3

# Recent Applications of Quantitative Structure-Activity Relationships in Drug Design

Omar Deeb
*Al-Quds University, Faculty of Pharmacy, Jerusalem*
*Palestine*

## 1. Introduction

One of the most important challenges that face medicinal chemists today is the design of new drugs with improved properties and diminished side-effects for treating human disease such as AIDS and others. Medicinal chemists began the process by taking a lead structure and then finding analogs exhibiting the preferred biological activities. Next, they used their experience and chemical insight to eventually choose a nominee analog for further development. This process is difficult, expensive and took a long time. The conventional methods of drug discovery are now being supplemented by shortest approaches made possible by the accepting of the molecular processes involved in the original disease. In this view, the preliminary point in drug design is the molecular target which is receptor or enzyme in the body as an option of the existence of known lead structure.

The lock-and-key concepts at present are considered in drug design. Samples of protein targets were isolated and X-ray crystallography discovered their molecular structural design. Molecules are conceived either on the basis of similarities with recognized reference structures or on the basis of their complementarily with the three dimensional (3D) structure of well-known active sites.

The techniques currently on hand provide widespread insight into exact molecular features that are in charge for the regulation of biological processes: molecular geometries, electronic features and others. All these structural characteristics are of crucial importance in the understanding of structure-activity relationships and in rational drug design.

Rational drug design is based on the belief that the biological properties of drugs are related to their actual structural features. What has changed along the years is the way molecules are perceived and defined. In the past, medicinal chemists considered molecules as simple two-dimensional (2D) entities with related chemical and physicochemical properties. Quantitative structure-activity relationships (QSAR) concepts began to be considered and became very accepted.

However, most of these properties have not been well represented by the basic numerical parameters considered to characterize these features: the interactions between a ligand and a protein require much more information than the ones included in substituent indexes characterizing the molecular properties. Now, it has shown that consideration of the full detailed properties in 3D is necessary in allowing the understated stereochemical features to be respected.

The effective design of chemical structures with the desirable therapeutic properties is directed towards computer aided-drug design (CADD) a well established area of computer aided molecular design (CAMD). These techniques cover new methodologies, such as molecular modeling and quantitative structure-activity relationships (QSAR). Molecular modeling can be simply considered as a range of automated techniques based on theoretical chemistry methods and experimental data that can be used to predict molecular and biological properties.

The main applications of CAMD are the clarification of the basic requirements for a compound to obtain a determined activity, the simulation of the binding between a ligand and the receptor, the discovery of new active compounds and the prediction of activities for non-synthesised analogues. These applications convert CAMD to be used in drug design.

Computer aided molecular design (CAMD) is predictable to contribute to the discovery of "bright" molecules conceived on the basis of exact three-dimensional details. Two major modeling strategies right now used in the designing of new drugs. In the first strategy, the three-dimensional features of a known receptor site are directly considered whereas in the second strategy, the design is based on the comparative analysis of the structural features of known active and inactive molecules that are interpreted in terms of their complementarily with a supposed receptor site model.

The improvements in computer speed and capacity increased the number of lead compounds available for further research. But not only the number of feasible drug candidates increased, but also the costs and time devoted in various drug discovery processes was reduced, improving the effectiveness of the drug development.

One of the initial approaches to decrease these costs were attempted by correlating the biological function of a compound with its chemical structure, expressed in terms of molecular structural descriptors, by means of QSAR techniques. This discipline was promoted by Hansch and his group (Fujita, 1990). It was based on the determination of mathematical equations expressing the biological activities as a function of molecular parameters.

QSAR believe that the biological activity of a compound is a result of its chemical structure. Within the QSAR approach, the descriptor variable are not physically measured but computed, therefore, they are easy and cheap to generate even for large molecular sets.

## 2. Quantitative structure activity relationships (QSAR)

QSAR is a way of finding a simple equation that can be used to calculate some property from the molecular structure of a compound. QSAR attempt to correlate structural molecular features (descriptors) with physicochemical properties such as biological activities for a set of compounds, by means of statistical methods. As a result, a simple mathematical relationship is established.

Applications of QSAR can be extended to any molecular design purpose, including prediction of different kinds of biological activities, lead compound optimization and prediction of novel structural leads in drug discovery. The process of building a QSAR model is similar, apart from what type of property is being predicted. It consists of several steps which hopefully lead to the design of new compounds with the desired activity profile.

The first step in building a QSAR model is to select a training set of compounds with their experimental activities. Ideally, each of these activities should cover the range of possible values for that activity. The next step is to compute descriptors that contain sufficient relevant information about the biological phenomenon. However, it is difficult to predict in advance which descriptor variables will be valuable. Once descriptors have been calculated, it is necessary to pick which should be included in the QSAR model. A correlation coefficient gives a quantitative measure of how well each descriptor describes the activity. Thus, the descriptor with the highest correlation coefficient can be picked. Next step, a data analysis is needed to calculate the best mathematical expression linking together the descriptors and biological activities, in which information relating the essential features of the chemical and biological data structure is obtained. In the final step, validation and predictions for non-tested compounds will take place. However, the predictive capability of the model first is verified experimentally. This is talented by biological testing of some additional compounds (test set) in the same way as the training set and then comparing the experimental finding with the values predicted by the QSAR model. If the QSAR predicts within acceptable restrictions, it may be used for a more extensive prediction of more compounds. An interpretation of results should be done for the proposal and design of new compounds with the desired activity outline.

### 2.1 History of QSAR

Crum-Brown and Fraser expressed the suggestion that the physiological action of a substance was a function of its chemical composition. Later, in 1893, Richet showed that the cytotoxicities of a dissimilar set of uncomplicated organic compounds were inversely related to their corresponding water solubility. After that, Meyer and Overton independently recommended that the narcotic action of a group of organic molecules correlated with their olive oil/water partition coefficients. The extensive work of Albert, and Bell and Roblin established the importance of ionization of bases and weak acids in bacteriostatic activity. In the physical organic border, great progress was being made in the clarification of substituent effects on organic reactions, led by the influential job of Hammett. Taft invented a way for separating polar, steric, and resonance effects and introducing the first steric parameter, *ES.*

The contributions of Hammett and Taft together laid the mechanistic source for the progress of the QSAR model by Hansch and Fujita. In 1962 Hansch et al. (Hansch et al., 1962) published their bright study on the structure-activity relationships of plant growth regulators and their dependency on Hammett constants and hydrophobicity. A Linear Free Energy Relationships (LFER) related model published by Fujita et al. and Hansch et al., (Fujita et. al., 1964, Hansch et. al., 1964) considered to be the official beginning for QSAR. Their fragment and additive group contribution idea added two things: the use of calculated properties to correlate with biological activities and the detection that multiple properties may influence the biological activity. For this purpose, they implemented the use of the computer to fit QSAR equations.

The so-called Hansch equation (Hansch, 1969) was developed to correlate physicochemical properties (descriptors) with biological activities is given in a general form by:

$$\log 1/C = a (\log P)^2 + b\log P + c\sigma + \ldots k \tag{1}$$

where $C$ is the molar concentration that produces the biological effect; $P$ is the octanol/water partition coefficient and $\sigma$ is the electronic Hammett constant.

Besides the Hansch approach, other methodologies were also developed to deal with structure- activity questions. The Free-Wilson approach (Free and Wilson, 1964) addresses structure-activity studies in a congeneric series in which the contribution of each structural feature was a parameter of interest. These parameters, also called indicator variables, codify the presence or absence of particular structural feature. They are assigned the binary values of 1 and 0, accordingly.

### 2.2 Descriptors

A common question in QSAR is how to describe molecules and their physicochemical properties (descriptors). The nature of the descriptors used and the extent to which they instruct the structural properties related to the biological activity is a critical part of a QSAR study (Downs, 2004). It has been estimated that thousands of molecular descriptors are now existing (Devillers and Balaban 1999; Karelson, 2000; Todeschini et. al., 2002). Most of them can be calculated by using commercial software packages such as CODESSA (Katritzky et. al., 2002), DRAGON (Todeschini et. al., 2002) and others. The various descriptors in use can be largely categorized as being constitutional, topological, electrostatic, geometrical, or quantum chemical.

Constitutional descriptors give a simple description of what is in the molecule. For example, the number of heteroatoms, the number of rings, the number of double bonds, etc. Constitutional descriptors often appear in a QSAR equation when the property being predicted varies with the size of the molecule.

Topological descriptors are numbers that give information about the bonding collection in a molecule. They are derived from graph representation of chemical structures; they attempt to encode the size, shape, or branching in the compound by handling of graph-theoretical aspects of the structures (Silipo and Vittoria, 1990). Some examples are Randic indices, Kier and Hall indices, Weiner index (sum of the chemical bonds existing between all pairs of heavy atoms in the molecule), the connectivity index and others.

Electrostatic descriptors are single values that give information about the molecular charge division. Some examples are polarity indices and polarizability. One of the most commonly used electrostatic descriptors is the topological polar surface area (TPSA), which gives an indication of the portion of the molecular surface composed of polar groups against nonpolar groups. Another deeply used descriptor is the octanol–water partition coefficient, which is designated by a specific prediction scheme such as ClogP or MlogP.

Geometrical descriptors are single values that describe the molecule's size and shape as well as the degree of complementarity of a ligand and the receptor. They are developed from three-dimensional models of molecules, and derived from molecular surface area calculations. Some examples are moments of inertia, molecular volume molecular surface area, and other parameters that describe length, height, and width.

Quantum chemical descriptors give information about the electronic structure of the molecule. They are obtained by molecular orbital calculations and they mainly describe electronic interaction. These includes, the energy of the highest occupied molecular orbital, $E_{HOMO}$, which is a quantitative measure for the chemical reactivity of the compound-ionization potential of a molecule, the energy of the lowest unoccupied molecular orbital, $E_{LUMO}$, which accounts for the electron affinity, refractivity, and total energy. The $E_{HOMO}$–$E_{LUMO}$ gap or ionization potential can be important descriptors for predicting how molecules will react.

New nodal angle quantum descriptors - the Frontier Orbital Phase Angles - suggested by Clare (Clare, 1998) which considered as novel QSAR descriptors for benzene derivatives will be discussed in the application part.

### 2.3 Statistical methods

Statistical methods are the mathematical basis for the development of QSAR models. Chemometric methods (Eriksson et al., 2001) are used to extract information from QSAR data using tools of statistics and mathematics. The applications of these methods are combined with the important goal of explanation and prediction of non-synthesised test compounds. Many different statistical methods are available in the literature and the selection of the appropriate method is critical (Xu and Zhang, 2001).

**Multiple Linear Regression (MLR)** (Montgomery and Peck, 1992) can be considered as an easy interpretable regression-based method. Regression analysis correlates independent X variables or descriptors **(**physicochemical parameters) with dependent Y variables **(**biological data). The regression model assumes a linear relationship between *m* molecular descriptors and the response (biological activity) variable. This relationship can be expressed with the single multiple-term linear equation:

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_mX_m + e \qquad (2)$$

The MLR analysis calculates the regression coefficients, $b_i$, by minimizing the residuals, **e**, which quantify the deviations between the data (Y) and the model (Y'), as in the case of simple linear regression.

**Partial Least Squares (PLS)** (World et al., 1993) which in turn decrease the information content of data matrices. It projects multivariate data into a space of lower size, and certainly providing insight to see and model huge sets of data. The Partial Least Squares (PLS) regression method carries out regression using latent variables from the independent and dependent data that are along their axes of most variation and are highly correlated. It is applied when the numbers of independent variables are more than the number of observations. Under these circumstances, it gives a more strong QSAR equation than multiple linear regressions. Thus, PLS is able to examine complex structure-activity problems and to examine data in a more realistic way. PLS gives a condensed statistically strong solution and, in fact, it contains MLR as a special case when a MLR be present.

Another way to reduce the dimensionality of the data set descriptors X is the so called **Principal Component Analysis** (PCA) technique (Jolliffe, 1986). It seeks to find out a new set of variables named Principal Components (PC) showing the data in order of decreasing variance with the aim to state the main information in the variables by the principal components of X**.** The primary Principal Component (PC1) describes the maximum deviation in the whole data set. The subsequent principal component (PC2) describes the maximum remaining variance, and so forth, with each axis linearly independent, to the preceding axis. Some of the last components may be discarded to decrease the size of the model and stay away from over-fitting.

The **Principal Components Regression (PCR)** method uses linear regression to generate a model by means of the principal components as independent descriptors. PCR applies the scores from PCA as regressors in the QSAR model. Therefore, a multiple-term linear equation is generated and derived from a principal components analysis transformation of the independent variables.

**Artificial Neural Networks (ANN)** method (Tetko, 1996; Novi et al., 1997; Duprat et al., 1998) is non-linear technique inspired in the human brain, composed of many simple processing units called neurons. This method is also recognized as learning algorithms. The aim is to simulate the various shells of the neurones, where each neuron is connected to a number of neighbouring neurones with variable coefficients of connectivity that signify the strength of these associations. The learning process consists of adjusting the coefficient so that the network provided as an output the suitable results. In neural networks, a training set is used to train the network, and then the network is used to predict the property (biological activity) that it was trained to predict. This technique can be associated with principal components analysis in which it is referred as **PC-ANN.**

**Support Vector Machine (SVM)** can be applied to regression by the introduction of an alternative loss function. In support vector regression (SVR) (Gunn, 1997), the basic idea is to map the data X into a higher-dimensional feature space via a nonlinear mapping and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set.

### 2.4 Validation of QSAR models

After the model equation is obtained, moreover the stability and the goodness of fit of the model, it is also significant to estimate the power and the validity of the model before using it to predict the biological activity. Validity is to establish the reliability and significance of the method for a particular use. Therefore, validation of a QSAR model must be done. There are two validation methods used for a QSAR model: internal and external validation techniques to establish the confidence and strength of the model.

#### 2.4.1 Internal validation

Internal validation uses the dataset from which the model is built and checks for internal stability. **Cross-Validation** (**CV**) **technique** is widely employed as an internal validation method of statistical models (Allen, 1974; World, 1978, 1991). Usually, one compound of the set is extracted each time, and then the model is recalculated using as training set the *n-1* (where n is number of compounds) remaining compounds, so that the biological activity value for the extracted compound is predicted once for all compounds. This process is repeated *n* times for all the compounds of the initial set, thus obtaining a prediction for each object. This process referred as leave-**one-out (LOO**) method. Also an alternative method can be defined when leaving out more than a compound of the data set at each time. This method is called **leave n-out** or **Leave-many-Out (LMO) CV method or sometimes it is referred as leave-group-out (LGO)**. Calculation of the correlation coefficient of the cross-validation procedure, that is, the **coefficient of prediction** $q^2$ must be done and it is by definition smaller or equal than the overall $r^2$ (correlation coefficient) for a QSAR equation. It is used as an investigative tool to estimate the predictive power of an equation obtained by using a regression method. Another procedure to test the validity of the model is the **randomization test.** Even with a huge number of compounds and a small number of descriptors, an equation can still have very poor predictive power. One way to test for this is by **randomization** of the compounds. The set of biological activity values is re-assigned arbitrarily to different compounds, and a new regression is done. This process is repeated many times. If the random models' biological activity prediction is analogous to the original equation within a given estimated self-confidence level, this means that the original model was obtained by chance. The random test analyses the ability of the model to derive actual structure-activity relationships.

### 2.4.2 External validation

A QSAR model with excellent goodness of fit and acceptable predictions may be deficient in real relationship between structural descriptors and biological activity. The perfect validity of the model is examined by **external validation**, which evaluates how well the model generalizes. If a sufficiently huge series of compounds with known activity is obtainable, the original data set can be divided into two subgroups, the **training set** and the **test set.** The training or calibration set is used to derive a calibration model that will be used later to predict the activities of the test or validation set compounds. On the other hand, an external test set that has not been included in any stage of the building of the model can be used as **test set.**

The obtained predictions of the new generated model for the test set determine the validity of the model. The parameters quantifying the superiority of prediction of the external test set may be the same used for the internal validation. The Sum of Squares Prediction Errors (SSPE) is extensively used to account for the inconsistency.

## 3. Recent applications of QSAR in drug design

### 3.1 Nodal angle quantum descriptors and flip regression

When implementing QSAR on flat, symmetrical, usually aromatic molecules, symmetry considerations often affirm that alternative orientations should be inspected. For phenethylamines, for example, there are five substitution sites on the benzene nucleus. If substituents with property $P_i$ were introduced to site i (i = 2-6), an equation may be formulated:

$$log\ A = \sum_i C_i P_i + C_0 \tag{3}$$

where A is activity and $C_i$ are constants to be determined by regression techniques (Clare and Supuran, 2005a). Hence, both 2-methoxyphenethylamine and 6-methoxyphenethylamine may be predicted to have the same activity for both molecules. Apparently, it may appear that $C_2$ must equal $C_6$ and $C_3$ must equal $C_5$, but it can be shown that this is not the case when considering 2,3,4-trimethoxyamphetamine and 2,4,5-trimethoxyamphetamine. Equating the C values would predict the same activity for both substances, but experimentally one of these is a potent hallucinogen and the other is inactive (Shulgin et al., 1991).

At the molecular level, the flat aromatic molecule may lay in two ways on the receptor, corresponding to the 5- or 6-membered rings swapping positions, or flipping. All combinations of each drug flipped and unflipped must be considered. In the absence of structural data the only way in which we can proceed is to carry out regressions with every combination of each drug in both orientations and find which regression fits best. For the case of N drugs, $2^N$ regressions must be considered. A full treatment is possible only to the smallest groups of compounds, so the approach used is to employ simulated annealing as a method of combinatorial optimization. This problem was first addressed by Kishida and Manabe (Kishida and Manabe, 1980), in perspective of QSAR of substituted benzenedisulfonamides. In a study by Clare (Clare, 1998), a descriptor for QSAR calculations on benzene derivatives was proposed, and shown to be highly effective in correlating activities in humans of a large class of phenylalkylamine hallucinogens.

Moreover, in a number of studies (Clare, 1998; Clare, 2000; Clare, 2001; Clare, 2002; Clare and Supuran, 2004), it has been shown that a small number of descriptors can account for the activity of diverse aromatic drugs, and a method for dealing with the symmetry nature in some groups of planar aromatic molecules has also been outlined. Particularly, it has been verified that in most cases the orientations of nodes in π -like orbitals of aromatic molecules are a significantly important feature in understanding their activity. This was first established in phenylalkylamine hallucinogens (Clare, 1998), and then also in benzenoid and heteroaromatic carbonic anhydrase, trypsin, thrombin and bacterial collagenase inhibitors (Clare and Supuran, 2004), as well as in tryptamine hallucinogens (Clare, 2004).

The descriptors are based on the similarity of the frontier orbitals of the molecule in question to those of benzene and involves an analytical least squares fitting of the molecules frontier orbitals, calculated by any semiempirical or ab initio method to those carefully calculated for unsubstituted benzene. Both the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) of benzene are degenerate, and each consists of two components that may be mixed in any proportion with normalization to form an infinity of equally acceptable frontier orbitals. In benzene itself, each of these mixtures is equivalent. When the benzene is substituted, the degeneracy is lifted, and each of the resulting separate orbitals may be considered as being approximately derived from one particular linear combination of the original two components.

The significance of orbital symmetry in the interactions of atoms to form molecules has been known for a long time. It appears that this is directly transferable to the association of molecules in pharmacology, at least insofar as π orbitals are involved. Many QSAR studies on aromatic molecules have involved the HOMO and LUMO energies or their sum or difference as descriptors. Consideration of the nodal angles, especially if the aromatic moiety is benzene could profit any of these studies. The π-like orbitals involved are standing waves of probability of finding an electron in a given location in the field of the atomic nuclei, and have no classical counterpart. Therefore, the dependence of activity on these variables (Clare, 2000; Clare, 2001) is perhaps the best indication yet of the essential quantum mechanical nature of drug-receptor interactions. Conventional 3D-QSAR programs, which employ classical interactions, such as coulombic charge-charge forces and empirical van der Waals interactions, may benefit from the incorporation of π orbital wave mechanical interactions such as those discussed in (Clare, 2000; Clare, 2001).

The calculation of nodal orientation is performed with the program NODANGLE (Clare and Supuran, 2005b). NODANGLE calculates the angle between the nodes in π -like orbitals and a reference point on the aromatic ring. NODANGLE works by comparing the coefficients of the $p_z$ atomic orbitals on a 5- or 6-membered ring with those of the cyclopentadienide anion (for a 5-membered ring) or the benzene molecule (for a 6-membered ring), of known nodal orientation.

### 3.1.1 QSAR of protein tyrosine kinase inhibitory activity of flavonoid analogues

Flavonoids are a group of low molecular weight plant (Wang and Wang, 2002; Cronin et al, 1998) products, based on the parent compound, flavone (2-phenylchromone) and have shown potential for application in a variety of pharmacological targets. A large number of natural and synthetic flavonoids are being tested as specific inhibitors of protein tyrosine kinase (PTK). The flavonoid-inhibitory activity is expressed as log 1 / $IC_{50}$, which is the molar concentration of the flavonoid necessary to give half-maximal inhibition compared to the control assay carried out in the absence of inhibitor, but in the presence of dimethyl sulphoxide carrier. Clare and Deeb in (Deeb and Clare, 2007) have investigated the flavonoid-inhibitory activity of 54 analogues using the nodal angle descriptors (Clare, 2000; Clare and Supuran, 2005b) and flipstep regression analysis (Clare, 2000; Clare, 2001) mentioned above.

For the flavonoid, calculating the angles in the three rings can be accomplished by entering the atom as numbered in Figure 1. The three rings are 6-membered rings numbered 1–6, 5–10 and 11-16 for rings 1, 2 and 3 respectively. The angles calculated by NODANGLE (Clare and Supuran, 2005b) are then $\Theta_1$, $\Theta_2$ and $\Theta_3$ in the figure, measured at atoms 1, 5 and 11 respectively. A problem arises from the symmetry of the parent molecule; therefore, the

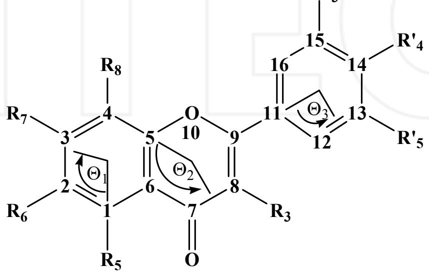


Fig. 1. Numbering of flavonoids skeleton used in MOPAC and NODANGLE calculations and angles used in the interpretations. The angle subscript indicates the ring number.

program FLIPSTEP, a component of the MARTHA statistical package (Clare and Supuran, 2005a) was used. This set of flavonoids separates into two parts (symmetry wise): the chromone moiety and the phenyl ring. The chromone ring system has no vertical mirror planes or axes. Hence, ring 1 cannot be flipped into ring 2. Thus for this part of the molecule flip regression is not applied. The phenyl ring has $C_{2v}$ symmetry, so flip regression is applicable to this. Thus only ring 3 should be flipped.

In the study carried out by Deeb and Clare (Deeb and Clare, 2007), it was demonstrated that the charge on $O_{10}$ proved to be the most important factor. Low charge on $O_{10}$ was found to be favourable to high activity. Furthermore, it was found that the charge on $C_7$ and the mean of absolute charge are significant variables. Moreover, it was shown that the orientation of the nodes on ring 3 are significant factors which indicate the importance of the electrostatic and quantum chemical descriptors for the interaction of flavonoids with the specific enzymatic active site plays an important role. Exactly which rings are involved becomes clear from the identity of the descriptors included in the regression equation:

$$\begin{aligned}\log 1/IC_{50} = {} & 7.4417\ (\pm 2.0652) - 0.65265\ (\pm 0.1569) \times HOP1 + 0.81601\ (\pm 0.1348) \times SHOP1 \\ & + 0.35316\ (\pm 0.1422) \times HOP3 - 0.32482\ (\pm 0.0974) \times S2\Theta 1H - 0.21638\ (\pm 0.0778) \times C2\Theta 1H \\ & - 0.00235\ (\pm 0.0008) \times Vol - 14.69500\ (\pm 6.1426) \times QC_7 - 27.77900\ (\pm 5.1297) \times QO_{10} \\ & + 14.14800\ (\pm 3.0611) \times Q_{mean} + 0.46973\ (\pm 0.1042) \times C2\Theta 3H + 1.57150\ (\pm 0.1779) \times S2\Theta 3H \\ & - 0.26624\ (\pm 0.0889) \times C4\Theta 3L - 12.75700\ (\pm 1.6295) \times S4\Theta 3L \end{aligned} \tag{4}$$

$$N = 54,\ R^2 = 0.8240,\ F = 14.403,\ S = 0.30537,\ Q^2 = 0.6612$$

where HOP1 is the highest occupied π orbital on ring 1, SHOP1 is the second highest occupied π orbital on ring 1, HOP3 is the highest occupied π orbital on ring 3, S2Θ1H is sin(2× the nodal angle in the highest occupied π orbital in ring 1), C2Θ1H is cos(2× the nodal angle in the highest occupied π orbital in ring 1), Vol is molecule volume, $QC_7$ is charge on $C_{7,}$ $QO_{10}$ is charge on $O_{10}$, $Q_{mean}$ is the mean absolute Mulliken charge, C2Θ3H is cos(2× the nodal angle in the highest occupied π orbital in ring 3), S2Θ3H is sin(2× the nodal angle in the highest occupied π orbital in ring 3), C4Θ3L is cos(4× the nodal angle in the lowest unoccupied π orbital in ring 3), S4Θ3L is sin (4× the nodal angle in the lowest unoccupied π orbital in ring 3), N is number of compounds, $R^2$ is the coefficient of determination, F is Fisher variance ratio, S is standard deviation and $Q^2$ is the square of the multiple correlation coefficients based on the leave-one-out residuals.The numbers in parentheses are the standard errors.

The work of Deeb and Clare (Deeb and Clare, 2007) demonstrated that the nodal orientation terms have a powerful explanatory importance in that they account for more of the variance in activity than is possible using the classical descriptors alone. However, a combination of the classical descriptors and the nodal orientation term gives even better explanatory of activity of the flavone analogues. The chromone moiety of the flavonoid structure is envisaged to be a mixed region for hydrophobic and electronic interactions, while the phenyl ring moiety, especially the substituents at the 3' and 4' position, are involved in electronic interactions with the enzyme. S4Θ3L, that is cos (4 × the nodal angle in the lowest unoccupied π orbital in ring 3), was identified to be an important descriptor.

### 3.1.2 QSAR of EGFR inhibitory activity of quinazoline analogues

Epidermal growth factor receptor (EGFR) that has been identified as a kind of PTK and has been demonstrated to be related to many human cancers such as breast and liver cancers, leading many to believe that EGFR is an attractive target for anti-tumor drug discovery (Yang et al., 2001).

Deeb and Clare in (Deeb and Clare, 2008a) have applied the flip regression procedure applied on classical and quantum nodal oreinetation angles descriptors to investigate the quinazoline-inhibitory activity of 63 analogues expressed as log $IC_{50}$. $IC_{50}$ is the effective concentration of the compound required to inhibit by 50% the phosphorylation of a 14-residue fragment of phosphorylase $C_{\gamma\text{-}1}$ (prepared from A431 human epidermoid carcinoma cells through immunoaffinity chromatography) by EGFR. For the quinazoline, calculating the angles in the three rings can be accomplished by entering the atom as numbered in Figure 2(a). The three rings are 6-membered rings numbered 1–6, 5–10 for ring 1 and 2, respectively. Ring 3 is also a 6-membered ring numbered 12–17. The angles calculated by NODANGLE are then $\Theta_1$, $\Theta_2$ and $\Theta_3$ in that figure, measured at atoms 1, 5 and 12 respectively.

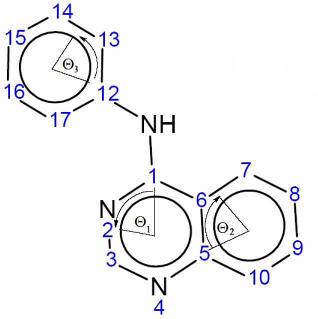


Fig. 2. (a) Numbering of quinazolines skeleton and angles used in the interpretations. The angle subscription indicates the ring number

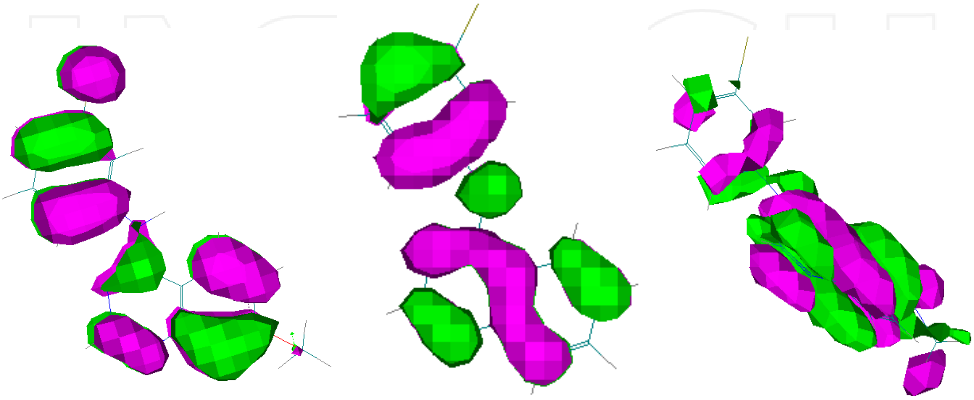


Fig. 2. (b) HOMO orbitals for quinazolines.

In this study (Deeb and Clare, 2008a) it is shown that the benzene rings of the quinazolines are interacting with aromatic systems on the receptor and that alignment occurs between the π orbital nodes on the pair. Exactly which rings are involved becomes clear from the identity of the descriptors included in the regression equation:

$$\begin{aligned} \log 1/IC_{50} = &-8.7912\,(2.80) - 0.30044\,(3.36)\,\text{SHOP2} \\ &+ 1.1489\,(7.10)\,\text{LUP3} \\ &- 1.2589\,(8.19)\,\text{SLUP3} \\ &- 0.47012\,(6.30)\,\text{C2}\Theta\text{1H} - 0.84000\,(7.73)\,\text{C4}\Theta\text{1L} \\ &- 0.27740\,(3.07)\,\text{S2}\Theta\text{2H} + 0.77819\,(7.80)\,P_{xx} \\ &- 0.55845\,(6.55)\,P_{zz} \\ &- 0.17269\,(3.34)\,\text{C2}\Theta\text{3H} + 0.03004\,(0.57)\,\text{S2}\Theta\text{3H} \\ &- 0.03340\,(0.51)\,\text{C4}\Theta\text{3L} + 0.5871\,(11.27)\,\text{S4}\Theta\text{3L} \qquad (5) \end{aligned}$$

$$N = 63,\ S = 0.49766,\ F = 38.21,\ R^2 = 0.9017,\ Q^2 = 0.8550$$

where SHOP2 is the second highest occupied π orbital on ring 2 (see Figure 2(b)), LUP3 is the lowest unoccupied π orbial in ring 3, SLUP3 is the second lowest unoccupied occupied π orbital on ring 3, C4Θ1L is cos(4× the nodal angle in the lowest unoccupied π orbital in ring 1), $P_{xx}$ is diagonal components of polarizability in x-direction, $P_{zz}$ is diagonal components of polarizability in z-direction and S2Θ2H is sin(2× the nodal angle in the highest occupied π orbital in ring 2). The numbers in parentheses are student's *t* values.

Equation (5) shows that the second lowest unoccupied π orbital on ring 3 was identified to be an important descriptor. The only classical variables found to be significant were the polarizability components. High polarizability in the highest inertia direction was found to be favorable to high activity, while high polarizability in the lowest inertia direction was detrimental.

### 3.1.3 QSAR of phenylisopropylamines MAO-inhibition: Comparison of AM1 and B3LYP-DFT

Monoamine oxidase plays a critical role in the regulation of monoamine neurotransmitters such as serotonin, noradrenaline, and dopamine. MAO isoenzymes are classified on the basis of their substrate preference, sensitivity toward specific inhibitors, and tissue distribution into MAO-A and MAO-B. Selective MAO-A inhibitors have been used clinically in the treatment of depression and anxiety, while MAO-B inhibitors have been used in the treatment of Parkinson's and Alzheimer's diseases. Many plant-derived and synthetic compounds such as isoquinolines and xanthones have been identified as MAO inhibitors. In (Deeb and clare, 2008b) the monoamine oxidase (MAO)-inhibitory activity of 46 phenylisopropylamines expressed as pIC50 is modeled with the orientations of nodes in π - like orbitals of the phenyl ring and some other descriptors using flip regression analysis. The authors aim to provide an initial clue regarding the scope and limitations of some state-of-the-art methods in computational chemistry, including semiempirical (AM1) and density functional theory (B3LYP), in the flip regression procedure applied to the inhibition of phenylisopropylamines.

Calculating the angles for the aromatic ring in the each phenylisopropylamines can be accomplished by entering the atoms as numbered in Figure 3. The ring is six-membered and is numbered 1 to 6. The angle calculated by NODANGLE is then Θ measured at atom 1. The flip regression is applicable to the phenyl ring of $C_{2v}$ symmetry.

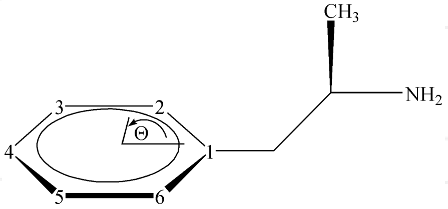


Fig. 3. Ring substitution pattern of the phenylisopropylamines.

Again, it was found that a combination of the classical descriptors and the nodal orientation terms gives better explanatory of activity of the phenylisopropylamines as it can be seen in the following regression equation:

$$\begin{aligned} pIC_{50} = {} & 14.961\ (\pm 2.2558) + 1.2957\ (\pm 0.28952)\ SHOP \\ & - 3.957\ (\pm 0.34587)\ LUP - 7.4769\ (\pm 2.52)\ LDI \\ & - 1.8473\ (\pm 0.22595)\ C2\Theta H \\ & - 1.8104\ (\pm 0.18915)\ S2\Theta H \\ & + 0.75002\ (\pm 0.10146)\ C4\Theta L \\ & + 1.1693\ (\pm 0.13404)\ S4\Theta L \end{aligned} \tag{6}$$

$$N = 32,\ R^2 = 0.9309,\ F = 46.222,\ S = 0.28690,\ Q^2 = 0.8587$$

where SHOP is the second highest occupied π orbital, LUP is the lowest unoccupied π orbital, LDI is the local dipole index, C2ΘH is cos(2* the nodal angle in the highest occupied π orbital), S2ΘH is sin(2* the nodal angle in the highest occupied π orbital) and S4ΘL is sin(4* the nodal angle in the lowest unoccupied π orbital).The numbers in parentheses are the standard errors.

From equation (6), it can be predicted that the phenyl moiety of phenylisopropylamines is involved in electronic interactions with the enzyme. Lowest unoccupied π energy was identified to be an important descriptor. Adding classical variables, improves the correlation. The classical variables found to be significant are LUP and LDI which were found to be favorable to high activity. This is based on the concept that the stability of stacked aromatic systems is highly orientation-dependent, and is also dependent on the energies of those orbitals in the two aromatic systems that resemble the degenerate HOMO and LUMO of benzene. Furthermore, the results show that the models established based on DFT-B3LYP method are better than those based on AM1 method. The B3LYP model gives more reasonable interpretation of phenylisopropylamines MAO inhibition activity.

## 3.2 Applications of QSAR using PC-ANN

### 3.2.1 Correlation ranking and stepwise regression procedures in principal components artificial neural networks modeling with application to predict toxic activity and human serum albumin binding affinity

A successful drug should be able to reach its target without generating toxic effects in addition to possessing intrinsic activity. Considering the substantial failure rate of drug candidates in late stage development and the expensive and time-consuming process of measuring toxic effects, predictive tools that eliminate inappropriate compounds become necessary. Prediction of toxicity from the structure of compounds can help in designing the new beneficial compounds and hence, screening of large number of chemicals for toxic effects as well as interpreting the mechanisms of toxicity. Development of QSARs relating toxicity potency and structural properties can be an alternative that has the advantage of high speed and low costs in comparison with experiments. Because most toxicology predictions engage a diverse set of compounds belonging to different classes and multiple toxic mechanisms, some nonlinear relations between the properties of compounds and their toxicity parameters are expected and linear regression approaches may not be accurate and can lead to imprecision.

Human serum albumin (HSA) is the most abundant protein in plasma. It is characterized by its surprising capacity to bind a large variety of drugs. Extensive biochemical studies resulted in the proposition of two main drug-binding sites in HSA, denoted as I and II. Site I was shown to prefer large heterocyclic and negatively charged compounds, while site II was the one for small aromatic carboxylic acids. When the crystal structures of HSA with ligands were available, these sites were localized at subdomains IIA and IIIA. Analytical techniques have been employed to measure drug-binding affinities to HSA. These techniques have low throughput and they require relatively large quantities of both drug and protein. The recent development of high-performance affinity chromatography (HPAC) columns with immobilized HSA has allowed the medium-throughput determination of drug binding to this protein in a way that requires small amounts of both drug and HSA. Developing a model for predicting the drug-binding affinity based on molecular structure is very important goal for medicinal chemist. Several studies aim to generate models that predict drug-binding affinities to HSA such as QSAR and molecular modeling.

Two data sets were used in this study (Deeb, 2010). The first was an extensive toxicity data set that contains 278 substituted benzenes (Feng et al., 2003). The logarithm of half maximal inhibitory concentration (log $IC_{50}$) toxicity to T. pyriformis is used as the toxicity end point. Another data set of 95 HSA drug and drug-like compounds and their binding affinities were reported by Colmenarejo (Colmenarejo, 2003). The 3D molecular structures of the compounds were optimized by Hyperchem software using the semiempirical AM1 level of theory. In this study, a total of 1233 and 698 molecular descriptors were calculated for each molecule of the substituted benzenes and HSA drug and drug-like compounds, respectively. These descriptors are belonging to 17 different types of theoretical descriptors (Table 1).

Dragon software was used to calculate 1217 and 684 descriptors gathered into 16 groups for the toxic compounds and HSA drug and drug-like compounds, respectively. A group of 16 and 14 quantum descriptors for the toxic and HSA drug and drug-like compounds, respectively, describing the electronic properties of molecules were calculated by Hyperchem software. SPSS Software was used for the simple principal component regression (PCR) analysis. PCA and ANN regressions were performed in the MATLAB environment.

| J * | Descriptors | No. of descriptors calculated for toxicity of substituted benzenes | No. of descriptors calculated for HSA binding affinity |
|---|---|---|---|
| 1 | Quantum descriptors | 16 | 14 |
| 2 | Constitutional descriptors | 34 | 31 |
| 3 | Topological descriptors | 228 | 89 |
| 4 | Molecular walk counts | 15 | 7 |
| 5 | Burden eigenvalue (BCUT) descriptors | 64 | 18 |
| 6 | Galvz topological charge indices 21 | 21 | 21 |
| 7 | 2D autocorrelations | 96 | 68 |
| 8 | Charge descriptors | 14 | 8 |
| 9 | Aromaticity indices | 4 | 0 |
| 10 | Randic molecular profiles | 41 | 5 |
| 11 | Geometrical descriptors | 38 | 29 |
| 12 | RDF descriptors | 142 | 51 |
| 13 | 3D-MoRSE descriptors | 160 | 95 |
| 14 | WHIM descriptors | 99 | 30 |
| 15 | GETAWAY descriptors | 197 | 131 |
| 16 | Functional group counts | 27 | 40 |
| 17 | Atom-centred fragments | 37 | 61 |

* J is the index of the group of descriptors.

Table 1. Types of descriptors used in this study.

Aiming to test the final model performances, the data set was divided into training (60%), validation (20%) and prediction (20%) sets based on descriptor spaces. For this purpose, the data matrix containing the total descriptors was subjected to PCA and the first two PCs were plotted against each other. PCA was run twice, once by grouping of descriptors and analysis of each group separately. This approach is referred to as the individual PCA approach, PCA(I). And once, by analysis of the entire set of calculated descriptors simultaneously. This approach is referred as the combined PCA approach, PCA(C). In PCA(I) procedure, each group of descriptors was subjected to PCA separately and the subset of PCs that explained 95% of the variances in the original descriptors data matrix were extracted from each set. The PCs extracted from this approach are named in the form "PCi–j" where "i" indicates the descriptors group and "j" indicates the PC number in the ith group which is related to its ranked eigenvalue. In a similar manner, in PCA(C) procedure all calculated PCs were collected in a single data matrix and the PCs were extracted. A feed-forward neural network with back-propagation of an error algorithm was constructed to model the structure–activity relationship. Our network has one input layer, one hidden layer, and one output layer. The input vectors were the set of PCs, selected according to PCA(I) and PCA(C) in combination with stepwise regression (SR) and correlation ranking (CR) procedures. The number of nodes in the input layer is dependent on the number of PCs introduced in the network. The number of nodes in the hidden layer is optimized through a learning procedure. Fgure 4 illustrates the four ANN analyses carried out in this study.

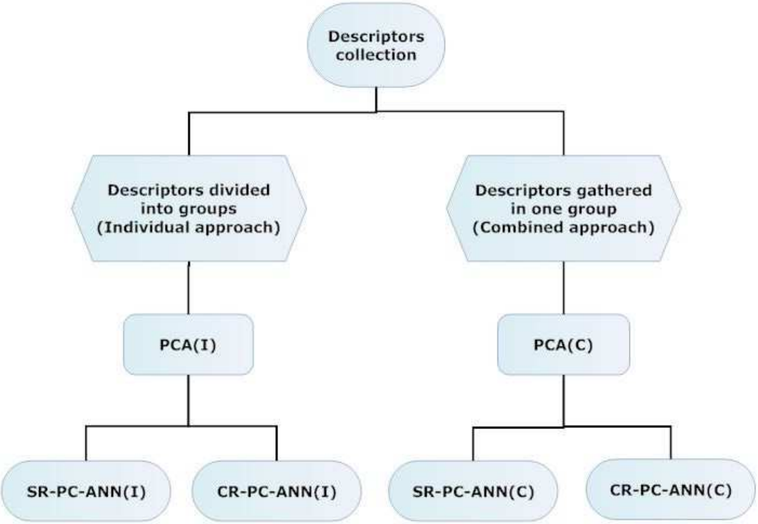

Descriptors collection
Descriptors divided into groups (Individual approach)
Descriptors gathered in one group (Combined approach)
PCA(I)
PCA(C)
SR-PC-ANN(I)
CR-PC-ANN(I)
SR-PC-ANN(C)
CR-PC-ANN(C)

Fig. 4. PC-ANN approaches used in this study.

The results of modeling the toxicity data indicates that the residual plots for the training, validation and test sets are not scattered and they do not warranty the stability of the models. There is a strong relationship between the residual and actual values which reflects that the obtained models have systematic error, therefore a correction scheme is done to correct this issue. The cross-validation parameters for the chosen models before and after correction are shown in Table 2. This table shows that the correction term improves the cross-validation parameters by lowering the RMSE and increasing the $R^2CV$ values. Considering the number of variables entered to the regression model for the SR-PC-ANN approach, PCA(I) based-model which has lower number of variables (3 variables) is superior over the PCR(C) based-model which has 13 variables due to chance correlation possibilities. It also shows that, after applying the correction scheme, model 3 obtained from the SR-PC-ANN(I) explains 58.5% of the data variances and has a 0.471 RMSE of prediction. This model has regression coefficients of 0.811 and 0.858 for the training and test sets, respectively. The RMSE of prediction for the CR-PC-ANN(I) model is less than that of the CR-PC-ANN(C) optimal model (0.571 and 0.590, respectively). The optimal model obtained using the CR-PC-ANN (I) procedure has correlation coefficients of 0.817 and 0.866 for the training and test sets, respectively and explains 53.4% of the data variances while the optimal model obtained from the CR-PC-ANN(C) procedure explains 51.1% of the data variances.

| **Approach used** | **PCs entered in the model** | **RMSE$^c$** | **R$^2{}_{CV}{}^c$** | **RMSE$^p$** |
|---|---|---|---|---|
| SR-PC-ANN(I) | PC2-1 + PC17-3 + PC13-2 | 0.657 | 0.496 | 0.668 |
| | | 0.498 | 0.585 | 0.471 |
| SR-PC-ANN(C) | PC2 + PC5 + PC6 + PC30 + PC29 + PC8 + PC3 + PC48 + PC1 + PC24 + PC11 + PC10 + PC18 | 0.618 | 0.558 | 0.619 |
| | | 0.404 | 0.734 | 0.451 |
| CR-PC-ANN(I) | PC2-1, PC17-3, PC16-2 | 0.691 | 0.443 | 0.691 |
| | | 0.525 | 0.534 | 0.571 |
| CR-PC-ANN(C) | PC2, PC1, PC3, PC12, PC4, PC5, PC45, PC34, PC24, PC26, PC48, PC43, PC27 | 0.682 | 0.452 | 0.643 |
| | | 0.551 | 0.511 | 0.590 |

Table 2. Cross validation parameters for the original models (grey background) and the corrected ones (white background) for the optimal ANN models of the different approaches used for modeling the toxicity data.

A randomization test was performed and the results obtained for five trails shows that the probability of obtaining chance models (with high correlation coefficients) from the PCA(C) approach is more than that for the PCA(I) approach. Furthermore, it is noticed that the chance correlation coefficients obtained for the CR-PCANN( I) approach are lower than those obtained for the SR-PC-ANN(I) approach. This indicates that model obtained from the CR-PC-ANN(I) approach is more accurate than the other models.

Following the same procedure used for modeling the toxicity of substituted benzenes, the SR-PC-ANN and CR-PC-ANN approaches were compared for modeling the HSA binding affinity with the PCs extracted according to PCA(I) and PCA(C) approaches. The results of this analysis are summarized in table 3.

| Approach used | PCs entered in the model | RMSE$^{c}$ | R$^{2}_{CV}{}^{c}$ | RMSE$^{p}$ |
|---|---|---|---|---|
| SR-PC-ANN(I) | PC16-1 + PC15-1 + PC6-1 + PC5-1 + PC4-1 + PC14-1 + PC1-1 + PC8-1 + PC2-2 + PC11-1 | 0.361 | 0.441 | 0.245 |
| | | 0.258 | 0.509 | 0.077 |
| SR-PC-ANN(C) | PC4 + PC8 + PC1 + PC33 + PC9 + PC3 + PC22 + PC12 + PC17 + PC31 + PC28 + PC7 + PC34 + PC10 + PC46 + PC20 + PC26 | 0.282 | 0.642 | 0.419 |
| | | 0.270 | 0.747 | 0.388 |
| CR-PC-ANN(I) | PC17-1, PC2-3, PC2-2, PC7-1, PC4-1, PC15-1, PC8-1, PC5-1, PC16-1, PC6-1, PC13-1, PC11-1, PC1-1, PC14-1 | 0.339 | 0.414 | 0.545 |
| | | 0.419 | 0.583 | 0.586 |
| CR-PC-ANN(C) | PC4 , PC8 , PC1 , PC33 , PC9 , PC3 , PC22 , PC12 , PC17 , PC31 , PC28 , PC7 , PC34 , PC10 , PC46 , PC20 , PC26 , PC27 , PC6 , PC25 | 0.284 | 0.633 | 0.454 |
| | | 0.258 | 0.764 | 0.321 |

Table 3. Cross validation parameters for the original models (grey background) and the corrected ones (white background) for the optimal ANN models of the different approaches used for modeling the HSA binding affinity.

It shows that the correction terms improve the cross-validation parameters by increasing the R$^{2}$CV and decreasing the RMSE values. This table shows also that the RMSE of prediction of the model obtained from the SR-PC ANN(I) approach is smaller than those for of the models obtained from the other approaches. The corrected model explains 50.9% of the data variances and has a RMSE of prediction of 0.077, regression coefficients of 0.714 and 0.670 for the training and test sets, respectively. On the other hand, the corrected model obtained from the CR-PC-ANN(I) approach explains 53.7% of the data variances and has a RMSE of prediction of 0.586, regression coefficients of 0.733 and 0.675 for the training and test sets, respectively.

In summary, the performance of the two novel QSAR algorithms, principal component-artificial neural network modeling method, named SR-PC- ANN and CR-PC-ANN, combined with two factor selection procedures, named PCA(I) and PCA(C), is compared. These methods are applied to predict the toxic activity of a large set of compounds (278 substituted benzenes) as well as HSA binding affinity (94 compounds). The optimal model for the toxicity data set has a prediction RMSE of 0.471 while the optimal model for the HSA binding affinity has a prediction RMSE of 0.077. Comparison of the models shows that the results obtained by the CR-PC-ANN procedure are more accurate than those obtained from the SR-PC-ANN procedure. Generally, the models obtained from the PCA(I) approach are better than those obtained from the PCA(C) approach regardless which approach was used to perform the ANN analysis. Both the external and cross-validation methods are used to validate the performances of the resulting models.

Randomization test is employed to check the suitability of the models and to investigate the possibility of obtaining chance models.

### 3.2.2 Exploring QSARs of some analgesic compounds

Analgesics are a class of drugs used to reduce pain. The pain relief induced by analgesics occurs either by blocking pain signals going to the brain or by interfering with the brain's interpretation of the signals, without loss of awareness. There are essentially two kinds of analgesics: non-narcotics and narcotics. Because of the potential to relieve pain, they play an important role in medical therapy. The dose required to produce analgesia frequently does not change the functions of central nervous system. Analgesia is believed to engage activation of μ-receptors largely at supraspinal sites and k-receptors mainly within the spinal cord. It has been demonstrated that log IC, where IC refers to the half maximal (50%) inhibitory concentration of a drug, can be successfully used to predict analgesic activity. The aim of this study is to apply PC-ANN with different molecular descriptors in the development of new statistically validated QSAR models. This model will predict the analgesic activity of the heterogeneous data set of different types of analgesics (narcotic, opioid, and non-opioid) as a whole without splitting them into categorizes. The strength and the predictive performance of the proposed models were verified using both internal (cross-validation and randomization) and external statistical validations.

In this study (Deeb and Drabh, 2010), a data set of 95 analgesic compounds and their analgesic activity (log IC) obtained from reference (Mathur, 2003) were used in this study. HyperChem software was used to optimize the structure of the different compounds on AM1 semi-empirical level. The optimization was preceded by the Polak-Rebiere algorithm to reach 0.01 root mean square gradient. In this study, a pool of molecular descriptors including constitutional, topological, chemical, quantum, and functional descriptors were calculated using Hyperchem and Dragon software. A condensed set of 150 descriptors were obtained by removal of highly intercorrelated ($r > 0.95$) descriptors in addition to descriptors having constant values. Descriptors that have zero for almost all the cases were also removed together with those descriptors that include outliers' values to enclose a set of 132 descriptors. This set was then declined to 24 descriptors by stepwise regression.

In the MLR analysis, different regression models were suggested in which the number of descriptors in these models varied between 1 and 20. The best correlation coefficient obtained is 0.760 for a regression model with 20 descriptors. The linear relationships according to MLR analysis provide models with poor cross-validation parameters. Therefore, ANNs algorithm was used to investigate non-linear relationships for the best MLR models according to the cross-validation coefficient of determination ($R^2CV$).

In PC-ANN, the inputs of the ANN were the subset of the descriptors used in different MLR models. From the correlation data matrix for these descriptors, some of them represent considerable degree of collinearity. Therefore, the PCA was performed first to classify the molecules into training, validation, and prediction (test) sets. Performing PCA on the whole data of 95 compounds, 132 descriptors and plotting the first and second principals (Figure 5), shows that 11 compounds behave differently (outliers) from other compounds with respect to both molecular structure (descriptors) and analgesic activity. Therefore, these compounds are not used in the future analysis (Figure 6).

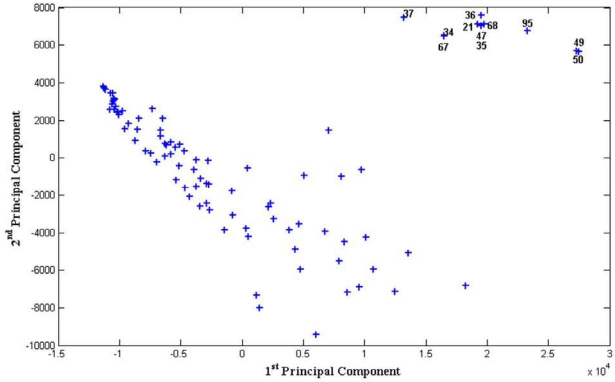


Fig. 5. First and second principal components for the factor spaces of the descriptors and analgesic activity data.

Checking the structure of the outlier compounds shown in Figure 6 reveals that they all morphine derivatives and belong to the same family of opiates analgesic.

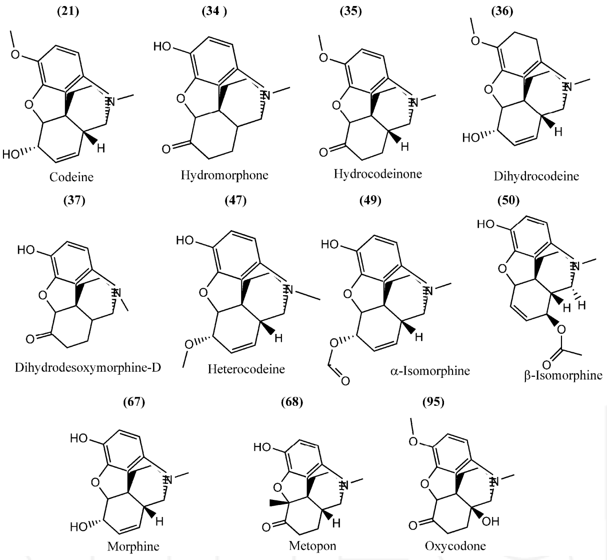


Fig. 6. Structure of outliers suggested from principal component analysis.

According to the pattern of the distribution of the data in factor spaces (Figure 5), the training, validation, and prediction compounds were selected homogenously, so that compounds in different zones of Figure 5 belong to all three subsets. After removing the outliers, the classified data were used as an input for the ANN. In this study, a three-layered feed-forward ANN model with back-propagation learning algorithm was employed. At first, non-linear relationship between the subset of descriptors selected by stepwise selection-based MLR and analgesic activity was preceded by PC-ANN models with similar structure. The number of hidden layer's nodes was set to 8 for all models, and the number of nodes in the input layer was the number of PCs extracted for each subset of descriptors. After that, for the best models, optimization of the number of hidden nodes was done.

Cross-validation parameters show that the prediction ability is improved for model the best model which has a relative standard error of prediction of 5.396% and the correlation coefficient of 0.834 ands 0.846 for the training and test sets, respectively. The cross-validation coefficient of determination is 0.656 which means that the six PCs selected by eigenvalue ranking procedure can explain at least 63.6% variances in log IC for the calibration. Consequently, the optimal performance occurs for the best model when using nine hidden nodes. A randomization test was performed to investigate the probability of chance correlation for the optimal model. The results of randomization test indicate that the correlation coefficients obtained by chance are low in general, while the predicted error values are high. This indicates that the model obtained from PCA-ANN is better than those obtained by chance.

In summary, the results obtained by principal component-artificial neural network give advanced regression models with good prediction ability using a relatively low number of principal components. A 0.834 correlation coefficient was obtained using principal component- artificial neural network with six extracted principal components.

### 3.2.3 QSAR Model of drug-binding to human serum albumin

In this study (Deeb and Hemmateenejad, 2007), a data set of 94 HSA drug and drug-like compounds and their binding affinities reported by Colmenarejo (Colmenarejo, 2003) are used in this study. HYPERCHEM software was used to optimize the structure of the different compounds on AM1 semi-empirical level. The optimization was preceded by the Polak-Rebiere algorithm to reach 0.01 root mean square gradient (298 K, gas phase). Esbelen (compound number 8) was dropped from this set because Se is not parameterized for AM1 semiempirical method. In this study, a set of 60 molecular descriptors including constitutional, topological, chemical, and quantum descriptors were calculated using Hyperchem and Dragon software.

Multiple linear regression analysis with stepwise selection and elimination of variables was employed to model the binding affinity (log K'hsa) relationships with different set of descriptors. The number of descriptors in the suggested MLR models is varied between 1 and 25. The best correlation coefficient obtained is 0.912 for a regression model with 25 descriptors. The number of descriptors is large according to the rule of the thumb, whereas the statistical parameters are not so high. Therefore, ANNs algorithm was used seeking for better regression model.

In PC-ANN, the inputs of the ANN were the subset of the descriptors used in different MLR models. The correlation data matrix for these descriptors indicated that some descriptors represent high degree of collinearity. Principal component analysis groups together descriptors that are collinear to form a composite indicator capable of capturing as much of common information of those descriptors as possible. Application of PCA on a descriptor data matrix results in a loading matrix containing factors or PCs, which are orthogonal. These factors used as the inputs of ANN instead of the original descriptors.

The procedure used in this study is similar to that used in the QSAR study of some analgesics compounds. Performing PCA on the whole data of 94 compounds and 60 descriptors and plotting the first and second principal, shows that compounds 62 and 91 are outliers. According to the pattern of the distribution of the data in factor spaces the training, validation, and prediction molecules were selected homogenously. After removing the outliers the classified data was used as an input for the ANN. A three-layered feed-forward ANN model with backpropagation learning algorithm was employed. At the first, the nonlinear relationship between the subset of descriptors selected by stepwise selection-based MLR and drug HSA binding constant was proceeded by PC-ANN models with similar structure. The number of hidden layer's nodes was set 3 for all models and the number of nodes in the input layer was the number of PCs extracted for each subset of descriptors. The results of PC-ANN modeling for MLR model numbers 15–25 shows that the best model which has almost the highest correlation coefficient for the external test set (0.8065) which indicates a high predictive power. This model has also a relatively low PRESS/SST ratio (0.4485) compared with other models which make it the most reasonable model among all. The $R^2$ values for the cross-validation and prediction for this model are 0.5515 and 0.5100, respectively, which means that the six PCs selected by eigenvalue ranking procedure can explain at least 55.2% and 51% variances in log K'hsa for the calibration and prediction, respectively. In order to optimize the performance of the ANN model , we trained the ANN using different number of hidden nodes starting from 1 hidden node to 20 hidden nodes. The results for the optimization indicate that an ANN with eight hidden nodes resulted in the optimum network model. Using eight hidden nodes, we obtained almost the highest correlation coefficient for both the training set (0.9218) and the prediction set (0.8302). This model gives the lowest PRESS/SST ratio (0.2757) which makes it the most reasonable. The results of the randomization test shows that the correlation coefficients obtained by chance are low in general while PRESS and PRESS/SST ratio are high. This indicates that the model obtained from PCA-ANN is better than those obtained by chance.

### 3.3 Exploring QSARs for inhibitory activity of non-peptide HIV-1 protease inhibitors by GA-PLS and GA-SVM

Human immunodeficiency virus (HIV), the causative agent of AIDS infects millions of people worldwide. Although a treatment has not been found yet for this serious disease, rapid advances in molecular biology along with the 3-D elucidation of HIV proteins have led to new drug-targeting approaches for designing antiviral agents that specifically bind to key regulatory proteins that are essential for HIV replication. Thus, by developing new inhibitors of HIV-1 protease activity, the treatment of AIDS can be advanced. Several peptidic inhibitors are currently under clinical trials and significant efforts to improve their pharmacology continues. In this study, we picked out small non-peptide HIV protease inhibitors with potentially better pharmacological characteristics based on the structural features of peptidic inhibitors bound to the enzyme, and performed QSAR study.

In this study (Deeb and Goodarzi, 2010), a data set of 46 non-peptide HIV-1 protease inhibitors and their inhibitory activity reported by Tummino et al. (Tummino et al. 1996) are used in this study. Molecular chemical structure was built using Hyperchem. AM1 method was applied to optimize the molecular structure of the compounds. All calculations were carried out at the restricted Hartree-Fock level with no configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient was 0.01 Kcal/mol.

One thousand four hundred eighty one descriptors belonging to eighteen different theoretical descriptors were calculated for each molecule. The calculated descriptors were first analyzed for the existence of constant or near constant variables. The detected ones were then removed. Correlation among descriptors with the activity of the molecules was examined and collinear descriptors (i.e. r > 0.9) were detected. Descriptors that contain a high percentage (> 90%) of identical values were discarded to decrease the redundancy existing in the descriptor data matrix. Among the collinear descriptors, the one presenting the highest correlation with the activity was retained and others were removed from the data matrix. The dataset was splitted into two sets based on activity range; training set (85%) with activity ranges from 3.921 to 8.444 and test set (15%) with activity ranges from 4.538 to 8.208. In this work, genetic algorithm (GA) variable subset selection method (Leardi et al., 1992) was used for the selection of the most relevant descriptors from the pool of remaining descriptors. These descriptors would be used as inputs for PLS and SVM in the construction of QSAR models.

In GA-PLS, model validation was achieved through leave-one-out cross-validation (LOO CV) to find the best number of latent variables (Lv) to be used in calibration and prediction. External validation (for a test set), and the predictive ability was statistically evaluated through the root mean square errors of calibration (RMSEC) and validation (RMSECV). The results indicate that four latent variables are the best number to make a model. The following equation represents the best model achieved by GA-PLS:

$$\begin{aligned} pIC_{50} = & -3.405737 (\pm 1.447) + 0.525607 (\pm 0.052)\ \text{TE2} \\ & + 0.911090 (\pm 0.236)\text{Ui} \\ & + 2.586873 (\pm 0.369)\ \text{GATS5e} \\ & - 47.069316 (\pm 8.558)\text{Mor13e} \\ & - 0.207581 (\pm 0.027)\ \text{ATS7m} \\ & + 13.338116 (\pm 3.599\ )\ \text{Ss} \\ & -0.001142 (\pm 0.000)\ \text{Mor27e} \\ & + 49.494231 (\pm 7.841)\ \text{RDF035e} \end{aligned} \tag{7}$$

The best model shown above reveals that the most significant contribution comes from the RDF035e. Table 4 gives brief description of these descriptors.

In GA-SVM, the quality of SVM for regression depends on several parameters namely, kernel type k, which determines the sample distribution in the mapping space, and its corresponding parameter σ, capacity parameter C, and ε-insensitive loss function. The three parameters were optimized in a systematic grid search-way and the final optimal model was determined. Six general statistical parameters were selected to evaluate the prediction ability of the constructed model. These parameters are: root mean square error of prediction

(RMSEP), relative standard error of prediction (RSEP), mean absolute error (MAE), square of correlation coefficient (R2), F-statistical and t test. Table 5 shows the results of GA-PLS and GA-SVM and the calculated statistical parameters. This table shows that the results of the GA-SVM are better than GA-PLS.

| No | Symbol | Class | Meaning |
|---|---|---|---|
| 1 | TE2 | Charge descriptors | Topographic electronic descriptor(bond resctricted) |
| 2 | Ui | Empirical descriptors | Unsaturation index |
| 3 | GATS5e | 2D autocorrelations | Geary autocorrelation-lag5/weighted by atomic Sanderson electronegativities |
| 4 | Mor13m | 3D-MoRSE descriptors | 3D MoRSE-signal13/weighted by atomic masses |
| 5 | ATS7m | 2D autocorrelations | Broto-Moreau autocorrelation of a topological structure- lag7/ weighted by atomic masses |
| 6 | Ss | Constitutional descriptors | Sum of Kier-Hall electrotopological States |
| 7 | Mor27e | 3D-MoRSE descriptors | 3D MoRSE-signal27/weighted by atomic Sanderson electronegativities |
| 8 | RDF035e | RDF descriptors | Radial Distribution function-3.5/weighted by atomic Sanderson electronegativities |

Table 4. Description of the selected descriptors in this study.

| Parameters | | GA-SVM | GA-PLS |
|---|---|---|---|
| NOC[a] | | | 4 |
| $Q^2$ LOO[b] | | 0.9672 | 0.8259 |
| $\sigma$ | | 0.5 | |
| $\varepsilon$ | | 0.06 | |
| C | | 8 | |
| RMSEP | Training set | 0.2027 | 0.3934 |
| | Test set | 0.2751 | 0.3962 |
| RSEP(%) | Training set | 3.1520 | 6.1156 |
| | Test set | 4.0216 | 5.7928 |
| MAE(%) | Training set | 6.5080 | 8.9351 |
| | Test set | 18.093 | 21.745 |
| $R^2$ | Training set | 0.9800 | 0.8935 |
| | Test set | 0.9355 | 0.8603 |
| F statistical | Training set | 1815.2 | 310.26 |
| | Test set | 72.481 | 30.792 |
| T test | Training set | 42.606 | 17.614 |
| | Test set | 8.5136 | 5.5491 |

[a] Number of components

[b] $Q^2$ Leave-one-out Cross-validation

Table 5. Results and statistical parameters of GA-PLS and GA-SVM.

Figure 7A shows calculated $pIC_{50}$ against experimental values, while Figure 7B shows their residual values against the experimental $pIC_{50}$ using GA-SVM.

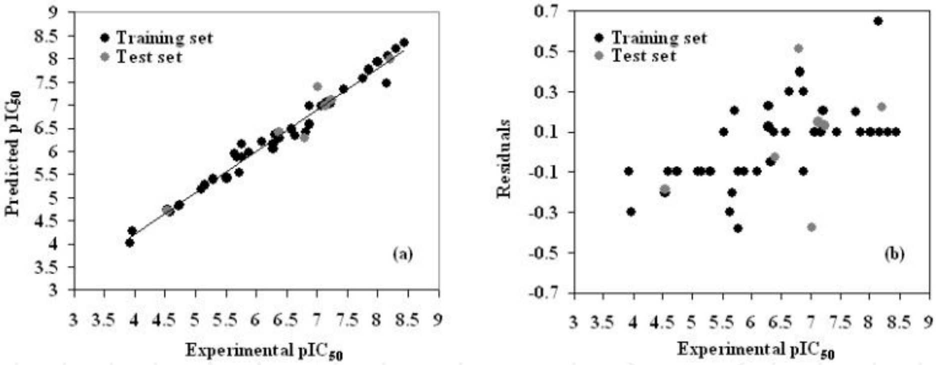


Fig. 7. A) Calculated $pIC_{50}$ against the experimental values using GA-SVM. B) Residual values against experimental $pIC_{50}$ using GA-SVM.

In summary, the support vector machine (SVM) and partial least square (PLS) methods were used to develop quantitative structure activity relationship (QSAR) models to predict the inhibitory activity of nonpeptide HIV-1 protease inhibitors. Genetic algorithm (GA) was employed to select variables that lead to the best-fitted models. A comparison between the obtained results using SVM with those of PLS revealed that the SVM model is much better than that of PLS. The root mean square errors of the training set and the test set for SVM model were calculated to be 0.2027, 0.2751, and the coefficients of determination ($R^2$) are 0.9800, 0.9355 respectively. Furthermore, the obtained statistical parameter of leave-one-out cross-validation test ($Q^2$) on SVM model was 0.9672, which proves the reliability of this model. Omar Deeb is thankful for Al-Quds University for financial support.

## 4. Acknowledgment

Omar Deeb is thankful for Al-Quds university for financial support.

## 5. References

Allen, D.M. (1974) The relationship between variable selection and data augmentation and a method for prediction. *Technometrics, 16*, 125-127.

Clare, B.W. (1998). The Frontier Orbital Phase Angles: Novel QSAR Descriptors for Benzene Derivatives, Applied to Phenylalkylamine Hallucinogens. *J. Med. Chem.* 41, 3845 – 3856.

Clare, B.W. (2000). The frontier orbital phase angles: a theoretical interpretation. *Theochem*, 507, 157-164.

Clare, B.W. (2001). Erratum to "The frontier orbital phase angles: a theoretical interpretation". *J. Mol. Struct. (Theochem)* 507 (2000) 157-167]. *Theochem*, 535 , 301-301.

Clare, B.W. (2002). QSAR of benzene derivatives: comparison of classical descriptors, quantum theoretic parameters and flip regression, exemplified by phenylalkylamine hallucinogens. *J. Comput.-Aided Mol. Des.* 16, 611-633.

Clare, B.W. (2004). A novel quantum theoretic QSAR for hallucinogenic tryptamines: a major factor is the orientation of π orbital nodes. *Theochem*, 712, 143-148.

Clare, B.W. and Supuran, C.T. (2004). Quantum Theoretic QSAR of Benzene Derivatives: Some Enzyme Inhibitors. *J. Enz. Inhib. Med. Chem.* 19, 237-248.

Clare, B.W. and Supuran, C.T. (2005a). Predictive Flip Regression: A Technique for QSAR of Derivatives of Symmetric Molecules, *J. Chem. Inf. Model.* 45, 1385-1391.

Clare, B.W. and Supuran, C.T (2005b). A physically interpretable quantum-theoretic QSAR for some carbonic anhydrase inhibitors with diverse aromatic rings, obtained by a new QSAR procedure. *Bioorg. Med. Chem.*, 13, 2197–2211.

Colmenarejo G. (2003). In silico prediction of drug-binding strengths to human serum albumin. *Med Res Rev*, 23, 275–301.

Cronin M.T.D., Gregory B.W. and Schultz T.W. (1998) Quantitative structure-activity analyses of nitrobenzene toxicity to Tetrahymena pyriformis. *Chem. Res. Toxicol.*, 11, 902–908.

Deeb O. and Hemmateenejad B., (2007). ANN-QSAR Model of Drug-binding to Human Serum Albumin, *Chem. Biol. Drug Des.* 70, 19-29.

Deeb O., (2010). Correlation ranking and stepwise regression procedures in principal components artificial neural networks modeling with application to predict toxic activity and human serum albumin binding affinity, *Chemometrics and Intelligent Laboratory Systems* 104, 181–194.

Deeb O. and Drabh M.,(2010). Exploring QSARs of some analgesic compounds by PC-ANN, *Chem Biol, Drug Des.* 76, 255-262.

Deeb O. and Goodarzi M., (2010), Exploring QSARs for Inhibitory Activity of Non-peptide HIV-1 Protease Inhibitors by GA-PLS and GA-SVM, *Chem. Biol. Drug Des.* 75, 506-514.

Deeb, O. and Clare, B.W., (2007). QSAR of Aromatic Substances: Protein Tyrosine Kinase Inhibitory Activity of Flavonoid Analogues. *Chem. Biol. Drug. Des.*, 70, 437–449.

Deeb, O. and Clare, B.W. (2008a) QSAR of aromatic substances: EGFR inhibitory activity of quinazoline Analogues. *J Enz. Inhib. Med. Chem.* 23, 763–775.

Deeb, O. and Clare, B.W.; (2008b) Comparison of AM1 and B3LYP-DFT for Inhibition of MAO-A by Phenylisopropylamines: A QSAR Study. *Chem. Biol. Drug. Des.*, 71, 352–362.

Devillers, J. and Balaban, A.T. (1999) *Topological Indices and Related Descriptors in QSAR and QSPR.* Gordon Breach Scientific Publishers: Amsterdam, 811.

Downs, G.M. (2004). Molecular Descriptors. In *Computational Medicinal Chemistry for Drug Discovery.* Bultinck, P.; De Winter, H.; Langenaeker, W.; Tollenaere, J. P. (Eds.). Marcel Dekker; New York, 515-538.

Duprat, A.F., Huynh, T. and Dreyfus, G.(1998). Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of log P. *J. Chem. Inf. Comput. Sci.*, *38*, 586-594.

Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold S.(2001). *Multi- and Megavariate Data Analysis. Principles and Applications.* Umetrics: Umea.

Feng J., Lurati L., Ouyang H, Robinson T., Wang Y, Yuan Y. and Young S. (2003) Predictive toxicology: benchmarking molecular descriptors and statistical methods, *J. Chem. Inf. Comput. Sci.* 43, 1463–1470.

Free S. M. and Wilson J. W. , (1964). A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.*, 7, 395-399.

Fujita, T., Iwasa, J. and Hansch, C. (1964). A new substituent constant,π, derived from partition coefficients. *J. Am. Chem. Soc.*, *86*, 5175-5180.

Fujita, T. (1990). The extrathermodynamic approach to drug design. In "*Comprehensive Medicinal Chemistry*" (C. Hansch, P. G. Sammes, and J. B. Taylor, eds.), Vol. 4, Pergamon, Elmsford, NY. pp. 497-560.

Gunn S.R. (1997) Support Vector Machines for Classification and Regression. UK: University of Southampton.

Hansch C., Maloney P. P., T. Fujita and Muir R. M. , (1962), Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature*, 194, 178.

Hansch, C. and Fujita, T. (1964). ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.*, *86*, 1616-1626.

Hansch, C. (1969). A quantitative approach to biochemical structure-activity relationships. *Acct. Chem.Res., 2*, 232-239.

Jolliffe, I.T. (1986) *Principal Component Analysis*. Springer-Verlag: New York

Karelson, M. *(2000). Molecular Descriptors in QSAR/QSPR*. Wiley-InterScience; New York.

Katritzky, A.R., Lobanov, V.S. and Karelson, M.(1994) CODESSA, Reference Manual. Gainesville, FL University of Florida. Available:
http://www.semichem.com/codessarefs.html

Kishida, K. and Manabe, R. (1980). The role of the hydrophobicity of the substituted groups of dichlorphenamide in the development of carbonic anhydrase inhibition. *Med. J. Osaka Univ*. 30, 95-100.

Leardi R., Boggia R. and Terrile M. (1992) Genetic algorithms as a strategy for feature selection. *J Chemom*. 6, 267–281.

Mathur K.C., Gupta S. and Khadikar P.V. (2003) Topological modeling of analgesia. *Bioorg Med Chem*, 11, 1915–1928.

Montgomery, D.C. and Peck, E.A.(1992) *Introduction to linear regression analysis*. Wiley: New York

Novi , M.; Nikolovska-Coleska, Z. and Solmajer, T.(1997) Quantitative structure-activity relationship of flavonoid p56 protein tyrosine kinase inhibitors. A neural network approach. *J. Chem. Inf. Comput. Sci.*, *37*, 990-998.

Shulgin, A., Shulgin A. and Pihkal A. (1991). A Chemical Love Story; Transform Press: Box 13675, Berkeley, CA, pp 864-869.

Silipo, C. and Vittoria, A. (1990) Three-Dimensional Structure of Drugs. In *Comprehensive Medicinal Chemistry. Vol 4. Quantitive Drug Design*. Hansch, C. Sammes, P.G.; Taylor, J.B.; eds. Pergamon Press, New York, 154-204.

Todeschini, R.; Consonni, V. and Pavan, M. (2001). DRAGON-Software for the Calculation of Molecular Descriptors. Release 1.12 for Windows. Available:
http://www.disat.unimib/chm

Tetko, I.V.; Alessandro, E.; Villa, P. and Livingstone, D.J. (1996) Neural network studies. 2. Variable selection. *J. Chem. Inf. Comput. Sci., 36*, 794-803.

Tummino P.J., Prasad J.V.N.V., Ferguson D., Nouhan C., Graham N., Domagala J.M., Ellsworth E. et al. (1996) Discovery and optimization of nonpeptide HIV-1 protease inhibitors. *Bioorg Med Chem.,* 4, 1401–1410.

Wang X., Yin C. and Wang L. (2002) Structure–activity relationships and response–surface analysis of nitroaromatics toxicity to the yeast (Saccharomyces cerevisiae). *Chemosphere*, 46, 1045–1051.

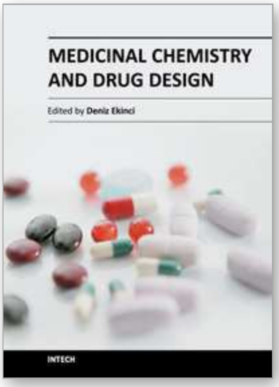
Wold, S. (1978) Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics, 20,* 397-405.

Wold, S. (1991) Validation of QSARs. *Quant. Struct.-Act. Relat., 10,* 191-193.

Wold, S., Johansson, E. and Cocchi, M.(1993) PLS—partial least squares projections to latent structures. In *3D-QSAR in Drug Design, Theory, Methods, and Applications.* Kubinyi, H. (Ed). ESCOM Science Publishers: Leiden, 523–550

Xu, L. and Zhang, W.J.(2001) Comparison of different methods for variable selection. *Anal. Chim. Acta, 446*, 477–483.

Yang E.B., Guo Y.J., Zhang K., Chen Y.Z. and Mack P. (2001) Inhibition of epidermal growth factor receptor tyrosine kinase by chalcone derivatives. *Biochim. Biophys. Acta.,* 1550, 144-152.



**Medicinal Chemistry and Drug Design**
Edited by Prof. Deniz Ekinci

ISBN 978-953-51-0513-8
Hard cover, 406 pages
**Publisher** InTech
**Published online** 16, May, 2012
**Published in print edition** May, 2012

Over the recent years, medicinal chemistry has become responsible for explaining interactions of chemical molecules processes such that many scientists in the life sciences from agronomy to medicine are engaged in medicinal research. This book contains an overview focusing on the research area of enzyme inhibitors, molecular aspects of drug metabolism, organic synthesis, prodrug synthesis, in silico studies and chemical compounds used in relevant approaches. The book deals with basic issues and some of the recent developments in medicinal chemistry and drug design. Particular emphasis is devoted to both theoretical and experimental aspect of modern drug design. The primary target audience for the book includes students, researchers, biologists, chemists, chemical engineers and professionals who are interested in associated areas. The textbook is written by international scientists with expertise in chemistry, protein biochemistry, enzymology, molecular biology and genetics many of which are active in biochemical and biomedical research. We hope that the textbook will enhance the knowledge of scientists in the complexities of some medicinal approaches; it will stimulate both professionals and students to dedicate part of their future research in understanding relevant mechanisms and applications of medicinal chemistry and drug design.

**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821