

This article was downloaded by:

On: 14 January 2011

Access details: *Access Details: Free Access*

Publisher *Taylor & Francis*

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Molecular Simulation

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t713644482>

QSAR study of oxazolidinone antibacterial agents using artificial neural networks

C. Zou^a; L. Zhou^a

^a Chemical Engineering Department, Sichuan University, Chengdu, People's Republic of China

To cite this Article Zou, C. and Zhou, L.(2007) 'QSAR study of oxazolidinone antibacterial agents using artificial neural networks', *Molecular Simulation*, 33: 6, 517 — 530

To link to this Article: DOI: 10.1080/08927020601188528

URL: <http://dx.doi.org/10.1080/08927020601188528>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

QSAR study of oxazolidinone antibacterial agents using artificial neural networks

C. ZOU and L. ZHOU*

Chemical Engineering Department, Sichuan University, Chengdu 610065, People's Republic of China

(Received July 2006; in final form December 2006)

The oxazolidinones antibacterial agents have been studied for their quantitative structure-activity relationships (QSAR). Molecules were represented by constitutional, topostructural, chemical and quantum chemical descriptors. Partial least square (PLS) regression was used to model the relationships between molecular descriptors and biological activity of molecules. The predictive ability of the acquired models was evaluated by the activity prediction of the prediction set compounds. Artificial neural network (ANN) was also employed to model the nonlinear structure-activity relationships. The results showed that the linear model does not perform as well as the nonlinear model in terms of predictive ability.

Keywords: Oxazolidinone; Partial least square regression; Artificial neural network; QSAR

1. Introduction

The wide use even the abuse of antibiotics has facilitated the development of the bacterial resistance to the currently available antibacterial agents, which has become a global healthy problem. Of particular concerns are the infections caused by multidrug-resistant Gram-positive pathogens, which are responsible for the significant morbidity and mortality in both hospital and community settings. The oxazolidinones are a new class of totally synthetic antibacterial agents, and unrelated to any other currently available agents. The oxazolidinones exemplified by eperezolid and linezolid is one such class of antibacterial agents with potent activity against Gram-positive organisms including methicillin-resistant *Staphylococcus aureus* (MRSA), methicillin-resistant *Staphylococcus epidermidis* (MRSE) and vancomycin resistant enterococci (VRE) [1,2]. These compounds have been shown to inhibit translation at the initiation phase of protein synthesis in bacteria by binding to the 50S ribosomal subunit [3]. At the 1987 Interscience Conference on Antimicrobial Agents and Chemotherapy (ICAAC) workers from the DuPont company formally reported the structure and antibacterial activity profiles of this antibacterial agents. After that, a number of SAR and QSAR studies on oxazolidinone derivatives are available [4–11]. In the recent past, some efforts have been made to

understand 3D-QSAR [10,11]. We aim at developing a model for the 2D-QSAR for a series of oxazolidinones in our study.

The most familiar standard approaches to QSAR are based on multiple linear regression (MLR) and partial least squares (PLS) regression. However, these approaches can capture only linear relationships between molecular characteristics and structural or functional features to be predicted. In contrast, artificial neural network (ANN) is capable of recognizing highly nonlinear relationships. The flexibility of ANN enables it to discover more complex relationships in experimental data, when it is compared with the traditional alternatives to the QSAR/QSPR studies [12–18]. In this study, PLS and ANN techniques were used for modeling the relationships between biological activity and molecular descriptors.

2. Materials and methods

2.1 Data set and descriptor calculation

The biological data for 118 oxazolidinone derivatives used in this study were obtained from Pharmacia Corporation (the compounds from number 1 to number 68 come from literature [19], number 69 to number 91 from literature [20], and number 92 to number 118 from literature [21]).

*Corresponding author. Email: zhouluscu@163.com

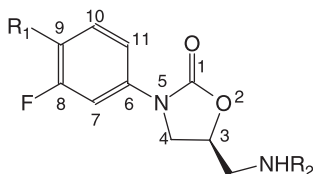


Figure 1. Basic structure of the oxazolidinone derivatives used in this study.

All the derivatives were tested for antibacterial activity *in vitro* against methicillin-susceptible *S. aureus*. The biological activity data, minimum inhibitory concentration (MIC, the concentration of drug required to kill the bacterial cells or inhibit their growth under standardized conditions) values, were determined using standard agar dilution methods. The MIC values were converted to logarithmic scale (pMIC) and then used for subsequent QSAR analysis as dependent variable. The basic structures of these compounds are shown in figure 1. The pMIC values of these compounds are listed in table 1.

Molecular descriptors define the molecular structure and physicochemical properties of molecules by a single number. A total of 105 calculated structure features were calculated for each of the 118 compounds including constitutional, quantum chemical descriptors, chemical and topostructural. Constitutional descriptors depend fundamentally on the composition of the molecule. The counts of atoms of different elements and the molecular weight reflect the composition. Quantum chemical descriptors were calculated using HyperChem 7.0 for Windows. The MM + molecular mechanics force field was first run to get close to the optimized geometry. The conformation obtained from molecular mechanics was subjected to a refined geometry optimization using AM1 semi-empirical quantum chemistry. Topological descriptors were calculated using two-dimensional representation of the molecules. The brief description of those descriptors, calculated for this study, is represented in table 2.

2.2 Rational division of the dataset

In order to obtain a reliable (validated) QSAR model, an available dataset should be divided into the training and prediction sets. Ideally, this division must be performed in a way so that the points representing both the training and prediction sets are distributed within the whole descriptor space occupied by the entire dataset, and each point of the prediction set is close to at least one point of the training set. In this study the division of a dataset into the training and prediction sets can be performed using clustering techniques. A cluster sampling algorithm would focus on densely occupied regions of the space and hence avoid outliers. After the clustering process, the structure closest to the centre of a cluster was selected as the representative structure. The data set was divided into training and prediction sets (10%) by a *K*-means clustering algorithm clustering on descriptors (*X*) and biological activity (*Y*)

values taken together [31]. Clustering on *X* and *Y* data together, rather than just on *X*, is our preferred method in that it clusters compounds according to all of the given information. This may lead to different prediction sets for different groups of indices but is appropriate when searching for the best model to represent a data set.

2.3 PLS regression modeling

The software package used for conducting PLS analysis was SAS 9.0. PLS regression method appears frequently in the literature [32–34]. It is good in avoiding the collinearity trouble. PLS is a bilinear modeling technique where information in the descriptor matrix *X* is projected onto a small number of latent variables (LV) called PLS components, referred to as PCs, which are linear combination of the original variables. The matrix *Y* is simultaneously used in estimating the “latent” variables in *X* that will be most relevant to predict the *Y* variables. All descriptor variables are preprocessed by autoscaling, using weights based on the variables’ standard deviation and the data are mean-centered prior to PLS processing. This method of scaling is necessary when the values have different orders of magnitude and different units, as is the case here.

Cross-validation was employed to select the used optimum number of LVs. With cross-validation, some samples were kept out of the calibration and used for prediction. The process was repeated so that each of the samples was kept out once. The predicted values of left-out samples were then compared to the observed values using predicted residual sum of squares (PRESS). The PRESS obtained in the cross-validation was calculated each time that a new LV was added to the model. The optimum number of LVs was concluded as the first local minimum in the PRESS versus LV plot. PRESS is defined as

$$\text{PRESS} = \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (1)$$

where \hat{y}_i is the estimated value of the *i*th object and y_i is the corresponding reference value of this object.

2.4 Descriptor selection

In order to determine which variables were significantly correlated with activity, regression coefficients (*B*) [35,36] and variable importance for the projection (VIP) [36,37] of each molecular descriptor was used to unravel which descriptor variables were the most relevant to explain pMIC. This PLS regression creates matrixes of loadings, inner relations and weights that are used to obtain a vector of linear multiple regression. High values of the regression coefficients signify that the descriptor are important to the regression. The VIP represents the value of each predictor in fitting the model for both predictors and responses. If a predictor has a relatively small coefficient and a small value of VIP, then it is a prime candidate for deletion. To decrease the redundancy existing in the descriptors

Table 1. Activity (pMIC[†]) data of oxazolidinone derivatives and the corresponding predicted values by PLS and ANN models.


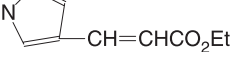
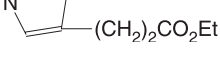

| Compound No | R1 | R2 | pMIC | Predicted | |
|-----------------|---|----|------|-----------|------|
| | | | | PLS | ANN |
| 1 |  | AC | 5.52 | 5.37 | 5.18 |
| 2 |  | AC | 6.16 | 5.34 | 5.91 |
| 3 [‡] |  | AC | 5.76 | 5.40 | 5.36 |
| 4 |  | AC | 5.59 | 5.34 | 6.81 |
| 5 [‡] |  | AC | 5.86 | 5.21 | 5.69 |
| 6 [‡] |  | AC | 5.56 | 5.84 | 5.59 |
| 7 |  | AC | 5.29 | 5.02 | 4.91 |
| 8 |  | AC | 4.37 | 4.85 | 4.93 |
| 9 |  | AC | 5.03 | 4.99 | 5.04 |
| 10 |  | AC | 5.04 | 5.17 | 5.14 |
| 11 [‡] |  | AC | 4.99 | 4.69 | 4.74 |
| 12 |  | AC | 4.77 | 4.62 | 4.72 |
| 13 |  | AC | 5.27 | 5.48 | 5.39 |
| 14 |  | AC | 4.37 | 4.80 | 4.31 |
| 15 |  | AC | 5.29 | 5.42 | 5.38 |

Table 1 – continued


| Compound | | | | Predicted | |
|----------|--|----|------|-----------|------|
| No | R1 | R2 | pMIC | PLS | ANN |
| 16 |  <chem>CN1C=CC=C1C(=O)OC</chem> | AC | 5.01 | 5.22 | 4.73 |
| 17 |  <chem>C1=CN=CN=C1</chem> | AC | 5.22 | 5.32 | 4.91 |
| 18 |  <chem>CCOC(=O)C1=CN=CN=C1</chem> | AC | 5.01 | 4.42 | 4.98 |
| 19 |  <chem>NC(=O)C1=CN=CN=C1</chem> | AC | 4.37 | 4.81 | 4.75 |
| 20 |  <chem>CNC(=O)C1=CN=CN=C1</chem> | AC | 4.39 | 4.60 | 4.37 |
| 21 |  <chem>N#CC1=CN=CN=C1</chem> | AC | 5.86 | 4.90 | 5.71 |
| 22 |  <chem>Ic1c[nH]cn1</chem> | AC | 5.36 | 5.51 | 5.21 |
| 23 |  <chem>CC(C)(C)OC(=O)Nc1c[nH]cn1</chem> | AC | 4.75 | 4.92 | 4.72 |
| 24 |  <chem>CC(=O)Nc1c[nH]cn1</chem> | AC | 4.41 | 4.53 | 4.57 |
| 25 |  <chem>CC(C)([Si](C)(C)C)C(=O)Nc1c[nH]cn1</chem> | AC | 5.03 | 5.05 | 4.85 |
| 26 |  <chem>CC(=O)Nc1c[nH]cn1</chem> | AC | 5.25 | 4.93 | 5.55 |
| 27 |  <chem>FC(F)(F)c1c[nH]cn1</chem> | AC | 5.00 | 5.05 | 4.93 |
| 28 |  <chem>CC(C)(C)OC(=O)Nc1c[nH]cn1</chem> | AC | 5.05 | 5.01 | 4.94 |
| 29 |  <chem>Nc1c[nH]cn1</chem> | AC | 5.24 | 5.34 | 5.37 |

Table 1 – continued

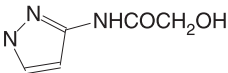
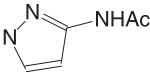
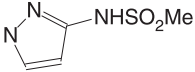
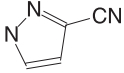
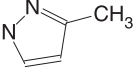
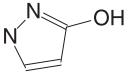
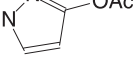
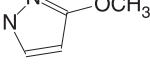

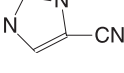

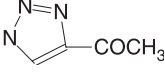
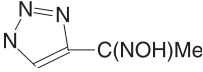
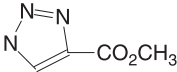
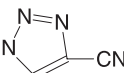
| Compound | | | | Predicted | |
|-----------------|---|----|------|-----------|------|
| No | R1 | R2 | pMIC | PLS | ANN |
| 30 |  | AC | 4.41 | 4.49 | 4.53 |
| 31 |  | AC | 4.71 | 4.79 | 4.94 |
| 32 |  | AC | 4.43 | 4.63 | 4.48 |
| 33 |  | AC | 5.56 | 5.81 | 5.49 |
| 34 |  | AC | 4.94 | 5.28 | 5.37 |
| 35 [‡] |  | AC | 4.64 | 4.87 | 4.72 |
| 36 |  | AC | 4.71 | 4.81 | 4.76 |
| 37 |  | AC | 4.96 | 4.79 | 5.17 |
| 38 |  | AC | 5.22 | 5.03 | 5.38 |
| 39 |  | AC | 4.95 | 5.23 | 5.42 |
| 40 |  | AC | 5.53 | 5.20 | 5.53 |
| 41 [‡] |  | AC | 5.58 | 5.08 | 5.25 |
| 42 |  | AC | 4.99 | 5.33 | 4.66 |
| 43 |  | AC | 4.99 | 4.47 | 5.21 |
| 44 |  | AC | 5.56 | 4.99 | 5.31 |

Table 1 – continued

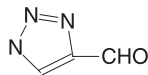
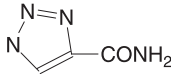
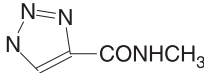
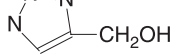
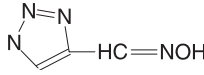
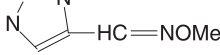

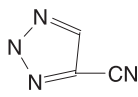

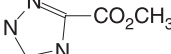
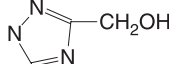
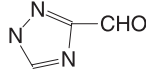
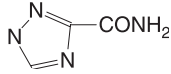
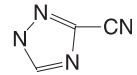
| Compound | | | | Predicted | |
|-----------------|---|----|------|-----------|------|
| No | R1 | R2 | pMIC | PLS | ANN |
| 45 |  | AC | 5.86 | 4.98 | 5.69 |
| 46 |  | AC | 4.98 | 5.07 | 4.99 |
| 47 |  | AC | 4.69 | 4.80 | 4.74 |
| 48 |  | AC | 4.96 | 4.60 | 5.23 |
| 49 |  | AC | 4.68 | 4.79 | 4.64 |
| 50 |  | AC | 4.99 | 5.00 | 5.37 |
| 51 |  | AC | 4.92 | 4.46 | 5.04 |
| 52 [‡] |  | AC | 5.86 | 5.39 | 5.58 |
| 53 |  | AC | 4.92 | 4.88 | 5.12 |
| 54 |  | AC | 4.39 | 4.52 | 4.76 |
| 55 |  | AC | 4.36 | 4.56 | 4.42 |
| 56 [‡] |  | AC | 4.96 | 4.75 | 5.20 |
| 57 |  | AC | 4.37 | 4.92 | 4.61 |
| 58 |  | AC | 4.96 | 4.83 | 4.83 |

Table 1 – continued

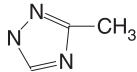
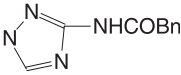
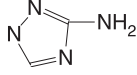
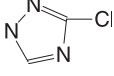


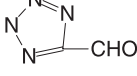
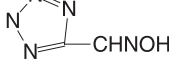
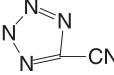

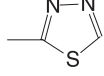
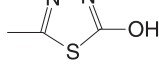
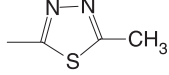
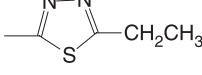
| Compound | | | | Predicted | |
|----------|---|--------------------|------|-----------|------|
| No | R1 | R2 | pMIC | PLS | ANN |
| 59 |  | AC | 4.64 | 5.04 | 4.67 |
| 60 |  | AC | 4.47 | 4.56 | 4.79 |
| 61 |  | AC | 4.64 | 4.85 | 4.72 |
| 62 |  | AC | 4.97 | 4.66 | 4.84 |
| 63 |  | AC | 4.32 | 4.58 | 4.58 |
| 64 |  | AC | 5.23 | 5.36 | 5.15 |
| 65 |  | AC | 4.66 | 5.17 | 5.17 |
| 66 |  | AC | 4.98 | 4.69 | 4.59 |
| 67 |  | AC | 5.26 | 5.22 | 5.40 |
| 68 |  | AC | 4.93 | 4.87 | 5.81 |
| 69 |  | —COCH ₃ | 5.53 | 5.28 | 5.15 |
| 70 |  | —COCH ₃ | 5.55 | 5.48 | 5.55 |
| 71 |  | —COCH ₃ | 5.55 | 5.45 | 5.71 |
| 72 |  | —COCH ₃ | 5.56 | 5.28 | 5.38 |

Table 1 – continued

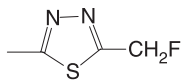
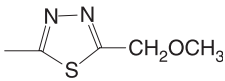
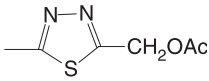
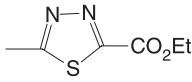
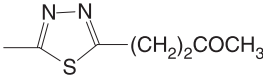
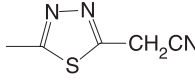
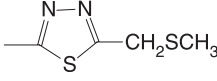
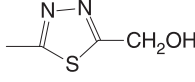
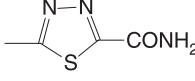
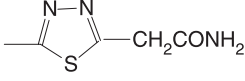
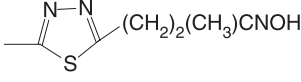
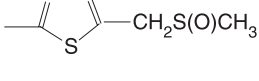
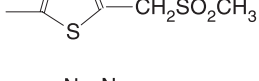
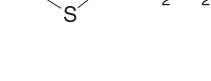
| Compound | | | | Predicted | |
|-----------------|---|--------------------|------|-----------|------|
| No | R1 | R2 | pMIC | PLS | ANN |
| 73 [‡] |  | —COCH ₃ | 5.57 | 5.85 | 5.54 |
| 74 |  | —COCH ₃ | 5.28 | 4.91 | 5.28 |
| 75 [‡] |  | —COCH ₃ | 5.63 | 5.72 | 5.54 |
| 76 |  | —COCH ₃ | 4.71 | 5.16 | 4.69 |
| 77 |  | —COCH ₃ | 5.61 | 5.40 | 5.61 |
| 78 |  | —COCH ₃ | 5.27 | 5.03 | 5.37 |
| 79 |  | —COCH ₃ | 5.60 | 5.68 | 5.47 |
| 80 |  | —COCH ₃ | 5.56 | 5.15 | 5.59 |
| 81 |  | —COCH ₃ | 5.88 | 5.47 | 5.60 |
| 82 |  | —COCH ₃ | 4.99 | 5.18 | 4.83 |
| 83 |  | —COCH ₃ | 5.32 | 5.10 | 5.34 |
| 84 |  | —COCH ₃ | 4.71 | 4.62 | 4.80 |
| 85 |  | —COCH ₃ | 5.03 | 4.84 | 4.99 |
| 86 |  | —COCH ₃ | 4.66 | 5.13 | 4.90 |

Table 1 – continued

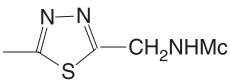
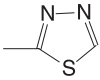
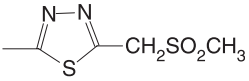
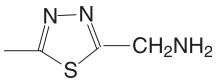
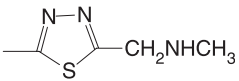

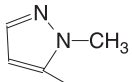
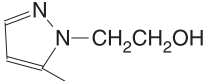
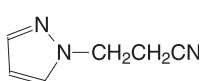
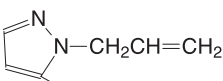
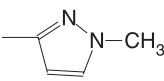
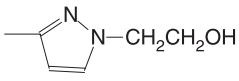
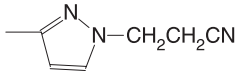
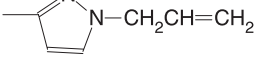
| Compound | | | | Predicted | |
|----------|---|--------------------|------|-----------|------|
| No | R1 | R2 | pMIC | PLS | ANN |
| 87 |  | —COCH ₃ | 4.98 | 5.21 | 5.05 |
| 88 |  | —COCH ₃ | 6.15 | 6.23 | 6.19 |
| 89 |  | —CSCH ₃ | 6.25 | 5.89 | 5.90 |
| 90 |  | —CSCH ₃ | 5.88 | 5.61 | 6.18 |
| 91 |  | —CSCH ₃ | 5.90 | 5.75 | 5.95 |
| 92 |  | —COCH ₃ | 4.90 | 5.08 | 5.00 |
| 93 |  | —COCH ₃ | 5.52 | 5.57 | 5.32 |
| 94 |  | —COCH ₃ | 4.36 | 4.55 | 4.47 |
| 95 |  | —COCH ₃ | 4.37 | 4.50 | 4.58 |
| 96 |  | —COCH ₃ | 4.65 | 4.89 | 4.54 |
| 97 |  | —COCH ₃ | 4.62 | 5.23 | 4.63 |
| 98 |  | —COCH ₃ | 4.36 | 4.58 | 4.44 |
| 99 |  | —COCH ₃ | 4.97 | 5.20 | 4.92 |
| 100 |  | —COCH ₃ | 4.65 | 4.61 | 4.64 |

Table 1 – continued

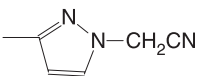
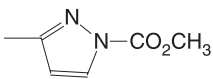
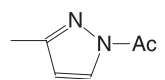
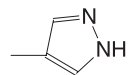
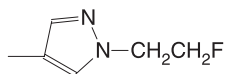
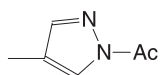
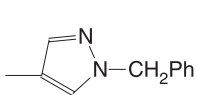
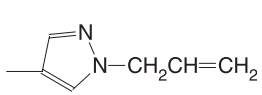
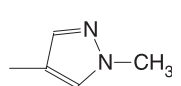
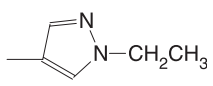
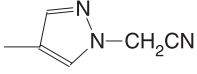
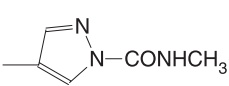
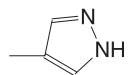
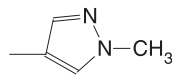
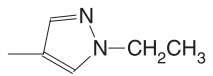
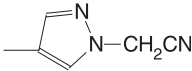
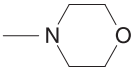
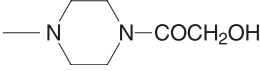
| Compound | | | | Predicted | |
|------------------|---|--------------------|------|-----------|------|
| No | R1 | R2 | pMIC | PLS | ANN |
| 101 |  | —COCH ₃ | 5.25 | 4.87 | 5.24 |
| 102 |  | —COCH ₃ | 4.67 | 4.66 | 4.68 |
| 103 [‡] |  | —COCH ₃ | 4.97 | 5.28 | 4.56 |
| 104 |  | —COCH ₃ | 5.20 | 4.86 | 5.10 |
| 105 |  | —COCH ₃ | 5.26 | 5.14 | 5.43 |
| 106 |  | —COCH ₃ | 5.28 | 5.11 | 5.10 |
| 107 |  | —COCH ₃ | 4.71 | 5.05 | 4.62 |
| 108 |  | —COCH ₃ | 4.95 | 5.12 | 5.03 |
| 109 |  | —COCH ₃ | 5.52 | 5.25 | 5.62 |
| 110 |  | —COCH ₃ | 4.94 | 5.28 | 4.90 |
| 111 |  | —COCH ₃ | 5.25 | 4.99 | 5.23 |
| 112 |  | —COCH ₃ | 5.59 | 5.55 | 5.51 |
| 113 |  | —CSCH ₃ | 5.83 | 5.84 | 5.79 |
| 114 [‡] |  | —CSCH ₃ | 5.84 | 6.44 | 5.66 |
| 115 |  | —CSCH ₃ | 5.56 | 5.60 | 5.66 |

Table 1 – continued

| Compound No | <i>R</i> 1 | <i>R</i> 2 | <i>p</i> MIC | Predicted | |
|-------------|---|--------------------|--------------|-----------|------|
| | | | | PLS | ANN |
| 116 |  | —CSCH ₃ | 6.17 | 5.68 | 5.99 |
| linezolid |  | AC | 4.93 | 5.26 | 5.12 |
| Eperezolid |  | AC | 4.99 | 5.11 | 5.00 |

[†] The unit of MIC is mol l⁻¹.

[‡] The compounds used in the prediction set.

data matrix, the correlation between the selected descriptors was examined, and those descriptors with low colinearity were considered for ANN modeling.

2.5 ANN modeling

In this study, ANN calculations were performed with Matlab 6.5. The ANN employed in this study was the back-propagation (BP) neural network. Of all neural networks, BP network is the most wide used model. Briefly, BP neural networks are made up of a number of processing units presented in three types of layers with appropriate connection: an input layer which distributes the descriptor data; one or more hidden layers with a variable number of units; an output layer which is trained to match a target set values. Weights are iteratively re-evaluated from the target error in the output layer by using a BP procedure according to the so-called ‘delta rule’ [38]. The network inputs were selected descriptors by PLS analysis, the signal of the output node represents the *p*MIC value for interested compounds and the number of nodes in the hidden layer would be optimized. In order to evaluate the performance of the ANN model, the mean square error (MSE) between the predicted and observed activities of compounds was used. For the evaluation of the predictive power of the networks, the trained ANN model was used to calculate *p*MIC of the molecules included in the prediction set.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (2)$$

where *n* is the number of patterns being evaluated.

3. Results and discussion

The data used in these experiments, consisted of 118 oxazolidinone derivatives. Antibacterial activity was measured by *p*MIC. By using cluster technique [31], this data set was divided into a training set of 106 compounds for developing the PLS and ANN models and a prediction

set of 12 compounds for evaluating the predictive ability of the models. The prediction set was the same for the two methods. The data set and corresponding observed and predicted values of the *p*MIC of all molecules studied by PLS and ANN in this work are shown in table 1. The predictive model building abilities of two methods, PLS and ANN, were compared.

The PLS model included all descriptors. To choose the number of PLS components we used some form of cross-validation. The PRESS statistic is based on the generated residuals. The cross-validation resulted in nine LVs was the optimal number with a minimum PRESS. The values of the accumulated variance of the model with all the autoscaled independent variables, using up to 20 LVs, are listed in table 3. The greater the number of LVs are, the better the predictive capability of the model is. The accumulated variance of the LVs decreases significantly at four LVs (table 3). The model showed a low correlation

Table 2. The calculated chemical descriptors used in this study.

| Descriptor type | Molecular descriptors |
|---------------------|--|
| Constitutional | Molecular weight, number of atoms, number of non-H atoms, number of heteroatoms, number of multiple bonds, number of aromatic bonds, number of functional groups, number of rings, number of H-bond donors, number of H-bond acceptors |
| Topological indices | Randic connectivity indices [22], Kier and Hall connectivity indices and valence connectivity indices [23,24], Wiener index (<i>W</i>) [25], Zagreb indices (<i>M</i>) [26,27], Balaban index (<i>J</i>) [28], Hosoya index (<i>Z</i>) [29] |
| Chemical | log <i>P</i> , hydration energy (<i>E</i> _{hydr}), molar refractivity (MR), Polarizability (Pol), molecular surface area (SA), molecular volume (<i>V</i>) |
| Quantum chemical | Dipole moment (DM), HOMO and LUMO energies, heat of formation (<i>H</i> _{form}), total energy (<i>E</i> _{total}), electronic energy (<i>E</i> _{ele}), the local charges at each atom of the base unit of basic structure (LC _i), most positive charge (MPC), most negative charge (MNC), sum of squares of charges (SSC), hardness (<i>η</i>), softness (<i>S</i>), electronegativity (<i>χ</i>), chemical potential (<i>μ</i>), and electrophilicity (<i>ω</i>) [30] |

Table 3. Values of the accumulated variances of the PLS model with all descriptors and 20 LVs for the independent and dependent variables blocks.

| LV number [#] | Independent | | Dependent | |
|------------------------|-------------|-------|-----------|-------|
| | This LV | Total | This LV | Total |
| 1 | 20.03 | 20.03 | 32.68 | 32.68 |
| 2 | 8.96 | 28.99 | 12.37 | 45.06 |
| 3 | 30.41 | 59.40 | 1.19 | 46.24 |
| 4 | 12.06 | 71.46 | 1.81 | 48.06 |
| 5 | 4.74 | 76.21 | 3.04 | 51.10 |
| 6 | 2.34 | 78.55 | 1.73 | 52.83 |
| 7 | 2.63 | 81.17 | 1.13 | 53.96 |
| 8 | 1.93 | 83.11 | 0.81 | 54.77 |
| 9 | 1.75 | 84.86 | 0.71 | 55.48 |
| 10 | 2.23 | 87.09 | 0.60 | 56.08 |
| 11 | 1.30 | 88.39 | 0.99 | 57.06 |
| 12 | 1.50 | 89.89 | 0.45 | 57.52 |
| 13 | 1.11 | 91.00 | 0.65 | 58.17 |
| 14 | 1.23 | 92.23 | 0.67 | 58.84 |
| 15 | 1.63 | 93.86 | 0.40 | 59.23 |
| 16 | 0.95 | 94.81 | 0.32 | 59.56 |
| 17 | 0.66 | 95.47 | 0.28 | 59.84 |
| 18 | 0.77 | 96.23 | 0.24 | 60.08 |
| 19 | 0.75 | 96.99 | 0.21 | 60.29 |
| 20 | 0.72 | 97.70 | 0.15 | 60.45 |

coefficient with R of 0.760 (MSE of 0.093) for training set and 0.598 (MSE of 0.150) for prediction set.

For predicting antibacterial activity of oxazolidinone derivatives, descriptor selection is important in ANN modeling. The variables that produce the best model were chosen using the values of regression coefficients and VIP as previously mentioned [35–37]. After the PLS analysis of this descriptor space, it is noteworthy that there is no significant intercorrelation between these selected descriptors. Finally, nine descriptors were remained to predict activity with a nonlinear model. All of them appeared a relatively high regression coefficient (B) and a high value of VIP and no significant intercorrelation, including electronic energy (E_{ele}), heat of molecular formation (H_{form}), n -octanol/water partition ($\log P$), molecular dipole moment at z direction (DM_z), LC_9 , hydration energy (E_{hydr}), softness (S), most negative charge (MNC), chain of cycle terms of 5th order ($^5\chi_{\text{CH}}$).

To process the nonlinear relationships existing between the activity and the descriptors, ANN with BP learning algorithm was used, as described in the previous section. All networks were of the three-layered type, containing a bias neuron in each layer and a single neuron in the output layer. A sigmoid transfer function was employed in all neurons. The initial weights of network were randomly selected from a uniform distribution that ranged between -1 to 1 . These values were optimized during the network training. The values of each input was normalized into $[-1, 1]$, to bring the values of the input variables into the dynamic range of the sigmoid transfer function in the ANN. Learning rate η set at 0.02 in the start and momentum parameter set at 0.1.

The avoidance of overfitting and overtraining has been shown to be an important factor for improvement of generalization ability in neural network studies [39]. In the

present study, a subdivision of the initial training set of 106 compounds into a learning set ($n = 88$) and into a validation set ($n = 18$) was done. The first set was used to train the network, whereas the second set was used to monitor the training process. The optimal training endpoint and network architecture were determined on the basis of this validation set. The network architecture and training endpoint gave the lowest MSE, for the predictions of the validation set was then used. In order to study the effect of network parameters on its performances, networks with different configurations were built. To ensure that the results obtained were not due to chance, the predictions were repeated 1000 times with different initial weights in the network and the averaged pMIC values were calculated for each model. The network with seven neurons in the hidden layer gave the best performance, as given in table 4. A sufficient training level was not reached with smaller number of neurons (< 7) and overfitting exist with a larger number of neurons (> 7) in the hidden layer, respectively. The optimal training ANN training endpoint required 8500 training epochs when the ANN architecture 9-7-1 was used. Neural networks were able to calculate quite accurately. For all ANN models, the R_s in the training set range from 0.8594 to 0.8940, and in the prediction set range from 0.7853 to 0.8763, while the MSEs in the training set range from 0.041 to 0.0682, and in the prediction set range from 0.0607 to 0.0856. These results reflect the generalization performance and high predictive ability of the network.

Compared with the PLS analysis, the improved predictive performance was observed through ANN approaches. The results of the analysis for two methods, PLS and ANN, are summarized in table 3 and plots of the observed versus predicted values are shown in figure 2. It is obvious to see from figure that the ability of the PLS model to predict properly the activity of the data set is poorer in comparison with the ANN model. The model obtained with ANN has better correlation with experiment than that obtained with the PLS. Furthermore, the data reveals that the proposed model has higher prediction ability with lower relative errors than PLS model. In the case of PLS, the maximum relative error for predicted pMIC is 16.34% for number 21 compound, and the minimum value is 0.11% for number 56 compound. For ANN, the maximum relative error for predicted pMIC is

Table 4. The results for the resulted ANN models in training set and prediction set with different configurations.

| n_H^{\dagger} | MSE (training) | R^{\ddagger} (training) | MSE (prediction) | R (prediction) |
|-----------------|-------------------|------------------------------|---------------------|------------------|
| 4 | 0.0682 | 0.8661 | 0.0794 | 0.8021 |
| 5 | 0.0513 | 0.8730 | 0.0696 | 0.8553 |
| 6 | 0.0410 | 0.8857 | 0.0645 | 0.8532 |
| 7 | 0.0427 | 0.8940 | 0.0607 | 0.8763 |
| 8 | 0.0623 | 0.8594 | 0.0749 | 0.8331 |
| 9 | 0.0471 | 0.8774 | 0.0856 | 0.7853 |

[†] The number of the nodes in the hidden layer.

[‡] The correlation coefficient between the experimental and prediction values.

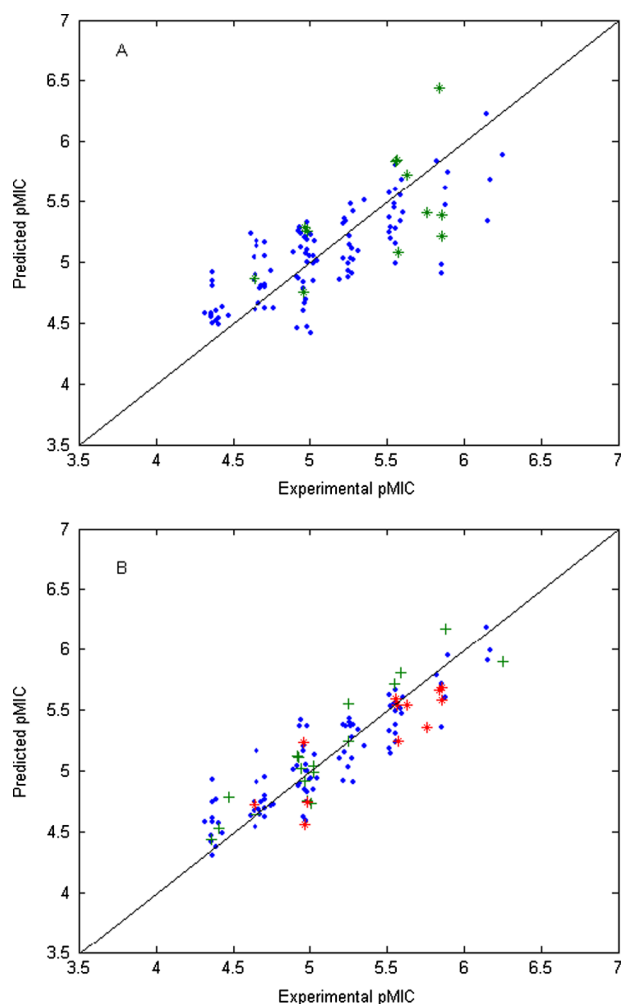


Figure 2. The plots of predicted activity by PLS (A) and ANN (B) against the experimental activity for the data used in the training set (•), validation set (+), and prediction set (*).

11.35% for number 8 compound, and the minimum value is 0.01% for number 40 compound. The average value of relative error between the calculated and experimental pMIC for PLS and ANN are 4.96 and 3.25%, respectively.

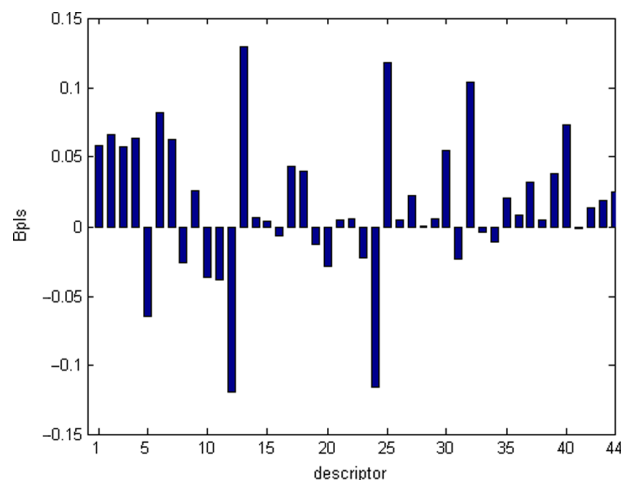


Figure 3. Weighted PLS regression coefficients (B_{PLS}) for pMIC. Numbers correspond to the descriptor variables.

Table 5. Comparison of the PLS and ANN models to estimate antibacterial activity of oxazolidinones.

| Model | Training set | | Prediction set | |
|------------------|--------------|--------|----------------|-------|
| | Mse | R | Mse | R |
| PLS | 0.093 | 0.7604 | 0.150 | 0.598 |
| ANN [†] | 0.043 | 0.894 | 0.061 | 0.876 |

[†]This ANN model has 7 nodes in the hidden layer.

A qualitative interpretation of the QSAR models can be made on the basis of the PLS weighted regression coefficients (B), as given in figure 3. For there are so many descriptors used in PLS modeling that we cannot list all of the descriptors, we have to use numbers as representation. The top five descriptors with the highest values are log *P*, hydration energy, softness, hardness and MNC, corresponding to the number 13, 12, 25, 24, 32. For ANN models, it is well known that the ANN can be envisaged as a nonlinear black box model [40,41], and it is not possible simply by inspection to determine the influence that one input variable has on one output variable. Therefore, ANN models have the advantage of giving accurate results, and this without requiring a formal model structure, but at the expense of a loss of model transparency.

4. Conclusions

The QSAR model PLS and ANN were employed to study the antibacterial activity of oxazolidinone derivatives. The goal of the project was to create QSAR models, which were both, interpretable as well as having good predictive ability. The linear regression model was found to be statistically valid, and the PLS routine enabled an investigation of the effects of each descriptor in the model.

The ANN models were found to be more successful than PLS analysis, reflecting that the relationship between descriptors and antibacterial activity of oxazolidinone derivatives is nonlinear. Usually, each one of the descriptors does not have clear correlation with the biological activity. The PLS method considers the inner relation among the independent variables to obtain good results of regression. Unfortunately, this method of regression does not suggest if the values of the descriptors should be high or low for maximizing the pharmacological potency. Therefore, PLS regression is a simple but powerful method to obtain a subset of significant input variables but it does not account for non-linear relationships.

References

- [1] B.H. Yagi, G.E. Zurenko. *In vitro* activity of linezolid and eperezolid, two novel oxazolidinone antimicrobial agents, against anaerobic bacteria. *Anaerobe*, **3**, 301 (1997).
- [2] G. Corti, R. Cinelli, F. Paradisi. Clinical and microbiologic efficacy and safety profile of linezolid, a new oxazolidinone antibiotic. *Int. J. Antimicrob. Agents*, **16**, 527 (2000).

- [3] B. Bozdogan, P.C. Appelbaum. Oxazolidinones: activity, mode of action, and mechanism of resistance. *Int. J. Antimicrob. Agents*, **23**, 113 (2004).
- [4] C.H. Park, D.R. Brittelli, C.L.J. Wang, F.D. Marsh, W.A. Gregory, M.A. Wuonola, R.J. McRipley, V.S. Eberly, A.M. Slee, M. Forbes. Antibacterials. Synthesis and structure-activity studies of 3-aryl-toxooxazolidines. 4. multiply-substituted aryl derivatives. *J. Med. Chem.*, **35**, 1156 (1992).
- [5] B. Das, S. Rudra, A. Yadav, A. Ray, A.V.S.R. Rao, A.S.S.V. Srinivas, A. Soni, S. Saini, S. Shukla, M. Pandya, P. Bhateja, S. Malhotra, T. Mathur, S.K. Arora, A. Rattan, A. Mehta. Synthesis and SAR of novel oxazolidinones: discovery of ranbezolid. *Bioorg. Med. Chem. Lett.*, **15**, 4261 (2005).
- [6] B.B. Lohray, V.B. Lohray, B.K. Srivastava, S. Gupta, M. Solanki, P. Kapadnis, V. Takale, P. Pandya. Oxazolidinone: search for highly potent antibacterial. *Bioorg. Med. Chem. Lett.*, **14**, 3139 (2004).
- [7] M.R. Barbachyn, C.W. Ford. Oxazolidinone structure-activity relationships leading to linezolid. *Angew. Chem. Int. Ed.*, **42**, 2010 (2003).
- [8] A.R. Renslo, P. Jaishankar, R. Venkatachalam, C. Hackbarth, S. Lopez, D.V. Patel, M.F. Gordeev. Conformational constraint in oxazolidinone antibacterials. synthesis and structure-activity studies of (azabicyclo[3.1.0]hexylphenyl) oxazolidinones. *J. Med. Chem.*, **48**, 5009 (2005).
- [9] A.R. Renslo, H. Gao, P. Jaishankar, R. Venkatachalam, M. Go'mez, J. Blais, M. Huband, J.V.N.V. Prasadb, M.F. Gordeev. Conformational constraint in oxazolidinone antibacterials. Part 2: synthesis and structure-activity studies of oxa-, aza-, and thiabicyclo[3.1.0]hexylphenyl oxazolidinones. *Bioorg. Med. Chem. Lett.*, **16**, 1126 (2006).
- [10] A.N. Pae, S.Y. Kim, H.Y. Kim, H.J. Joo, Y.S. Cho, K. Choi, J.H. Choi, H.Y. Koh. 3D QSAR studies on new oxazolidinone antibacterial agents by comparative molecular field analysis. *Bioorg. Med. Chem. Lett.*, **9**, 2685 (1999).
- [11] B. Gopalakrishnan, A. Khandelwal, S.N. Selvakumar, J. Das, S. Trehan, M.S. Kumarb. Three-dimensional quantitative structure-activity relationship (3D-QSAR) studies of tricyclic oxazolidinones as antibacterial agents. *Bioorg. Med. Chem.*, **11**, 2569 (2003).
- [12] D.T. Manallack, D.J. Livingstone. Neural networks in drug discovery: have they lived up to their promise? *Eur. J. Med. Chem.*, **34**, 195 (1999).
- [13] G. Schneider, P. Wrede. Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.*, **70**, 175 (1998).
- [14] K.L.E. Kaiser. The use of neural networks in QSARs for acute aquatic toxicological endpoints. *J. Mol. Struct. (Theochem)*, **622**, 85 (2003).
- [15] A.P. Borosy, K. Keseru, P. Matyus. Application of nonlinear and local modeling methods for 3D QSAR study of class I antiarrhythmics. *Chemometr. Intell. Lab. Syst.*, **54**, 107 (2000).
- [16] D.G. Arjona, G.L. Perez, A.G. Gonzalez. Non-linear QSAR modeling by using multilayer perceptron feedforward neural networks trained by back-propagation. *Talanta*, **56**, 79 (2002).
- [17] K. Hasegawa, T. Deushi, O. Yargashi, Y. Miyashita, S. Sasaki. Artificial neural network studies in quantitative structure-activity relationships of antifungal azoxy compounds. *Eur. J. Med. Chem.*, **30**, 569 (1995).
- [18] J.V. Turner, D.J. Maddalena, D.J. Cutler. Pharmacokinetic parameter prediction from drug structure using artificial neural networks. *Int. J. Pharma.*, **270**, 209 (2004).
- [19] M.J. Genin, D.A. Allwine, D.J. Anderson, M.R. Barbachyn, D.E. Emmert, S.A. Garmon, D.R. Graber, K.C. Grega, J.B. Hester, D.K. Hutchinson, J. Morris, R.J. Reischer, C.W. Ford, G.E. Zurenko, J.C. Hamel, R.D. Schaadt, D. Stapert, B.H. Yagi. Substituent effects on the antibacterial activity of nitrogen-carbon-linked (azolyphenyl) oxazolidinones with expanded activity against the fastidious gram-negative organisms *Haemophilus influenzae* and *Moraxella catarrhalis*. *J. Med. Chem.*, **43**, 953 (2000).
- [20] C.S. Lee, D.A. Allwine, M.R. Barbachyn, K.C. Grega, L.A. Dolak, C.W. Ford, R.M. Jensen, E.P. Seest, J.C. Hamel, R.D. Schaadt, D. Stapert, B.H. Yagi, G.E. Zurenko, M.J. Genin. Carbon-carbon-linked (pyrazolyphenyl) oxazolidinones with antibacterial activity against multiple drug resistant gram-positive and fastidious gram-negative bacteria. *Bioorg. Med. Chem.*, **9**, 3243 (2001).
- [21] L.M. Thomasco, R.C. Gadwood, E.A. Weaver, J.M. Ochoada, C.W. Ford, G.E. Zurenko, J.C. Hamel, D. Stapert, J.K. Moerman, R.D. Schaadt, B.H. Yagi. The synthesis and antibacterial activity of 1,3,4-phenyl oxazolidinone analogues. *Bioorg. Med. Chem. Lett.*, **13**, 4193 (2003).
- [22] M. Randic. On the characterization of molecular branching. *J. Am. Chem. Soc.*, **97**, 6609 (1975).
- [23] L.B. Kier, L.H. Hall. *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, NY (1976).
- [24] L.B. Kier, L.H. Hall. *Molecular Connectivity in Structure Activity Analysis*, John Wiley & Sons, New York, NY (1986).
- [25] H. Wiener. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, **69**, 17 (1947).
- [26] I. Gutman, B. Russic, N. Trinajstic, C.F. Wicox Jr. Graph theory and molecular orbitals. Part 12. Acyclic polyenes. *J. Chem. Phys.*, **62**, 3399 (1975).
- [27] I. Gutman, M. Randic. Algebraic characterization of skeletal branching. *Chem. Phys. Lett.*, **47**, 15 (1977).
- [28] A.T. Balaban. Topological index *J* for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)*, **21**, 115 (1986).
- [29] H. Hosoya. Topological index. A proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.*, **44**, 2332 (1971).
- [30] P. Thanikaivelan, V. Subramanian, J.R. Rao, B.U. Nair. Application of quantum chemical descriptor in quantitative structure activity and structure property relationship. *Chem. Phys. Lett.*, **323**, 59 (2000).
- [31] F.R. Burden. Robust QSAR models using bayesian regularized neural networks. *J. Med. Chem.*, **42**, 3183 (1999).
- [32] E.G. Borges, Y. Takahata. QSAR study of anti-ulcer compounds using calculated parameters. *J. Mol. Struct. (Theochem)*, **539**, 245 (2001).
- [33] L.O. Hansson, M.D. Ennis, P. Stjernlof. Quantitative structure-activity relationships in the 8-amino-6,7,8,9-tetrahydro-3H-benz[e]indole ring system. Analysis of serotonin 5-HT_{1A} effects *in vivo* and *in vitro* via partial least squares regression. *Eur. J. Med. Chem.*, **32**, 571 (1997).
- [34] N.A. Kratochwil, W. Huber, F. Muller, M. Kansy, P.R. Gerber. Predicting plasma protein binding of drugs: a new approach. *Biochem. Pharm.*, **64**, 1355 (2002).
- [35] S. Wold, M. Sjostrom, L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab. Syst.*, **58**, 109 (2001).
- [36] G. Chong, C.H. Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometr. Intell. Lab. Syst.*, **78**, 103 (2005).
- [37] S. Wold. In *Chemometric Methods in Molecular Design*, H. van de Waterbeemd (Ed.), Vol. 2, Chapter 4.4 pp. 195-218, VCH, Weinheim (1995).
- [38] D.E. Rummelhart, J.L. McClelland. *Parallel Distributed Processing, Volume I: Foundations*, MIT Press, Cambridge, MA.
- [39] I.V. Tetko, D.J. Livingstone, A.I. Luik. Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.*, **35**, 826 (1995).
- [40] M. Smith. *Neural Networks for Statistical Modeling*, Van Nostrand Reinhold, New York, USA (1994).
- [41] J.A. Anderson. *An Introduction to Neural Networks*, p. 650, MIT, Cambridge, MA (1995).