journal of Chemical Information and Computer Sciences, 44*(4), 1257-1266.

Yasri, A., and Hartsough, D. (2001). Toward an Optimal Procedure for Variable Selection and QSAR Model Building. [doi: 10.1021/ci010291a]. *Journal of Chemical Information and Computer Sciences, 41*(5), 1218-1227.

Yueying Ren, Huanxiang Liu , Chunxia Xue, Xiaojun Yao, Mancang Liu, and Fan, B. (2006). Classification study of skin sensitizers based on support vector machine and linear discriminant analysis. *Analytica Chimica Acta 572*, 272–282.

Yueying Rena, Huanxiang Liu , Chunxia Xue, Xiaojun Yao, Mancang Liu, and Fan, B. (2006). Classification study of skin sensitizers based on support vector machine and linear discriminant analysis. *Analytica Chimica Acta 572*, 272–282.

Zarzo, M., and Ferrer, A. (2004). Batch process diagnosis: PLS with variable selection versus block-wise PCR. *Chemometrics and Intelligent Laboratory Systems, 73*(1), 15-27.

Zeng, X.-L., Wang, H.-J., and Wang, Y. (2012). QSPR models of n-octanol/water partition coefficients and aqueous solubility of halogenated methyl-phenyl ethers by DFT method. *Chemosphere, 86*(6), 619-625.

Zhang, S., Golbraikh, A., Oloff, S., Kohn, H., and Tropsha, A. (2006). A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. *J. Chem. Inf. Model., 46*(5), 1984-1995.

Zhang, S., Wei, L., Bastow, K., Zheng, W., Bross, A., Lee, K.-H., et al. (2007). Antitumor Agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. *J Comput Aided Mol Des, 21*, 97–112.