# Data Analytics – Computer Modelling of Metabolic Rates

SUNDAY TAREKAKPO ODOBAI[1], NAZIFI LAWAL BASHIR[2]
[1] *Niger Delta University*
[2] *Department of Petroleum Resources*

***Abstract-*** ***Artificial Neural Networks (ANNs) and Multiple Linear Regression (MLR) based Quantitative Structure-Activity Relationships (QSARs) models were developed to predict enzymatic activities, that is, the Michaelis-Menten constant (Km) and the maximum reaction rate (Vmax) for reactions involving the biotransformation of xenobiotics, catalysed by three classes of enzymes present in the mammalian livers. The enzymes we have studied here are alcohol dehydrogenase (ADH), aldehyde dehydrogenase (ALDH), and Flavin-containing monooxygenase (FMO). Data for enzymatic constants were collected from the literature and the computation of potential predictors was done for all xenobiotics to include for hundreds of molecular descriptors. The best predictor variables were selected (maximum of seven and a minimum of two descriptors) using the Microsoft excel correlation function for each enzyme class. Each dataset was divided into three sets, the divisions were training, cross-validation, and test sets in the ratio of 70%, 15%, and 15% respectively for both the ANNs and the MLR models to build the QSARs. The MATLAB programming language was employed to implement the writing and running of the learning algorithms. The predictive strengths of the models were assessed through the correlation of their predictions relative to the target outcomes for the three divisions and the mean square errors were computed, after fitting the resulting models with the entire dataset for each enzyme class. The ANNs model appeared best as it was seen to be relatively stable in performance through the training, cross-validation, and test sets of the data than the MLR model. For the prediction of Km, the most influential descriptors were partition coefficients and functional groups or fragments for compounds metabolised by ADH, ALDH, and FMO. Size, shape, symmetry, and atom distribution are those properties that mostly influenced the prediction of Vmax. This study is valuable in predicting Km and Vmax and for understanding the principles behind biotransformation by the liver enzymes; which in turn can be useful in taking proactive and remedial actions on issues regarding industrial activities affecting environmental wellbeing. It also finds relevance when guidance is needed for selecting an appropriate analytical model for a given dataset.***

***Indexed Terms-*** ***Machine Learning, Supervised Learning, Artificial Neural Network, Multiple Linear Regression, Quantitative Structure-Activity Relationships, Xenobiotic, Michaelis-Menten Constant.***

## I. INTRODUCTION

Metabolic activities in living organisms are responsible for the natural biotransformation of edibles, xenobiotic, poisonous substances, and medications which precede the consumption of substances that are useful to the biological systems and the removal of undesired or toxic substances from the systems; usually accompanied by the release of energy. The major organs that carry out metabolism in mammals are kidney, skin, liver, lung, gastrointestinal tract, and endothelial cells of the blood-brain barrier, with the primary ones being the liver, kidney, and intestines (BioFoundations, 2018). The liver carries out the following functions: ammonia filtration from the gastrointestinal tract drained blood, detoxification of endotoxins, filtration of other bacteria-derived substances, and xenobiotics filtration via the portal vein, glucose homeostasis, collecting and uptake of cholesterol, proteins assembly, and secretion of bile. The external origins of xenobiotic which are present in living organisms could result from human or natural actions which have direct or indirect effects on the natural ecosystem. In addition to components sourced from chemicals that could cause damages to the liver,

some naturally poisonous substances that can be found in the environment are peptides of *Amanita phalloides*, the *pyrrolizidine alkaloids*, and the toxin of the *cycad nut* (Ramaiah and Banerjee, 2015). As noted by Ramaiah and Banerjee in their research titled 'Liver Toxicity of Chemical Warfare Agents', mammals can also be contacted by toxic materials through other means such as unaware ingestion of mycotoxins through edibles that were contaminated as a result of environmental conditions that are beneficial to the growth of fungus and cyanobacterial polluted water. The liver cells possess the capability to stock up poisonous metals and extra vitamins that may result in toxic damage. Although, the notion of mammal renal UDP-glucuronosyltransferase (UGT) and cytochrome P450 (CYP) catalyst enzymes and the roles which they perform in the biotransformation of xenobiotics and endo-biotics are quite minute relative to liver-related metabolic actions on chemicals and drugs, evidence tell that the mammalian kidney possesses an excellent capacity for metabolic activities (Knights, K. *et al.*, 2013). The kidneys also, possess the ability to carry out prolonged red-ox, conjugation, and hydrolysis reactions (Lash, 1994). It is, therefore, pertinent to maintain a healthy level of enzymatic activities and to aid poor state of them in mammals and other existing organisms for the maintenance of the ecosystem. The preservation and improvement of such activities can be realised by effecting positive changes to consumables (food and drugs) to satisfy requirements. The available volume of data and the continuous expansion of the volume of the database make it necessary for insightful semi-analytical estimations leading to rational characterisation and description of trends in open data, which is vital for the purpose of decision-making. A big data can be defined as a collection of any dataset that is so large in volume and which needs a significant effort of processing via common programming devices that suppose that every information is available in memory (Dmitrij Martynenko, 2015). We may also define a big data as an object of human individual, and likewise a collected information which is generated and shared usually within the digital domain, where virtually everything can be measured and recorded by means of electronic devices and in so doing transformed into data (Sivarajah, *et al.*, 2016) – the process is also called 'datafication' (Mayer-Schönberger and Cukier, 2014). Accordingly, "data analytics -computer modelling of mammalian metabolism" can be said to be the analytics of a big data considering that these data were not originally present in the memory of the programming functions and require extraction from different sources. This primarily involves the modelling of mammalian cells metabolism using theoretical molecular descriptors as independent variables that characterise the structures and molecules of various substrate compounds to be metabolised by the mammalian enzymes. This method is the Quantitative-Structure Activity Relationships (QSARs), which is a widely applicable approach in metabolism studies based on analytical tools such as regression, decision trees, support vectors, discriminant analysis, etc.

In this work, we used the Artificial Neural Networks (ANNs) and Multiple Linear Regression (MLR) machine learning algorithms to model metabolism in mammalian tissues based on theoretical molecular descriptor features, using existing data. This was achieved by employing the convenience of the MATLAB programming language. The metabolism study here is that which concerns the biotransformation of the various xenobiotic in the environment, by some key enzymes in the mammalian livers.

## II. METHODOLOGY

The methods employed in this work are analytical and computational. The experimental data for the enzyme properties were obtained from the Braunschweig Enzyme Database (BRENDA) – an online experimental database and other reviews (Scheer, *et al.*, 2011; Hansch, *et al.*, 2004). The primary data which originated from the BRENDA database and other reviewed sources followed by thorough checks were collected from the supporting information of a publication (Pirovano, *et al.*, 2015). BRENDA is a comprehensive database which contains a plethora of experimental information about enzymes including those of metabolism (that is the Michalis-Menten constant, $K_m$ and maximum reaction rate, $V_{max}$) which are relevant for QSARs metabolism studies.

The theoretical molecular descriptors of the compounds metabolised by the various isoenzymes for

each of the three categories of enzymes considered were computed using the Online Chemical Modelling Environment (OCHEM) for descriptors such as WHIM, GETAWAY, 3D Morse, etc. Compounds were represented as SMILES (simplified molecular-input line-entry system) before the computation of the descriptors. The enzyme classes are Alcohol dehydrogenase (ADH), Aldehyde dehydrogenase (ALDH), and Flavin-containing monooxygenase (FMO); with each catalysing reaction for a combination of mammals (Human, pig, horse, rat, and mouse). OCHEM is a web-based platform which is a widely used platform that automatically computes a variety of descriptors employed for QSARs studies (Sushko, *et al*., 2011).

Correlations of the descriptors with the enzyme properties ($K_m$ and $V_{max}$) were done with Microsoft excel and descriptors with the best values of correlation coefficient were extracted to ensure reliable models.

The QSARs models were developed using Artificial Neural Networks (ANNs) and Multiple Linear Regression (MLR). The regression technique which is a widely applied statistical method employed for properties prediction and which finds relevance in many disciplines, had been used in xenobiotic metabolism prediction. The ANN offers an assuring model result, particularly for datasets with nonlinear relationships (Agatonovic-Kustrin and Beresford, 2000). ANNs are excellent pattern finding machine learning tools employed for too complicated or numerous patterns. The application of ANNs to predict metabolic activities on diverse xenobiotics in mammals is yet to be pronounced. Hence, this work seeks to exploit the predictability of the ANN algorithm in mammalian metabolic modelling with the intention of comparing its level of accuracy with that of MLR in this regard.

The chosen machine learning algorithms were written and run on the MATLAB programming language to develop the predictive models. MATLAB means Matrix Laboratory. It is a high-level programming language that directly expresses matrices and array mathematics and provides an environment for numerical calculations with suitable computation, visualisation, and other in-built tools (Chern, 2015). It is specially created for easy and fast scientific calculations, with many in-built functions and toolboxes that are applicable for researches in engineering, statistics, optimisation, partial differential equations, and data analytics (Gerritsen, 2006).

Finally, the qualities of the prediction tools were demonstrated using the root-mean-square errors and the correlation coefficients between measured and predicted outcomes.

The anticipated problem that later surfaced while carrying out this work was that of determining stable molecular descriptors: certainly, there was the need to explore numerous descriptors software before settling for stable descriptors with acceptable values of correlation with respect to the expected outcomes for precise predictions in the analysis. This was very tasking and time consuming. Most descriptors computed for this work had average correlations with the enzymatic constants. Nevertheless, this work can serve as a guide in further studies.

## III. SUMMARY OF INPUT DATA

Each of the datasets used for the models' input was divided into training, cross-validation, and test sets in the ratio of 70%, 15%, and 15% respectively for the analysis of both machine learning methods. The tables under this section give the summary of the entire input data used for developing the models.

### 3.1 Physical Interpretation of the Descriptors

For this QSARs study, the theoretical descriptor variables that were selected are presented in the table below:

Table 1 – Descriptors by group.

| Descriptor | Group |
|---|---|
| AlogPS_logP | CDK |
| AlogPS_logS | CDK |
| AlogP | CDK |
| XlogP | CDK |
| Autocorr2D | RDKIT |
| Autocorr3D | RDKIT |
| Morse | RDKIT |
| Apol | CDK |

| | |
|---|---|
| nAtom | CDK |
| SMR_VSA10 | RDKIT |
| AMR | CDK |
| Whim | RDKIT |
| Getaway | CDK |

These descriptors are selected based on their correlations with Log($1/K_m$) and Log($V_{max}$) and sometimes, with one another. Hence, the selected independent variables were found to have the best values of correlation coefficients with the target variables they were used to predict.

The AlogPS_logP, AlogPS_logS, AlogP, and XlogP are molecular hydrophobicity (lipophilicity) descriptors, with the P being the partition coefficient. They are used for estimating the hydrophobicity and pharmacokinetic properties of chemical compounds. The LogP is a measure of the molecular hydrophobicity, with P being the partition coefficient obtained from the distribution of a drug between two non-miscible solvents, mainly 1-octanol and water (Kujawski, *et al*., 2011). For ADH, AlogPS_logP is 0.6 correlated with Log($1/K_m$) and highly correlated with Autocorr2D8 and Morse129 (0.9 and 0.8 respectively). For ALDH, AlogPS_logP is 0.56 correlated with Log($1/K_m$), highly correlated with XLogP (R=0.94) and Apol (R=0.81). For FMO, AlogPS_logP is negatively correlated with AlogP (R= -0.77).

The two and three-dimensional autocorrelation (Autocorr2D and Autocorr3D) descriptors are size and shape, and functional or fragment descriptors respectively encoded with the relative positions of atoms or properties. They do so, by computing the separation, in terms of bond count (Autocorr2D) and Euclidean distance (Autocorr3D), between pairs of atoms (Sliwoski, *et al.*, 2015). The Autocorr3D21 was found to correlate poorly, with Log($V_{max}$).

The 3D-Morse quantifies the representation of molecular structures based on electron diffraction descriptors; the descriptors have a wide range of application, predominantly in QSARs studies. They contain information about the atomic mass, van der Waals volume, polarizability, electronegativity, and atomic partial charge of molecules (Devinyak, *et al*., 2014). For ADH, Morse129 had an average correlation with Log($1/K_m$) and a high correlation with Autocorr2D8 (R=0.95).

The Apol descriptor gives information about the sums of the polarizabilities (together with implicit hydrogen) of atoms (CCG). It was averagely correlated (R=0.54) with Log($1/K_m$).

The nAtom descriptor provides information about the number of atoms (including implicit hydrogen) in a molecule (CCG). It was averagely correlated (R=0.6) with Log($1/K_m$) and highly correlated (R=0.99) with Apol.

The SMR_VSA descriptor provides information on the refractivity of a molecule (including implicit hydrogen) together with the subdivided surface area based on the van der Waals surface area approximation (CCG).

The AMR is a molecular properties descriptor encoded with information on the Ghose-Crippen molar refractivity of molecules (DRAGON).

The Whim descriptor incorporates the entire information of the 3D, that is, size, shape, symmetry, and atom distribution as well as information on the electrostatic potential, hydrogen bonding capacity, and hydrophobicity of molecules (Bravi and Wikel, 2000). Getaway descriptors contain the information on the 3D structure and weights of the molecule atoms by their masses, that is, size and shape (Consonni, *et al.*, 2002).

### 3.2 Ranges of Actual Values of the Descriptors

The areas of application of the QSARs, which is in line with the Organisation for Economic Co-operation and Development (OECD), 2006 QSAR validation principles, are presented in ranges (minimum and maximum) of values of the theoretical molecular descriptors that were used to develop the models (Zvinavashe, *et al.*, 2008).

Table 2 – The data for log($1/K_m$) and the descriptors showing the range of values (minimum, maximum) for each enzyme class.

| Enzyme | Name | Range | Range (for MLR) Training | Cross-validation | Test |
|---|---|---|---|---|---|
| ADH | AlogPS_logP | (-1.52, 5.78) | (-1.52, 5.78) | (-1.52, 5.78) | (-1.52, 5.78) |
| | Autocorr2D8 | (0.54, 3.27) | (0.54, 2.90) | (0.54, 3.27) | (0.54, 2.90) |
| | Morse129 | (0.32, 15.2) | (0.32, 11) | (0.32, 15.2) | (0.32, 11) |
| | Log($1/K_m$) | (-6.48, 0) | (-6.48, 0) | (-5.18, -0.60) | (-5.34, -0.70) |
| ALDH | AlogPS_logP | (-2.69, 8.20) | (-2.69, 8.20) | (-0.69, 4.43) | (-1.01, 2.60) |
| | XlogP | (-0.70, 10.2) | (-0.70, 10.2) | (0.02, 4.57) | (-0.70, 2.86) |
| | Morse71 | (-4.34, 0.06) | (-4.34, 0.06) | (-1.01, 0.06) | (-0.95, 0.06) |
| | Apol | (3.90, 103) | (3.90, 103) | (3.90, 31.7) | (3.90, 29.9) |
| | nAtom | (4.0, 87.0) | (4.0, 87.0) | (4.0, 31.0) | (4.0, 26.0) |
| | Log($1/K_m$) | (-4.0, 3.40) | (-4.0, 3.40) | (-3.38, 0.70) | (-3.51, 1.0) |
| FMO | AlogPS_logS | (-8.45, 1.21) | (-8.45, 1.21) | (-5.85, 1.04) | (-5.64, 1.04) |
| | AlogP | (-2.07, 5.05) | (-2.07, 5.05) | (-0.77, 3.97) | (-2.07, 5.05) |
| | SMR_VSA10 | (0.0, 45.20) | (0.0, 44.6) | (0.0, 45.2) | (0.0, 40.6) |
| | Log($1/K_m$) | (-4.60, -0.04) | (-4.60, -0.04) | (-3.88, -0.30) | (-3.90, -0.15) |

Table 3 – The data for log($V_{max}$) and the descriptors showing the range of values (minimum, maximum) for each enzyme class.

| Enzyme | Name | Range | Range (for MLR) Training | Cross-validation | Test |
|---|---|---|---|---|---|
| ADH | Getaway255 | (0.35, 19.9) | (0.35, 19.9) | (0.35, 19.9) | (0.35, 16.7) |
| | Getaway264 | (0.22, 17.4) | (0.22, 17.4) | (0.22, 17.4) | (0.22, 13.1) |
| | Log($V_{max}$) | (-2.0, 1.94) | (-2.0, 1.94) | (-0.82, 1.93) | (-1.05, 0.74) |
| ALDH | Morse203 | (0.41, 19.1) | (0.41, 19.1) | (0.41, 10.1) | (0.41, 9.07) |
| | Whim8 | (0.42, 1.0) | (0.42, 1.0) | (0.43, 1.0) | (0.46, 1.0) |
| | Log($V_{max}$) | (-2.0, 1.23) | (-2.0, 1.23) | (-2.0, 0.997) | (-1.70, 0.23) |
| FMO | Whim1 | (0.13, 7.06) | (0.13, 5.93) | (0.16, 7.06) | (1.61, 5.84) |
| | Whim3 | (0.49, 0.96) | (0.49, 0.96) | (0.49, 0.93) | (0.49, 0.93) |
| | Whim4 | (0.04, 0.49) | (0.04, 0.49) | (0.05, 0.49) | (0.05, 0.49) |
| | Whim25 | (0.48, 0.96) | (0.48, 0.96) | (0.48, 0.93) | (0.49, 0.93) |
| | Whim26 | (0.04, 0.49) | (0.04, 0.49) | (0.05, 0.49) | (0.05, 0.49) |
| | Log($V_{max}$) | (-1.52, 0.40) | (-1.52, 0.25) | (-0.70, 0.40) | (-0.92, 0.37) |

## IV. MODEL DEVELOPMENT

The QSARs models were built with the ANN and the MLR. For both models (ANN and MLR), all datasets were divided into training sets – those used for estimating the model parameters, cross-validation sets, and test sets. After testing the model, the estimated parameters were finally used to fit the whole dataset to estimate the model performance on the entire dataset.

4.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are based on the following sigmoid models:

$$h_\theta(x) = P(y = 1|x; \theta) = \frac{1}{1+e^{-z}} \quad (1)$$

Where $0 \leq h_\theta(x) \leq 1$.

$$z = x\theta^T \quad (2)$$

Basically, the threshold is 0.5 but in practice it is usually being raised to ensure reasonable level of certainty.

A scaled form of the sigmoid function is the hyperbolic tangent function which have an output range of -1 to +1, with a basic threshold of 0.

$$f(x) = \tanh(x) = \frac{2}{1+e^{-x}} - 1 \quad (3)$$

The models have been intensely studied and they are very popular learning techniques among others in *in-silico* modelling. ANNs have been utilised in medicinal chemistry for classifying compounds, QSARs modelling, primary virtual screening of compounds, identification of potential drug targets, and localisation of structural and functional characteristics of biopolymers (Patel and Chaudhari, 2005). ANN techniques have also been applied in the fields of robotics, pattern recognition, psychology, physics, computer science, biology, and others (Fogel, 2008).

ANN came up in an attempt to simulate the structure and function of the human brain. Nevertheless, besides any neurological interpretation, they can be considered as a class of general, flexible, nonlinear regression models (Haykin, 1999). The network is made up of simple units, known as neurons, arranged in a certain topology, and connected to each other. Neurons are organized into layers. A typical Network comprises of an input layer and one output layer, with a single or more hidden layers. The accuracy of an ANN increase as the number of hidden layers and hidden neurons increases, likewise the cost of computation. An ANN in which the neurons are connected only to those in the preceding layers are called the feedforward networks, this group contains multiplayer perceptron (MLP), radial basis function (RBF) networks, and Kohonen's self-organizing maps (Kohonen's SOM). Conversely, if recursive connections exist between neurons in different layers, it is a feed-back network. The forward propagation computes the activation functions of the hidden units and the output, while the back-propagation algorithm computes the cost function of a neural network with respect to the weights or the fitting parameters. A simple Network with two hidden layers is shown below:

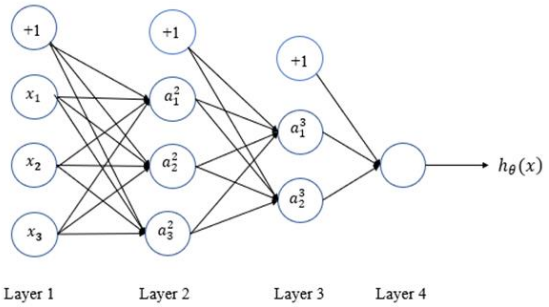


Figure 1 – A simple neural networks.

A neuron consists of a linear activator followed by a nonlinear inhibiting function. The activation function computes the sigmoid of the sums of the products of its input data and the parameters plus that of an independent term from a bias unit with an input of +1. The signal level of the sum is captured by the nonlinear retarding function. The most familiar activation hypotheses are the hyperbolic tangent, step, and sigmoid functions. The act of improving the parameters of fit with available data is known as "training of the network" and the data used for this purpose, the 'training dataset'. The algorithm mostly used for the network training is the back-propagation which is essentially a gradient descent method that minimises the computational cost function (the mean square error), it basically minimises the mean square error difference between the model outcomes and the target values of the training dataset to arrive at the parameters of best fit.

A common problem with ANNs in predictive analytics is that the classification models produced are not always interpretable physically or chemically, this issue is usually called the 'black box' nature of ANNs. However, the main benefit of ANNs is the capacity to arrest and simulate nonlinear trends in data (Lavecchia, 2015).

Considering the four-layered network illustrated in figure 1: The network consists of three input units representing the features (independent variables), two hidden layers with the first hidden layer having three

hidden units and the second hidden layer having two hidden units, and one output unit.

Each node or neuron consists of a linear activation function, which is basically a sigmoid function, followed by a nonlinear inhibiting function.

4.1.1 Feed Forward Propagation Model

The feed forward propagation accomplishes the computation of the linear activation functions which is essentially the sigmoid of the sums of the products of its input data and the parameters, plus that of an independent term from a bias unit with an input of plus one for the hidden units and the expected outcomes, as the nonlinear inhibiting function attempts to arrest the signal level of the sum having trained a network.

For the network in figure 1, the activation functions of the hidden layers are computed as follows:

$$a_1^2 = g(\theta_{10}^1 x_0 + \theta_{11}^1 x_1 + \theta_{12}^1 x_2 + \theta_{13}^1 x_3) \quad (4)$$
$$a_2^2 = g(\theta_{20}^1 x_0 + \theta_{21}^1 x_1 + \theta_{22}^1 x_2 + \theta_{23}^1 x_3) \quad (5)$$
$$a_3^2 = g(\theta_{30}^1 x_0 + \theta_{31}^1 x_1 + \theta_{32}^1 x_2 + \theta_{33}^1 x_3) \quad (6)$$
$$a_1^3 = g(\theta_{10}^2 x_0 + \theta_{11}^2 x_1 + \theta_{12}^2 x_2 + \theta_{13}^2 x_3) \quad (7)$$
$$a_2^3 = g(\theta_{20}^2 x_0 + \theta_{21}^2 x_1 + \theta_{22}^2 x_2 + \theta_{23}^2 x_3) \quad (8)$$

The output function *h(x)*, is given by:

$$h_\theta(x) = a_1^4 = g(\theta_{10}^3 a_0^3 + \theta_{11}^3 a_1^3 + \theta_{12}^3 a_2^3) \quad (9)$$

Where $a_i^j$ is the activation function of unit *i* in layer *j* and $\theta^j$ is the matrix of the parameters controlling function mapping from layer *j* to layer *j*+1, $x_0$ is 1 and $a_0^3$ is 1.

For the hidden units $\theta^1, \theta^2 \in \mathbb{R}^{3\times4}$ ,while for the output unit $\theta^3 \in \mathbb{R}^{1\times4}$.

4.1.2 Back Propagation Model

The back propagation algorithm computes the cost function of a neural network. The algorithm does the training of the network by adjusting the parameters to find the parameters which best fit the training dataset. The adjustment is done through an iterative gradient descent process to minimise the computation cost (the squared error function).

The cost function of a neural network is given as:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} log h_\theta(x^i)_k + (1 - y_k^{(i)}) \log\left(1 - h_\theta(x^i)\right)_k\right] + \frac{\lambda}{2m}\sum_{i=1}^{m}\sum_{k=1}^{s_{l+1}}\left(\theta_{ik}^{(l)}\right)^2 \quad (10)$$

Where $\boldsymbol{\lambda}$ is the regularisation parameter, *m* is the length of the training set, *k* is the number of units in a given layer, and *l* denotes the number of layers.

The gradient functions are computed by back propagation alongside the parameters of fit to obtain those parameters which give the minimum computation cost. The parameters can also be arrived at by some advanced optimisation algorithm such as FMINUNC and FMINCG.

Let $\delta_i^l$ be the deviation of a prediction at node *k* and layer *l* from a target value, considering a four-layer network like that of figure 2.4, we have the following:

$$\delta_k^4 = a_k^4 - y_k \quad (11)$$
$$\delta^3 = (\theta^3)^T \delta^4 .* g'(z^3) \quad (12)$$
$$\delta^2 = (\theta^2)^T \delta^3 .* g'(z^2) \quad (13)$$

Where

$$g'(z^l) = a^l .* (1 - a^l) \quad (14)$$

At every *i* training examples, the gradient is computed as:

$$\frac{\partial}{\partial \theta_{ik}^l} J(\theta) = \frac{1}{m}(a_k^l \delta_i^{l+1} + \lambda \theta_{ik}^l) \quad (15)$$

And the *a'*s are the activations earlier computed for the nodes in the layers other than the input layer using feed forward propagation.

As part of debugging, gradient checking is usually done by computing the numerical estimates of the gradients using the function:

$$\frac{d}{d(\theta)} J(\theta) \approx \frac{J(\theta+\varepsilon) - J(\theta-\varepsilon)}{2\varepsilon} \quad (16)$$

Where $\varepsilon$ is a very small value of about $10^{-4}$.

A simple three-layer (with one hidden layer) network is sufficient to train a neural network. But for this study, the networks are trained with five layers (having three hidden layers) with each of the hidden layers having thirty hidden units to increase the level of certainty.

4.2 Multiple Linear Regression Model

The linear regression analysis is a statistical approach which is performed to predict the values of a target variable, *y*, given some predictor variables ($x_1$, $x_2$, …., x*n)*. This method of analysis is employed in QSARs modelling of the relationship between one or more molecular descriptors (independent variables or features) and a continuous outcome/target (dependent variable). In metabolism modelling, this outcome can be the metabolic rate ($V_{max}$) or the affinity between an enzyme and a substrate ($K_m$). A linear regression model could be a simple linear equation, equation with multiple independent variables or a polynomial function.

The multiple linear regression hypothesis is expressed as follows:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots + \theta_n x_n \quad (17)$$

Where $h_\theta(x)$ is the dependent variable which represents the predicted biological activities, that is, the Michaelis-Menten constant ($K_m$) and the maximum reaction rate ($V_{max}$) that we predicted for the enzymatic activities of the four classes of enzymes, $x_1$, $x_2$, $x_3$, $x_4$,. . . $x_n$ are the features representing the theoretical molecular descriptor values, and $\theta_0$, $\theta_1$, $\theta_2$, $\theta_3$, $\theta_4$, . . . $\theta_n$ are the parameters of best fit which are to be learnt with the training set of each dataset. In this study, these parameters were determined using the method of Least Squares.

In matrix form, (17) is expressed as follows:

$$h_\theta(x) = \begin{pmatrix} h_{\theta 0} \\ h_{\theta 1} \\ h_{\theta 2} \\ h_{\theta 3} \\ \vdots \\ h_{\theta n} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & x_{31} \ldots x_{n1} \\ 1 & x_{12} & x_{22} & x_{32} \ldots x_{n2} \\ 1 & x_{13} & x_{23} & x_{33} \ldots x_{n3} \\ 1 & x_{14} & x_{24} & x_{34} \ldots x_{n4} \\ \vdots & \vdots & \vdots & \vdots \ddots \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} \ldots x_{nn} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \vdots \\ \theta_n \end{pmatrix} \quad (18)$$

The objective function which is a function of the loss or the difference between the model outcome $h_\theta(x^i)$ and the measured dependent variable $y^i$ is given by the squared error function:

$$L = \sum_{i=1}^{m} \left( y^i - h_\theta(x^i) \right)^2 \quad (19)$$

The parameters can be estimated using the method of least squares, with the intention of minimising the objective:

$$L = (y - X\,\theta)^T (y - X\,\theta) \quad (20)$$

By expansion

$$L = y^T y - 2\,\theta^T X^T y + \theta^T X^T X \theta \quad (21)$$

The minimum value of $L$ is obtained when $\partial L / \partial \theta = 0$, as such:

$$\frac{\partial L}{\partial \theta} = -2\,X^T y + 2 X^T X \theta = 0 \quad (22)$$

$$(X^T X)\theta = X^T y \quad (23)$$

Therefore,

$$\theta = (X^T X)^{-1} X^T y \quad (24)$$

The gradient descent approach was further used to double-check the estimated parameters.

## V. RESULTS

The most influential descriptors (that is, those with the highest values of correlation with $Log(1/K_m)$) for ADH were ALogPS_logP and Autocorr2D8, that is, partition coefficient and functional group or fragment respectively, with all having positive correlations. The most influential descriptors for ALDH were ALogPS_logP and XLogP having positive correlation coefficients. The best descriptor for FMO was ALogP with positive correlation.

For the QSAR modelling of $Log(V_{max})$ prediction, the best-correlated descriptors are Getaway264, Whim8, and Whim1, for ADH, ALDH, and FMO respectively. For each model predictions of $Log(1/K_m)$ and $Log(V_{max})$ for the various enzyme, the models' performances (the root mean square deviations and the Pearson's correlation coefficients) were recorded. The variable $1/K_m$ is a reflection of the enzyme affinity for substrate: a high $K_m$ suggests a low binding affinity. The correlation coefficient (R) and the root-mean-

square error (RMSE) revealed the performances of the models on each dataset, showing the relationships between the models' outcomes and measured values. The performance on the test datasets is of concern here because, those tell how well the models will perform on an unseen data, although consistency matters still. The RMSE is only presented for models' predictions when fitted with the whole datasets.

Detailed necessary discussions on the results obtained from the learning algorithms are as follows:

### 5.1 ANNs Results

The five-layer network (three hidden layers, each having thirty hidden units) was trained for each of the dataset over ten random division of the datasets for the prediction of Log($1/K_m$) and Log ($V_{max}$) and performances were averaged as presented in the following subsections:

#### 5.1.1 ANN Prediction of Michaelis-Menten Constant ($K_m$)

In all, the average performances of the model on each division showed insignificant differences. The model's prediction strengths (R) were about 64% for ADH and ALDH, and 54% for FMO. The model was seen to produce good and consistent correlations for all enzyme classes as shown by the performance plots in figures 2 – 4 below.

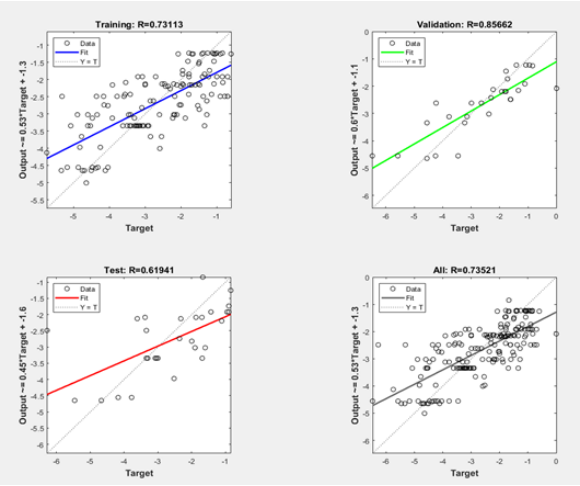


Figure 2 – ANN prediction of Log($1/K_m$) plot for ADH.

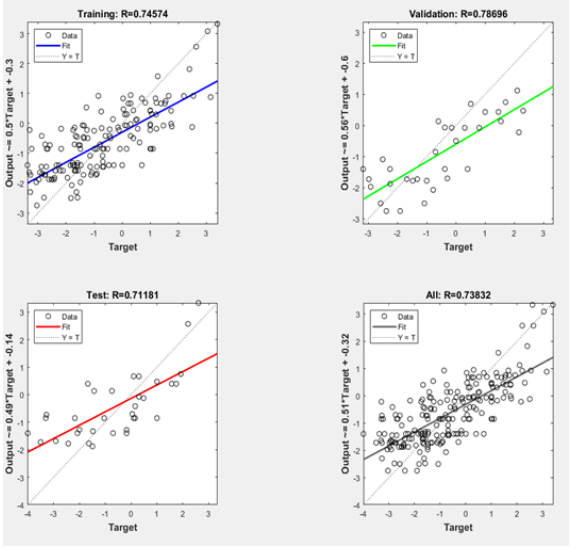


Figure 3 – ANN prediction of Log($1/K_m$) for ALDH.

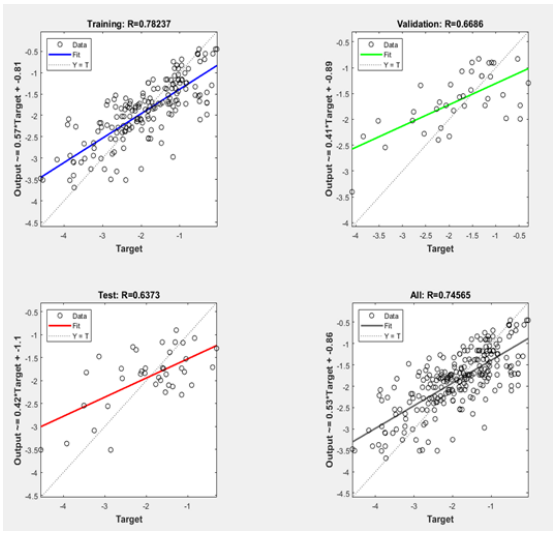


Figure 4 – ANN prediction of Log($1/K_m$) for FMO.

#### 5.1.2 ANN Prediction of Maximum Reaction Rate ($V_{max}$)

The average performances (R) of the model on the test sets for ADH and ALDH appear to be of insignificant differences relative to the performances on the training sets. But the said performances show substantial differences, as well as poor results in the case of FMO which can be easily traced to the correlation between the molecular descriptor values and Log($V_{max}$). The correlation for ADH is about 50%, about 40% for ALDH, and 11% for FMO. This showed fairly consistent correlation results for ADH and ALDH, but poor for FMO because of the inherent poor correlation

between the features and the target variables as can be seen in table 6.

Figures 5 – 7 below show the various performance plots for Log($V_{max}$) prediction.

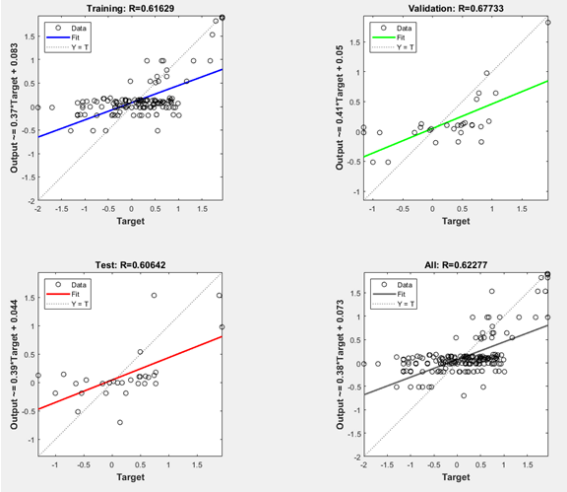


Figure 5 – ANN prediction of Log($V_{max}$) for ADH.

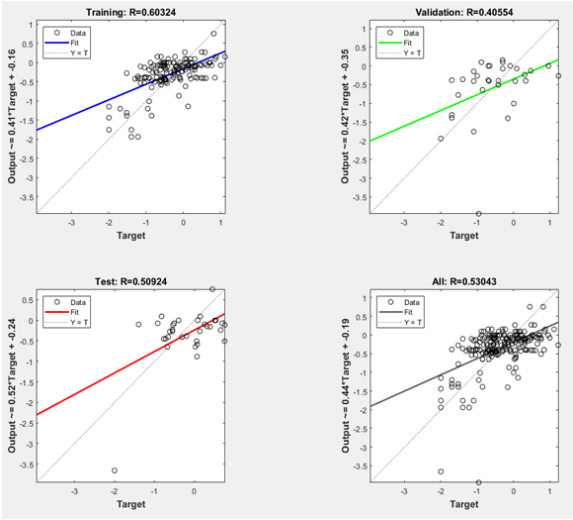


Figure 6 – ANN prediction of Log($V_{max}$) for ALDH.

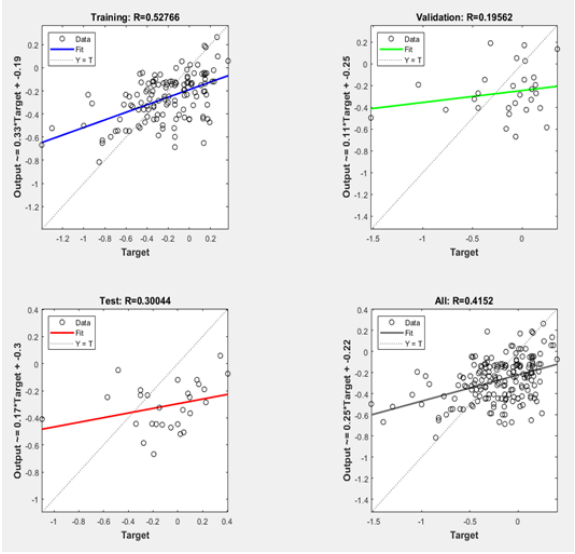


Figure 7 – ANN prediction of Log($V_{max}$) for FMO.

## 5.2 MLR Results

The performance of the MLR predictive model for the four classes of enzymes was examined using the same data that were used to check for performances on the ANN prediction model. This means that those descriptors that mostly influence the prediction of Log($1/K_m$) and Log($V_{max}$) in the ANN model, that is, partition coefficient and functional group for Log($1/K_m$) and size, shape, symmetry, and atom distribution for Log($V_{max}$), were still valid. Although the MLR model was run once on each of the datasets, performances in most cases appear to be lower than the worse in the case of the ANN model. The MLR model results are summarised in detail as follows:

### 5.2.1 MLR Prediction of Michaelis-Menten Constant ($K_m$)

For the MLR model, the best performance (R) was seen on the ADH dataset, but significant variations in the training, cross-validation, and test results, as well as lower R values, were observed. The test performance was about 66% for ADH, 40% for ALDH and FMO. Evidence of overfitting and underfitting, however, appear significant as shown in table 5; in which correlations appear fairly consistent for ADH but not for ALDH and FMO.

The estimated training parameters computed by the method of least squares are as follows:

ADH: $\theta = (-4.827, 1.585, 5.621, -4.442)^T$

ALDH: $\theta = (-0.895, 0.511, 0.361, -0.853, -0.153, 0.1132)^T$

FMO: $\theta = (-2.671, -0.095, 0.139, 0.019)^T$

The model's performance plots are presented in figures 8 – 10 below.

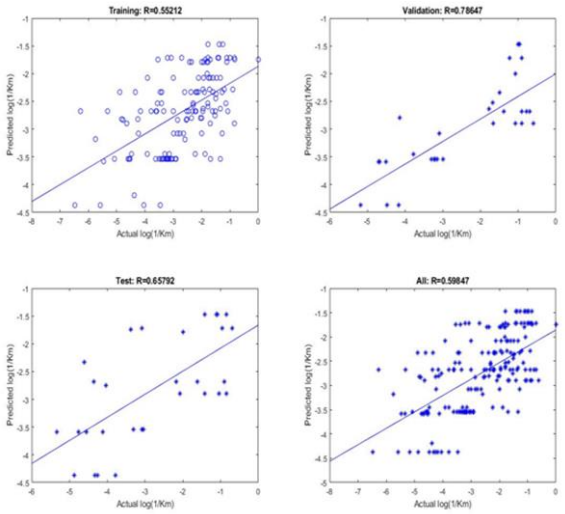


Figure 8 – MLR prediction of Log(1/ $K_m$) for ADH.

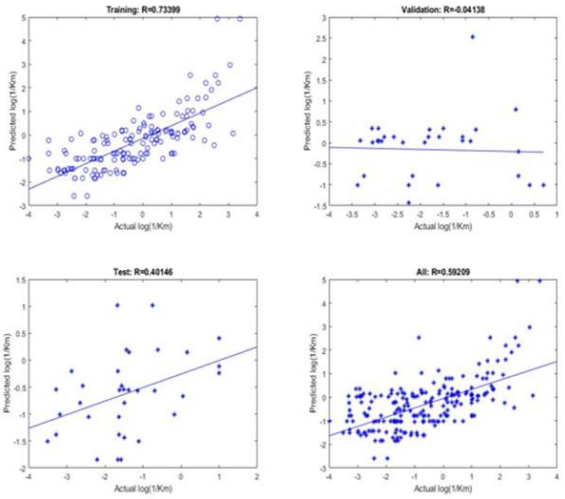


Figure 9 – MLR prediction of Log(1/ $K_m$) for ALDH.

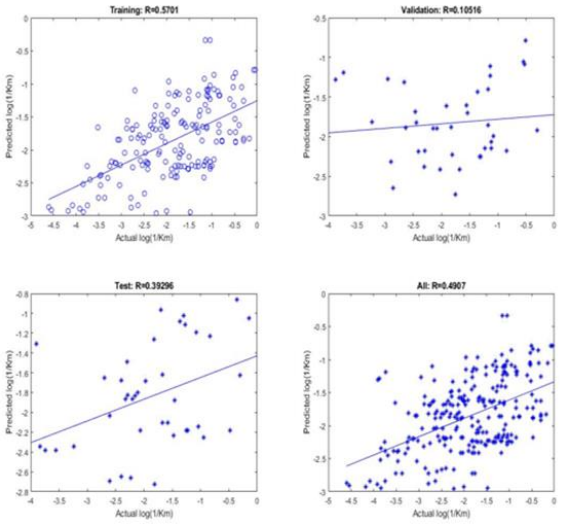


Figure 10 – MLR prediction of Log(1/ $K_m$) for FMO.

5.2.2 MLR Prediction of Maximum Reaction Rate ($V_{max}$)

Although the model performance (R) appears favourable in some instances, it fails the test of generalisation due to clear cases of overfitting and underfitting as revealed by the plots; generally showing inconsistent correlation values for all the enzyme classes. Test sets performance were about 27% for ADH, 35% for ALDH, and 47% for FMO.

The learning parameters computed for Log($V_{max}$) prediction by the method of least squares are given as follows:

ADH: $\theta = (-0.012, 0.005, 0.070)^T$

ALDH: $\theta = (-0.018, -0.028, -0.154)^T$

FMO: $\theta = (-1.636, 13.332, 13.278, 0.047, -11.663, -11.909)^T$

The Log($V_{max}$) prediction performance plots for the MLR model are presented in figures 11 - 13 as shown:

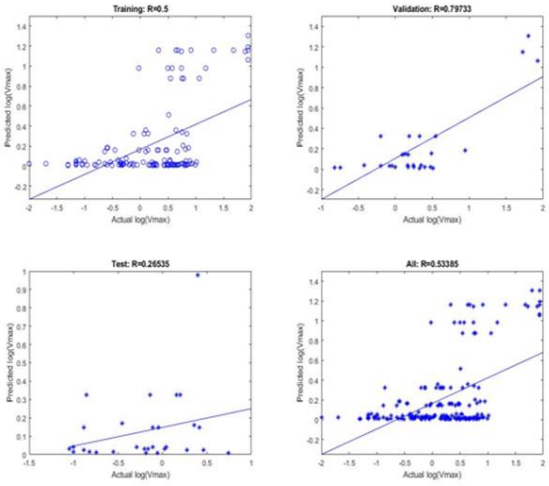


Figure 11 – MLR prediction of Log($V_{max}$) for ADH.

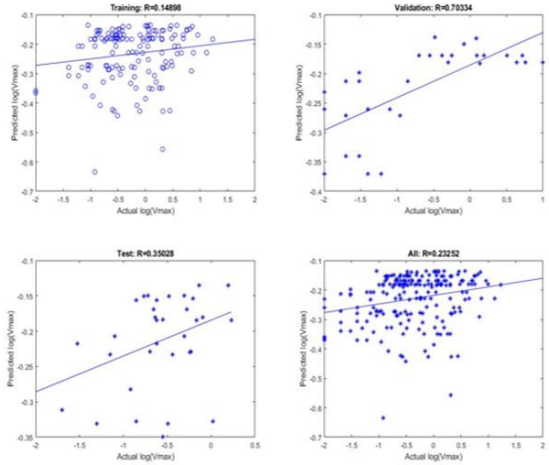


Figure 12 – MLR prediction of Log($V_{max}$) for ALDH.

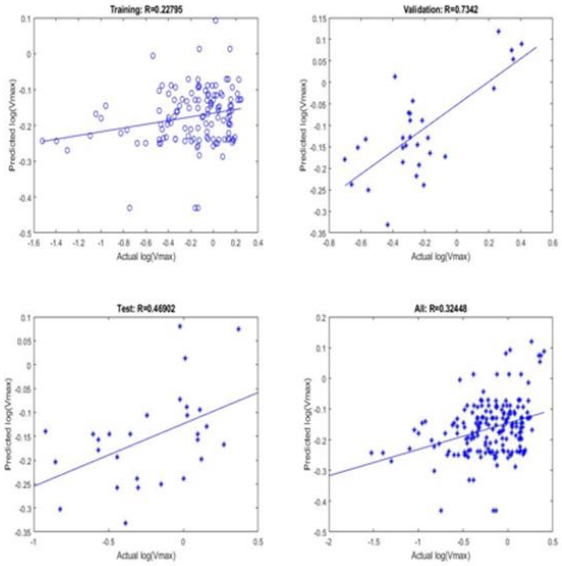


Figure 13 – MLR prediction of Log($V_{max}$) for FMO.

## VI. SUMMARY

Having developed the ANN and MLR based QSARs models to predict metabolism in mammals for the three enzyme classes, we have been able to achieve the aim of this study. First, the performance (R) of the models on the datasets appeared to be in order when compared to those reported in the literature especially for the Log($K_m$) prediction in which most of the correlation values meet the threshold (absolute 0.4). For Log($1/K_m$) prediction, both ANN and MLR have performance (R) on the test datasets in the following decreasing order: ADH, ALDH, and FMO. For the Log($V_{max}$) prediction in which most of the descriptors did not meet the threshold requirement, the ANN model still followed the order of performance but the MLR model did not. The tables below present the performances of the models on the datasets for the various enzyme classes, with RMSE presented for the whole datasets:

The QSARs model results obtained for the prediction of Log($1/K_m$) are summarised as follows:

Table 4 – ANN model's average performances for Log($1/K_m$) prediction.

| Enzyme | R | | | RMSE |
|---|---|---|---|---|
| | Training | Cross-validation | Test | All |
| ADH | 0.7364 | 0.6719 | 0.6414 | 0.7093 |
| 0.9593 | | | | |
| ALDH | 0.7373 | 0.7079 | 0.6360 | 0.7208 |
| 1.1397 | | | | |
| FMO | 0.7792 | 0.4979 | 0.5445 | 0.6943 |
| 0.7262 | | | | |

Table 5 – MLR model's performances for Log($1/ K_m$) prediction.

| Enzyme | R | | | RMSE |
|---|---|---|---|---|
| | Training | Cross-validation | Test | All |
| ADH | 0.5521 | 0.7865 | 0.6579 | 0.5985 |
| 1.0832 | | | | |
| ALDH | 0.7340 | -0.0414 | 0.4015 | 0.5921 |
| 1.3469 | | | | |
| FMO | 0.5701 | 0.1052 | 0.3930 | |
| 0.4907 | 0.8483 | | | |

The QSARs model results obtained for the prediction of Log($V_{max}$) are summarised as follows:

Table 6 – ANN model's average performances for Log($V_{max}$) prediction.

| Enzyme | R | | | RMSE |
|---|---|---|---|---|
| | Training | Cross-validation | Test | All |
| ADH | 0.6184 | 0.5332 | 0.4723 | 0.5597 |
| 0.6550 | | | | |
| ALDH | 0.5921 | 0.3061 | 0.4125 | 0.5185 |
| 0.5897 | | | | |
| FMO | 0.5488 | 0.1120 | 0.1140 | 0.3857 |
| 0.3201 | | | | |

Table 7 – MLR model's performances for Log($V_{max}$) prediction.

| Enzyme | R | | | RMSE |
|---|---|---|---|---|
| | Training | Cross-validation | Test | All |
| ADH | 0.5000 | 0.7973 | 0.2654 | 0.5339 |
| 0.6200 | | | | |
| ALDH | 0.1490 | 0.7033 | 0.3503 | 0.2325 |
| 0.6683 | | | | |
| FMO | 0.2280 | 0.7342 | 0.4690 | 0.3245 |
| 0.3200 | | | | |

Above all, despite the possibility of inherent noise in the data, the problems of overfitting and underfitting appeared more evident with the MLR model even when descriptors were relatively stable but less significant with the ANN model as revealed in tables 4 to 7 except for the Log($V_{max}$) cases in which correlations were relatively weak. This observation in particular implies that the ANN's model is able to learn better even with noisy data. Hence, more credible for generalisation. Therefore, the results on the tables do not imply that the MLR model is superior where its performance (R) are relatively higher.

The limitations of the models which necessitated poor performances in some instances on the datasets become clear when the data sources are taken into consideration. The fact that the experimental $K_m$ and $V_{max}$ values were obtained from the scientific literature implies that they resulted from different laboratory experiments that used different orders and employed conditions which vary. For instance, pH and temperature conditions will influence enzymatic activities (Garrett and Grisham, 2010). Furthermore, the rate data reported as either $V_{max}$ or $K_{cat}$ values required transformation to convert the rates into same units using conversion factors (Pirovano, *et al.*, 2015). Also, the merging of data for various mammals (that is, human, horse, rat, pig, mouse, and rabbit) and for the several isoenzymes is a likely cause of variations. Additionally, the correlation threshold of absolute 0.4 between descriptors and the $K_m$ and $V_{max}$ values reported by Pirovanol, *et al.*, 2015 could not be achieved in this work due to the limited descriptors software within reach. Therefore, the descriptors used in this work are of absolute correlation values in the range of 0.2 to 0.6, unlike those reported where correlations as high as 0.9 were achieved.

## CONCLUSION

The predictive strengths of two learning algorithms have been evaluated in this work, that is, those of Artificial Neural Networks and Multiple Linear Regression based Quantitative Structure-Activity Relationships, using existing data and accomplished with the MATLAB programming tool. The enzyme data utilized for achieving the objectives contained information for several xenobiotic compounds metabolized by the ADH, ALDH, and FMO, and for various mammalian species.

The main properties which determined the affinity coefficient ($1/K_m$) appeared to be enzyme specific. The partition coefficient and functional group were those that mostly influenced ADH, ALDH, and FMO. Size, shape, symmetry, and atom distribution were the most influential predictors for the maximum reaction rate ($V_{max}$). The constant $V_{max}$ is indicative of the speed of reaction of the catalysed process involving the interaction between substrate or xenobiotic and enzyme.

This study is useful for understanding the principles behind biotransformation by the liver enzymes and for predicting the enzymatic constants ($K_m$ and $V_{max}$) of the four main mammalian enzymes metabolizing various xenobiotics. It is also relevant for choice-making when confronted with the issue of selecting an appropriate model considering the nature of data available for analysis.

## REFERENCES

[1] Alessandra Pirovano, Stefan Brandmaier, Mark A. J. Huijbregts, Ad M. J. Ragas, Karin Veltman

and A. Jan Hendriks (2015). The utilisation of structural descriptors to predict metabolic constants of xenobiotics in mammals. *Environmental Toxicology and Pharmacology*. 39: 247-258.

[2] Albert Chern (2015). "Introduction to MATLAB". *ACM11 Spring 2015*, California Institute of Technology.

[3] Alexandre Varnek and Igor Baskin (2011). Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis?. *Journal of Chemical Information and Modelling*, dx.doi.org/10.1021/ci200409x.

[4] Ammar Abdo, Beining Chen, Christoph Mueller, Naomie Salim, and Peter Willett. (2010). Ligand-based virtual screening using Bayesian networks. *J. Chem. Inf. Model*. 50 (6) 1012–1020.

[5] Andrea Mauri, Viviana Consonni, Manuela Pavan, and Roberto Todeschini (2006). Dragon software: an easy approach to molecular descriptor calculations. *MATCH Commun. Math. Comput. Chem.* 56: 237-248, ISSN 0340 – 6253.

[6] Andreas Karoly Gombert and Jens Nielsen (2000). Mathematical modelling of metabolism. *Current Opinion in Biotechnology*, 11: 180–186.

[7] Andrew Ng. (2018). *Coursera*. Stanford Online Machine Learning Lecture.

[8] Antonio Lavecchia (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today*, 20 (3) 318 – 331.

[9] Bailey J. E. (1998). Mathematical modelling and analysis in biochemical engineering: past accomplishments and future opportunities. *Biotechnology Prog*, 14: 8-20.

[10] Balaz, S. (2009). Modelling kinetics of subcellular disposition of chemicals. *Chem. Rev.* 109: 1793–1899.

[11] BioFoundations (2018). The Detoxification and Biotransformation System in the Human Body. https://biofoundations.org/the-detoxification-and-biotransformation-system-in-the-human-body/. *Extracted on 29th March 2018.*

[12] BRENDA: The Comprehensive Enzyme information system. https://www.brenda enzymes.org/.

[13] Byvatov, E. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci*. 43: 1882–1889.

[14] Chemical Computing Group. https://www.chemcomp.com/journal/descr.htm. *Extracted on the 26th of January 2019*.

[15] Cheng, T. *et al*. (2011). Binary classification of aqueous solubility using support vector machines with reduction and recombination feature selection. *J. Chem. Inf. Model.* 51: 229–236.

[16] Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I.,Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini,R., Consonni, V., Kuz'min, V.E., Cramer, R., Benigni, R., Yang,C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., Tropsha,A. (2013). QSAR modelling: where have you been? Where are you going to?. *J. Med. Chem.* 57: 4977–5010.

[17] Consonni, V., Todeschini, R. (2010). Molecular descriptors. *Recent Advances in QSAR Studies*. Springer, Dordrecht, the Netherlands, pp. 29–102.

[18] David A. Winkler and Frank R. Burden (2000). "Robust QSAR Models from Novel Descriptors and Bayesian Regularised Neural Networks". *Molecular Simulation*. 24: 4-6, 243-258, DOI: 10.1080/08927020008022374.

[19] Deconinck, E. *et al.* (2006). Classification tree models for the prediction of blood– brain barrier passage of drugs. *Journal of Chem. Inf. Model*. 46: 1410–1419.

[20] Dmitrij Martynenko (2015). "Big Data Analytics with MATLAB". http://www.mathworks.com/discovery/matlab-mapreduce-hadoop.html. *Extracted on 29th March 2018.*

[21] Emre Karakoc, S. Cenk Sahinalp, and Artem Cherkasov (2006). Comparative QSAR – and Fragments Distribution Analysis of Drugs, Drug-likes, Metabolic Substances, and Antimicrobial Compounds. *J. Chem. Inf. Model.* 46: 2167-2182.

[22] Fogel, G.B. (2008). Computational intelligence approaches for pattern discovery in biological systems. *Brief Bioinform*. 9: 307–316.

[23] Foody, G.M. and Mathur, A. (2006). The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by SVM. *Remote Sens. Environ*. 103: 179–189.

[24] Frank R. Burden (1999). Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.*, 42: 3183-3187.

[25] Frank, E. *et al.* (2000). Technical note: naive Bayes for regression. *Mach. Learn.* 41: 5–25

[26] Garrett, R., Grisham, C. M., (2010). Biochemistry, fourth ed. Brooks/Cole, Cengage Learning, Boston, MA, USA.

[27] GeorgeW. Bassel, Enrico Glaab, Julietta Marquez, Michael J. Holdsworth, and Jaume Bacardit (2011). Functional Network Construction in Arabidopsis Using Rule-Based Machine Learning on Large-Scale Data Sets. *Large-Scale Biology Article*, 23: 3101–3116.

[28] Gershenfeld N. A. (1999). *The Nature of Mathematical Modelling*. Cambridge: Cambridge University Press.

[29] Gianpaolo Bravi and James H. Wikel (2000). Application of MS-WHIM Descriptors: 1. Introduction of New Molecular Surface Properties and 2. Prediction of Binding Affinity Data. *Quant. Struct. Act. Relat.*, 19. https://doi.org/10.1002/(SICI)1521 3838(200002)19:1<29::AID-QSAR29>3.0.CO;2-P.

[30] Gleeson, M. P. *et al.* (2006). In silico human and rat Vss quantitative structure– activity relationship models. *J. Med. Chem.* 49: 1953–1963.

[31] Gregory Sliwoski, Jeffrey Mendenhall, and Jens Meiler (2015). Autocorrelation descriptor improvements for QSAR: 2DA_Sign and 3DA_Sign. *J. Comput. Aided Mol Des*. DOI 10.1007/s10822-015-9893-9.

[32] Haiping Lu, Konstantinos N. Plataniotis, Anastasios N. Venetsanopoulos (2011). A survey of multilinear subspace learning for tensor data. *Pattern Recognition*. 44: 1540–1551.

[33] Hansch, C., Mekapati, S.B., Kurup, A., Verma, R.P., (2004). QSAR of cytochrome P450. *Drug. Metab. Rev*. 36: 105–156.

[34] Haykin, S. S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall.

[35] Ho, T. K. (1998). The random subspace method for constructing decision forests. *ITPAM* 20: 832–844.

[36] Hou, T. *et al.* (2007). ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *Journal of Chem. Inf. Model*. 47: 2408–2415.

[37] Iurii Sushko, Sergii Novotarskyi, Robert Ko¨rner, Anil Kumar Pandey, Matthias Rupp, Wolfram Teetz, Stefan Brandmaier, Ahmed Abdelaziz, Volodymyr V. Prokopenko, Vsevolod Y. Tanchuk, Roberto Todeschini, Alexandre Varnek, Gilles Marcou, Peter Ertl, Vladimir Potemkin, Maria Grishina, Johann Gasteiger, Christof Schwab, Igor I. Baskin, Vladimir A. Palyulin, Eugene V. Radchenko, William J. Welsh, Vladyslav Kholodovych, Dmitriy Chekmarev, Artem Cherkasov, Joao Aires-de-Sousa, Qing-You Zhang, Andreas Bender, Florian Nigsch, Luc Patiny, Antony Williams, Valery Tkachenko, Igor V. Tetko (2011). Online chemical modelling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol Des.* 25: 533–554, DOI 10.1007/s10822-011-9440-2.

[38] Jacek Kujawski, Marek K. Bernard, Anna Janusz, and Weronika Kuzma (2011). Prediction of log P: ALOGPS Application in Medicinal Chemistry Education. *J. Chem. Educ*. 2012, 89, 64–67. dx.doi.org/10.1021/ed100444h.

[39] Johannes Kirchmair, Andreas H. Göller, Dieter Lang, Jens Kunze, Bernard Testa, Ian D. Wilson, Robert C. Glen and Gisbert Schneider (2015). Predicting drug metabolism: experiment and/or computation?. *PERSPECTIVES*. 14: 389-404.

[40] Jun Zhang, Zhi-hui Zhan, Ying Lin, Ni Chen, Yue-jiao Gong, Jing-hui Zhong, Henry S.H. Chung, Yun Li, Yu-hui Shi (2011). Evolutionary Computation Meets Machine Learning: A Survey. *IEEE Computational Intelligence Magazine*, pp 68-75.

[41] Kathleen M. Knights, Andrew Rowland, and John O. Miners (2013). Renal drug metabolism in humans: The potential for drug-endobiotic interactions. *British Journal of Clinical Pharmacology*. 76 (4) 587-602.

[42] Kauffman, G. W. and Jurs, P.C. (2001). QSAR and k-nearest neighbour classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *Journal of Chem. Inf.* Comp. Sci. 41: 1553–1560.

[43] Krueger, S.K., Williams, D.E., (2005). Mammalian flavin-containing monooxygenases: structure/function, genetic polymorphisms and role in drug metabolism. *Pharmacol. Ther.* 106: 357–387.

[44] Lamanna, C. *et al.* (2008). Straightforward recursive partitioning model for discarding insoluble compounds in the drug discovery process. *Journal of Med. Chem*. 51, 2891–2897

[45] Lash, Lawrence H. (1994). "Role of Renal Metabolism in Risk". *Environmental Health Perspectives.* 102 (11) 75-79.

[46] Lewis, D.F.V., (1999). Frontier orbitals in chemical and biological activity: quantitative relationships and mechanistic implication. *Drug. Metab. Rev.* 31: 755–816.

[47] List of molecular descriptors calculated by DRAGON. http://www.talete.mi.it/products/dragon_molecular_descriptor_list.pdf. *Extracted on 30th January 2019.*

[48] Lowe, R. *et al.* (2012). Predicting the mechanism of phospholipidosis. *Journal of Cheminformatics* 4: 2.

[49] Marco Chiarandini. "Machine Learning: Linear Regression and Neural Networks". *Introduction to Computer Science.* Department of Mathematics & Computer Science University of Southern Denmark.

[50] Margot Gerritsen (2006). "A brief introduction to MATLAB". *Linear Algebra with Application to Engineering Computations, Autumn 2006 Handout 3*.

[51] MathWorks (2016). "Introducing Machine Learning". mathwork.com/trademarks. *Extracted on 2nd April 2018.*

[52] Mayer-Schönberger, V., and Cukier, K. (2014). Big data: A revolution that will transform how. *American Journal of Epidemiology*. 179 (9) 1143–1144.

[53] Mente, S. R. *et al*. (2005). A recursive-partitioning model for blood–brain barrier permeation. *J. Comput. Aided Mol. Des.* 19: 465–481:

[54] Nielsen J., Jørgensen H. S. (1996). A kinetic model for the penicillin biosynthetic pathway in Penicillium chrysogenum. *Control Eng Practice*, 4:765-771.

[55] Nigsch, F. *et al.* (2006). Melting point prediction employing k-nearest neighbour algorithms and genetic parameter optimization. *J. Chem. Inf. Model.* 46: 2412–2422.

[56] Oleg Devinyak, Dmytro Havrylyuk, and Roman Lesyk (2014). 3D-MoRSE Descriptors Explained. *Journal of Molecular Graphics and Modelling*. DOI: 10.1016/j.jmgm.2014.10.006.

[57] Patel, J. and Chaudhari, C. (2005). Introduction to the artificial neural networks and their applications in QSAR studies. *ALTEX.* 22: 271.

[58] Pirovano, A., Huijbregts, M.A.J., Ragas, A.M.J., Veltman, K.,Hendriks, A.J. (2014). Mechanistically-based QSARs to describe metabolic constants in mammals. *ATLA.* 42: 59–69.

[59] Pissara P. N., Nielsen J., Bazin M. J. (1996). Pathway kinetics and metabolic control analysis of a high-yielding strain of *Penicillium chrysogenum* during fed batch cultivations. *Biotechnology Bioeng*, 51:168-176.

[60] Rizzi M, Baltes M, Theobald U, Reuss M (1997). *In vivo* analysis of metabolic dynamics in *Saccharomyces cerevisiae*: II. mathematical model. *Biotechnol Bioeng*, 55:592-608.

[61] S. Agatonovic-Kustrin and R. Beresford (2000). Basic concepts of artificial neural network (ANN) modelling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*. 22 (5) 717-727.

[62] Sakiyama, Y. et al. (2008). Predicting human liver microsomal stability with machine learning techniques. *J. Mol. Graph. Model.* 26: 907–915.

[63] Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C.,Rother, M., Söhngen, C., Stelzer, M., Thiele, J., Schomburg, D. (2011). "BRENDA, the enzyme information system". *Nucleic Acids Res*. 39: D670–D676.

[64] Schilling C. H., Edwards J. S., Palsson B. O. (1999). Toward metabolic phenomics: analysis of genomic data using flux balances. *Biotechnol. Prog*. 15: 288-295.

[65] Shashi K. Ramaiah and Atrayee Banerjee (2015). "Liver Toxicity of Chemical Warfare Agents": Handbook of Toxicology of Chemical Warfare. *ScienceDirect*. Pp 615-626.

[66] Theilgaard H., Nielsen J. (1999). Metabolic control analysis of the penicillin biosynthetic pathway: the influence of the LLD-ACV: bis ACV ratio on the flux control. *Anton Leeuw Int J G*, 75: 145-154.

[67] Tiago M. Fragoso and Francisco Louzada Neto (2017). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*. 0(0)1–28. doi:10.1111/insr.12243.

[68] Tropsha, Alexander (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*. 29: 6-7: 476–488.

[69] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody (2016). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research.* http://dx.doi.org/10.1016/j.jbusres.2016.08.001.

[70] Vapnik, V. N. (1998). Statistical Learning Theory. Wiley.

[71] Vapnik, V. N. (2000). The Nature of Statistical Learning Theory. Springer

[72] Vasiliou, V., Pappa, A., Petersen, D.R., (2000). Role of aldehyde dehydrogenases in endogenous and xenobiotic metabolism. *Chem. Biol. Interact.* 129: 1–19.

[73] Viviana Consonni, Roberto Todeschini, Manuela Pavan, and Paola Gramatica (2002). Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. *J. Chem. Inf. Comput. Sci.* 42: 693-705.

[74] Von Korff, M. and Sander, T. (2006). Toxicity-indicating structural patterns. *J. Chem. Inf. Model.* 46: 536–544.

[75] Waller, C.L., Evans, M.V., and McKinney, J.D. (1996). Modelling the cytochrome P450-mediated metabolism of chlorinated volatile organic compounds. *Drug Metab. Dispos*.24: 203–210.

[76] Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*. 44: 92–107.

[77] Wilbert B. Copeland, Bryan A. Bartley, Deepak Chandran, Michal Galdzicki, Kyung H. Kim, Sean C. Sleight, Costas D. Maranas, Herbert M. Sauro (2012). Computational tools for metabolic engineering. *Metabolic Engineering*. 14: 270–280.

[78] Willett, P. *et al.* (2007). Prediction of ion channel activity using binary kernel discrimination. *J. Chem. Inf. Model.* 47: 1961–1966.

[79] Yousefinejad S. and Hemmateenejad B. (2015). "Chemometrics tools in QSAR/QSPR studies: A historical perspective". *Chemometric and Intelligent Laboratory Systems*. Part B, 149: 177–204.

[80] Zvinavashe, E., Murk, A.J., Rietjens, I.M.C.M., (2008). "Promises and pitfalls of quantitative structure–activity relationship approaches for predicting metabolism and toxicity". *Chem. Res. Toxicol.* 21, 2229–2236.