QSAR & Combinatorial Science
28,11-12, 1487–1499,2009

Is Feature Selection Essential for ANN Modeling?

Mohammad Goodarzi[a], Shreekant Deshpande[b], Vanangamudi Murugesan[b], Seturam B. Katti[b] and Yenamandra S. Prabhakar[b*]

[a]Department of Chemistry, Faculty of Sciences, Azad University, Arak, Iran; Young Researchers Club, Azad University, Arak, Iran

[b]Medicinal and Process chemistry Division, Central Drug Research Institute, Lucknow -226 001, India.

Key words. Artificial neural networks, Feature selection, DRAGON descriptors, Thiazolidin-4-ones, HIV-1 RT, Anilinoquinolines, Piperazinoquinolines, Antimalarials.

Received on…..

Type of manuscript: Full paper

ABSTRACT. In modeling approaches, artificial neural networks (ANNs) have a special place to address the nonlinear phenomena or curved manifold. Often one or other feature selection approach is used prior to ANN to feed the input variables for its models. The function of 'selected' versus 'arbitrary' features on the outcome of ANN models is investigated with a variety of objectively selected and arbitrarily chosen variables from chemical databases namely thiazolidinones, anilinoquinolines and piperazinoquinolines. For each database, its biological activity is considered as the dependent variable and the molecular descriptors from DRAGON software are used as explanatory variables. The selection sets are obtained from feature selection approaches namely, combinatorial protocol in multiple linear regression, stepwise regression and genetic algorithm. Apart from these, a large number of arbitrary sets have been created by randomly picking the descriptors from corresponding databases. The features of all sets have shown a variety of inter- and intra- set diversities. A three-layer back propagation ANN with Levenberg-Marquardt optimization algorithm has been used for modeling the phenomena. Regardless of the origin of the feature sets, the ANN models from a very large number of sets have well explained the activity and qualified themselves to be predictive models. Also, no specific pattern is apparent between the quality of ANN model and the origin of its input feature set. Since these results are unusual, the study is extended to a few more databases. All the results emphasized the innate ability of ANN in developing complex network of relations among features to estimate the target variable. This has prompted us to suggest that prior feature selection is not essential for ANN and it is a desirable option for meaningful outputs in terms of the rationale behind the inputs.

*Corresponding author, Phone:+91-522-2612411; Fax:+91-522-2623405; E-mail: yenpra@yahoo.com (Y.S. Prabhakar)

1 Introduction

In nature all systems are recognized, identified and characterized by their measurements, may be each one under different metrics. These measurements with respect to time and space assign characteristics to the behavior of the system. In this scenario, a character or behavior gets infused into the system when it is not isolated as well as not stationary with respect to time and/ or space. Whatever may be the nature of a system, the human endeavor has always been focused to tame, tailor and comprehend its behavior. All modeling and simulation studies provide a well defined course to understand the behavior of the systems. For these studies, parameterization of the system under investigation is very crucial. Most of the phenomena of interest are a result of different forces operating in a given environment. Modeling and simulations are common area of interest to many a cost and time intensive investigations. They deal with the processing of measurements to durable predictions. In this, correlation is a consequence of predictability and it declares the soundness of model. Drug research is one such explorative research area. This has prompted the development of different metrics and led to an exponential growth in the number of measurements that could be made on the objects (chemicals) and their environment [1-5]. However, apart from practical and economic considerations, one purpose of models is to forecast the events or phenomena of interest with far less number of observations (from the experimental domain) compared to the actual scope it holds. The increased number of features and their selection approaches has an important bearing on all kinds of modeling approaches as well as on their outcomes [6-13].

In modeling approaches, artificial neural networks (ANNs) have a special place to address the nonlinear phenomena or curved manifold [14-16]. An ANN consists of many pathways and nodes organized into a sequence of layers. The first layer is an input layer with one node for each variable or feature of the data. The last layer is an output layer consisting of one node for each variable to be investigated. In between these, there is a series of one or more hidden layer(s) with computing nodes (hidden), which are responsible for learning to predict the output. They emulate biological neural networks and are developed in an organized step-by-step process either to optimize an ‘end’ measurement or to follow some implicit internal constraint. Different ANN approaches are available to model non-linear phenomena [15]. The back propagation neural networks (BP-ANNs) are often used in a variety of analytical applications [15,16]. In drug research, they are widely applied to quantitative structure-activity relationship (QSAR) studies as a powerful non-linear modeling technique [17]. They can be applied to model unknown situations as well. Moreover, several reports show that ANN models give better statistical results, both in fitting and prediction, in comparison to the linear modeling approaches [17].

Feature selection approaches are commonly used in modeling studies to identify the most ‘relevant’ descriptors to the phenomena under investigation from a pool of measurements made on the object [6-13]. They give opportunity to focus onto a limited number of variables in modeling the phenomena. In general, modeling a phenomenon in ANN as such does not call for any prior feature selection. Its architecture sets no restrictions on the number of input variables to be used to model the output. The collection of descriptors for an object can be very large or infinite, but the number of critical descriptors to predict a phenomenon (associate with the object) is limited. Keeping this in view, often one or other feature selection approach is used

prior to ANN to feed the input variables for its models [9,10,17]. In this background the question of 'selected' versus 'arbitrarily picked' (picked at random) features on the outcome of ANN models has been investigated using different chemical structure databases namely thiazolidin-4-ones tested as HIV-1 RT inhibitors [18,19,20], and quinolines [21,22] tested as antimalarial agents (Figure 1). The molecular descriptors (explanatory variables) of these chemical databases have originated from DRAGON software [4]. These descriptors have been used to create the objectively selected as well as arbitrarily chosen descriptors subsets (feature sets) for predicting the phenomena associated with them in ANN. The feature sets were formed from two perspectives namely (a) feature sets with fixed number of descriptors and (b) feature sets with due consideration to "parsimony" effect in number of descriptors for accounting equal or comparable explained variance of target phenomenon. The ANN models from the feature sets are discovered using the BP-ANN with Levenberg-Marquardt algorithm [15,16,23]. Also simulations with random numbers (meaningless numbers forming false descriptors) in place of true feature sets in the input layer of ANN are investigated in order to differentiate the arbitrarily picked features (true descriptors) from that of meaningless random numbers. A report of the investigation is presented hereunder.

2 Materials and Methods

2.1 Structure Databases and Biological Activities

Three chemical structure databases namely a) thiazolidin-4-ones (termed as thiazolidinones; 184 compounds) tested as HIV-1 RT inhibitors (EC_{50}, effective concentration to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1 RT) [18,19,20], b) 4-(3′,5′-disubstituted anilino)quinolines (termed as anilinoquinolines; 74 compounds) tested as antimalarial agents (*Plasmodium falciparum, FcB1R*; IC_{50}, concentration required to achieve 50% inhibition) [21] and c) N^1-(7-chloro-4-quinolyl)-1,4-bis(3-aminopropyl)piperazine (termed as piperazinoquinolines; 44 compounds) tested as antimalarial agents (*P. falciparum, FcB1*; IC_{50}, concentration required to achieve 50% inhibition) [22] (Figure 1) were used in this investigation. In SYBYL [24] adopting standard procedure the structures of thiazolidinones and anilinoquinolines were generated using the crystal structures of 7-chloro-1-(2,6-difluorophenyl)-1H,3H-thiazolo[3,4-a] benzimidazole [25] and amodiaquine [26], respectively. In DRAGON software [4] these structure databases have respectively led to 999 (506 for 0D to 2D and 493 for 3D) and 1176 (490 for 0D to 2D and 686 for 3D) descriptors. For piperazinoquinolines, the molecular descriptors (483 descriptors) available in ref.22 were adopted as such. For the purpose of ANN investigation, each structure database was divided into training, validation and test set compounds. For thiazolidinones and anilinoquinolines, the cluster analysis of MACCS fingerprints (FP-BIT-MACCS) [27,28] was used to partition them into training, validation and test sets. In case of piperazinoquinolines, the prior compilation order of compounds as given in ref.22 was used to partition them into training, validation and test sets. From the list of piperazinoquinolines, alternative compounds were assigned to the training test. The remaining compounds of the list were again alternatively assigned to validation and test sets. Table 1 shows the composition of training, validation and test sets of three databases and activity distribution therein.

2.2 Feature Sets

2.2.1 Selection Sets

The feature selection approaches namely combinatorial protocol in multiple linear regression (CP-MLR) [12], stepwise regression (SR) [29] and genetic algorithm (GA) [30,31] were used for the identification of features of the selection sets. Prior to the application of feature selection procedures, all those descriptors showing a correlation of less than 0.1 with the dependent variable (descriptor *vs*. activity $r < 0.1$) were excluded. Only the training set compounds were used in the feature selection approaches for the identification of features of selection sets.

The CP-MLR procedure [12,32] was used for the identification of features of CPMLR set. The thrust of this procedure is in the embedded 'filters'. The details are discussed in some of the recent publications [12,32]. For the selection of features from 0D to 2D and 3D datasets of thiazolidinones in CP-MLR, the initial threshold of filter-1 was assigned as 0.3 and subsequently liberated it to 0.79 to boost the formation different seeds. The individual descriptors of the datasets have shown a correlation of less than 0.6 with the activity. Considering this, the search was started with two-variable seeds and with an initial filter-3 value of 0.6. The information rich descriptors were collected by successively incrementing the number of variables per seed as well as the threshold of filter-3 to the optimum *r*-bar value of the preceding generation. At the end of a search, five eight-parameter models were identified. Each one of these equations has explained more than 67 ($r^2 \geq 0.67$) per cent variance in the training set activity and showed test set r^2 values in the range of 0.440 to 0.519. They have together shared eleven descriptors amongst them. These variables were designated as a selection set from CP-MLR (CPMLR set) for thiazolidinones and were used as an input feature set for the ANN model.

The stepwise regression (SR) procedure [29] was used for the identification of features of SR set. In the forward selection procedure of SR, the 0D to 2D and 3D datasets of thiazolidinones have yielded, respectively, seventeen and nineteen descriptors as significant ones to explain the variance in the activity. The short-listed descriptors were merged and the forward selection procedure of SR was repeated on them to identify the optimum eleven descriptors to explain the variance in the activity. Since eleven descriptors were present in the thiazolidinones CPMLR set, the same numbers of descriptors were opted for its SR set and other sets from this database. In the linear regression, the eleven descriptors from SR (SR set) have explained 78.4 (r^2=0.784) per cent variance in the activity of thiazolidinones' training set compounds.

The genetic algorithm variable subset selection (GA-VSS) routine of MOBY DIGS [30,31] was used for the identification of features of GA set. For thiazolidinones, the selection in GA-VSS proceeded with an initial population of one hundred solutions (chromosomes) with maximum allowed variables in a model as eleven. The fitness for each chromosome was calculated based on leave-one-out (LOO) cross-validation (Q^2). The reproduction/mutation trade-off (T) value was set to 0.5. Based on the T value, the crossover and mutation values of GA were automatically fixed *in situ* in the computation. The optimum solution was identified at the end of one hundred generations of GA evolution process (selection, crossover and mutation). In the linear regression, the eleven descriptors from GA (GA set) have explained 75.7 (r^2=0.757) per cent variance in the activity of training set of thiazolidinones. Also for thiazolidinones, the three

feature selection approaches were used to identify the selection sets with due consideration to the "parsimony" effect in terms of number of features for explaining same or almost same variance in the activity.

For anilinoquinolines and piperazinoquinolines one selection set each was identified. The selection set of anilinoquinolines was identified through CP-MLR. It has five descriptors. In linear regression it has explained 73.5 per cent (r^2 is 0.735) variance in the activity of these compounds. The selection set of piperazinoquinolines is a subset of QSAR equation 4 reported in ref.22. This selection set is composed of three features. In linear regression this has explained 67 per cent (r^2 is 0.67) variance in the activity of the analogues.

2.2.2 Trial Sets

All descriptors of each database were used for the generation of respective arbitrary feature sets. Using in-house written program the features of arbitrary sets of each database were randomly picked from the corresponding molecular descriptors without bias. According to the number of features in the selection sets, a matching number of features were considered for the arbitrary sets (termed as trial sets). For thiazolidinones more than one hundred trial sets were generated. Three trial sets each were generated for anilinoquinolines and piperazinoquinolines. In stepwise regression, the trial sets of each database have explained very negligible variance in their assigned activity. None of the feature sets, selection as well as trial sets, have shown chance correlation with the activity in Y-scrambling test (the scrambled biological responses; Y-randomization study) [32,33].

2.2.3 Feature Sets Diversity

All the feature sets were quantitatively analyzed for their independence and relative separation in the descriptor space. For this the cumulative descriptor-descriptor distances were computed using the following equations.

The internal (within set) distance (*ID*)

$$ID = \frac{100\sum_{i}^{k}\sum_{j}^{k}\left(1 - r_{ij}^{2}\right)}{k^{2} - k} \qquad (1)$$

The external (in between sets) distance (*ED*)

$$ED = \frac{100\sum_{Ai}^{k}\sum_{Bj}^{k}\left(1 - r_{AiBj}^{2}\right)}{k^{2}} \qquad (2)$$

In equation 1, r_{ij} is correlation coefficient between the descriptors *i* and *j* of the same set whereas in equation 2, r_{AiBj} is correlation coefficient between i^{th} descriptor of A-set and j^{th} descriptor of B-set. In both the equations, *k* is the number of descriptors in the selection or trial set. In this, an *ID* (or *ED*) value of 100 denotes orthogonal descriptor set(s) and a zero value refers to fully

collinear descriptor set(s). They ascertain the degree of diversity within and in between feature sets. In addition to this, the descriptors of each set were analyzed for the multicollinearity using Variance Inflation Factor (*VIF*) (equation 3) [34, 35].

$$VIF_i = \frac{1}{(1-r^2)} \qquad (3)$$

For a subset with *k* features, VIF_i is the variance inflation factor of i^{th} feature; r^2 is the multiple regression correlation coefficient of i^{th} feature versus the remaining *k*-1 features of the subset. Literature reports recommend a *VIF* value greater than 5.0 as an indication of multicollinearity in the feature set. There are also reports suggesting higher cutoff values (e.g. 10.0) for *VIF* [35].

Since autoscaling (unit variance and zero mean) of descriptors give them equal weight in the regression, the similarity (or dissimilarity) of the subspaces spanned by the regression equations of selected and randomly picked feature sets were analyzed by comparing the normalized regression coefficients of autoscaled descriptors of each set. For the purpose of comparison of regression subspace of different sets, the normalized regression coefficients from each descriptor set were arranged in descending order of their fraction contribution to the regression in question. The dissimilarity of subspace (*DS*) between the regression equations were computed using the arithmetic differences of respective normalized regression coefficients (in descending order of the fraction contribution to the regression) of feature sets (equation 4).

$$DS = \sqrt{\frac{\sum (C_{Ai} - C_{Bi})^2}{r_A . r_B}} \qquad (4)$$

For descriptor sets with *k*-features each, r_A and r_B are multiple regression correlation coefficients of dependent variable Y with the *k*-features of sets *A* and *B*, and C_{Ai} and C_{Bi} are the i^{th} normalized regression coefficients (in descending order of the fraction contribution to the regression) of corresponding regression equations. A zero value for *DS* indicates identical regression space of sets *A* and *B*. A larger *DS* value indicates greater dissimilarity between the feature sets in question. Furthermore, the extent of relatedness of models emerged from different feature sets were analyzed using the intercorrelations of Y-residuals (ΔY) from each one of them.

2.3 ANN Procedure

A three-layer back propagation ANN (BP-ANN) (Figure 2) with a sigmoid transfer function was used in the investigation of feature sets. The descriptors from the training set were used for the model generation whereas the descriptors from the validation set were used to stop the overtraining of network. And the descriptors from the test set were used to verify the predictivity of the model. Before training the networks, the input and output values were normalized with auto-scaling of all data. The initial weights were selected randomly between (-0.3) and (0.3). For each database, the optimum number of nodes in hidden layer was found through a standard evaluation of its CP-MLR selection set with different numbers of hidden layer nodes [15,16]. For the purpose of comparison of results, the same number of hidden layer nodes was used for the

ANN models from all other feature sets of each database. The goal of training the network is to minimize the output errors by changing the weights between the layers (equation 5).

$$\Delta w_{ij,n} = F_n + \alpha \Delta w_{ij,n-1} \quad (5)$$

In this, Δw_{ij} is the change in the weight factor for each network node, α is the momentum factor, and F is a weight update function, which indicates how weights are changed during the learning process. The weights of hidden layer were optimized using the Levenberg-Marquardt algorithm, a second derivative optimization method [23].

2.3.1 Levenberg-Marquardt Algorithm

In Levenberg-Marquardt algorithm, the update function, F_n, is calculated using equations.

$$F_0 = -g_0 \quad (6)$$

$$g = J^T e \quad (7)$$

$$F_n = -[J^T \times J + \mu I]^{-1} \times J^T \times e \quad (8)$$

where g is gradient and J is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights, and e is a vector of network errors. The parameter μ is multiplied by some factor (λ) whenever a step would result in an increased e and when a step reduces e, μ is divided by λ.

2.3.2 Statistical Parameters

In training the network, the over fitting of data was controlled by comparing the root-mean-square errors (*RMSE*s) of training and validation sets. It measures the goodness of the output and is useful for the comparison of the target values. The training of the network for the prediction of target value was stopped when the *RMSE* of the validation set began to increase while that of training set continues to decrease. The goodness of fit of activity of the test set compounds was used to further validate the developed models. The predictive ability of the constructed models were assessed using different statistical measures namely, the training, validation and test sets' correlation coefficients (r^2), and corresponding root mean square error of prediction (*RMSEP*), relative standard error of prediction (*RSEP*) and mean absolute error (*MAE*) values. They are calculated using the following equations.

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (y_{pred} - y_{obs})^2}{\sum_{i=1}^{n} (y_{obs} - y_{mean})^2} \quad (9)$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_{pred} - y_{obs})^2}{n}} \quad (10)$$

$$RSEP(\%) = 100\sqrt{\frac{\sum_{i=1}^{n}(y_{pred} - y_{obs})^2}{\sum_{i=1}^{n}(y_{obs})^2}} \quad (11)$$

$$MAE(\%) = \frac{100}{n}\sqrt{\sum_{i=1}^{n}\left|(y_{pred} - y_{obs})\right|} \quad (12)$$

where y_{obs} is the observed activity, y_{mean} is the mean of observed activity and y_{pred} is the predicted activity of the compound in the sample and n is the number of samples in the concerned set.

All ANN models were reassessed from the scratch for any chance correlations through simulation runs with the scrambled biological response (Y-randomization test) [33]. For this, the activity values were scrambled and ANN was again performed from the scratch with the input features of the set under investigation to establish correlation with scrambled activity. The previously optimized network weights were not used in anyway in the Y-scrambling study. Each feature set was evaluated for one hundred times with repeatedly scrambled biological response. The emerging correlation coefficients of scrambled activity were considered. This has been used to compute the percent chance correlation [100*(number of Y-scrambling correlations which were more than or equal to the correlation from Y-original)/(number of Y-scrambling trials)], average chance correlation (r^2_{Yrand}) and maximum chance correlation of the feature set under examination. The ANN computations were carried out using the MATLAB 7.6 for windows [36]. All the computations were verified with the outputs emanating from the known data [9,12]. Systat package [29] has been used to cross-check the regression and the statistical parameters used in the investigation.

3 Results and Discussion

The scope of ‘selected’ versus ‘arbitrary’ features in the development of ANN models has been investigated using three different chemical databases in conjunction with their activity profiles. Among them, the feature sets from thiazolidinones have been extensively investigated. The names of descriptors of three selection sets and four trial sets (sets-1 to 4) of thiazolidinones used for investigating the scope of discovering ANN models of HIV inhibitory activity are shown in Table 2. Among these, the descriptors of first three sets have emerged from feature selection approaches namely, CP-MLR, stepwise regression (SR) and genetic algorithm (GA). In linear regression approach, these selection sets have explained 71.1 to 78.4 per cent variance in the activity of training set compounds. Two to three variables are common between each pair of these three selection sets’ descriptors.

The descriptors of the trial sets 1 to 4 (or simply sets-1 to 4; Table 2) are arbitrary picks from the whole dataset of explanatory variables of thiazolidinones. These descriptors are different from each other (Table 2). The only exception is that, one descriptor (IC1) of trial set-2 is also present in GA set (Table 2). In stepwise regression procedure, five out of eleven descriptors of trial set-2 have entered the regression equation and explained 48.7 (r^2=0.487) per cent variance in the activity. Under same conditions, none of the descriptors of trial set-4 have entered the regression to form an equation. The Y-residuals (ΔY) from the regression of trial sets have shown a higher order of correlation with Y itself (r is between 0.73 and 0.99) than with the Y-residuals from selection sets (r is between 0.47 and 0.76). This trend is clearly noticeable with sets-3 and 4. The Y-residuals of selection sets have shown correlations (r) with each other in the range of 0.72 to 0.76, whereas the same for the trial sets is in the range of 0.70 to 0.97. Table 3 shows the absolute values of minimum and maximum correlations of the activity with the individual descriptors of each selection and trial set and the average and maximum chance correlations from Y-scrambling study. Also, attempts to find a model involving the trial sets descriptors in conjunction with their squared terms did not lead to any equation showing worthwhile correlation in comparison to the trial set descriptors alone. The cumulative descriptor-descriptor internal distances within the sets (*ID*) and the external distance in between the sets (*ED*) have demonstrated that all selection and trial sets are heterogeneous and well separated with respect to each other in the parameter space. Furthermore, the *DS* values have indicated that the regression equations of selection and arbitrarily picked feature sets have represented distant spaces. The selection sets are relatively close to each other (*DS* is 10.51 to 21.57) when compared to that of arbitrary sets (15.41 to 286.80). The selection and arbitrary sets are separated in the range of 12.78 to 123.98. In multicollinearity assessment, two pairs of features in CPMLR set have shown *VIF* between 5.82 and 7.38. The *VIF* of remaining features of this set is below 3.0. A higher *VIF* for some features in this set is due to assembling it from five regression models. The features of SR and GA sets are free from multicollinearity. Among the trial sets, in Sets-2 and 4 the *VIF* of all features is below 3.0. Some degree of multicollinearity exists between two pairs of features in Sets-1 and 3. In summary, each one of the selection and trial sets has portrayed a distinct composition in the parameter space.

The ANN architecture and network parameters used for investigating the features are shown in Table 4. The statistics emerged from the ANN models of thiazolidinones' feature sets (Table 2) are shown in Table 5. In the domain of QSAR and QSPR studies, prediction of a phenomenon (biological response and/ or physicochemical property) with a test set r^2 value in the vicinity of 0.9 is often regarded as highly significant result. In this context, regardless of the origin of descriptor sets, the ANN models from all these seven sets have well explained the HIV inhibitory activity of the compounds (r^2 is in the range of 0.91 to 0.98 for training set and 0.90 to 0.97 for test set) and qualified themselves to be sound predictive models (Table 5). The plots of observed versus ANN predicted activities from all the selection and trial sets are shown in Figure 3. Among selection sets, the best predictions were observed for GA-set (training set r^2 is 0.943; test set r^2 is 0.931) followed by SR- and CPMLR-sets. Even though the descriptors of selection sets and trial sets have originated from different underlying principles, interestingly, all four trial sets have led to equally good ANN models (Table 5). Also the analysis of Y-residuals of ANN models of selection and trial sets has indicated that they are uncorrelated with each other (r^2 is

0.02 to 0.24). Outwardly no systematic relationship is apparent between the quality of ANN model and the origin of its input variables, say either selection set or trial set (Tables 3 and 5).

The probability of obtaining an ANN model from feature sets formed arbitrarily (at random), such as trial sets 1 to 4 (Table 3), has been further investigated with one hundred additional trial sets each one with eleven randomly picked features from the original descriptors of thiazolidinones. Among these 100 trial sets, the descriptors' match between any two sets is less than or equal to three. Altogether, only nineteen sets have shown a match of three features between sets. Also, the matched features are limited to separate pairs of sets only. The *ID* and *ED* values of the additional trial sets have confirmed their diversity in descriptor space. In terms of *DS* values, the selection and additional trial sets are separated in the range of 6.35 to 543.65 with a vast majority of them showing distances more than 20.0. The additional trial sets have shown even greater diversity within them (*DS*> 1000). In multicollinearity assessment of these additional trial sets, the *VIF* of features in forty one sets is well below 5.0 (for most of the features the value is less than 3.0); in sixteen sets the *VIF* of only one feature in each set is more than 5.0 but below 10.0 (for most of the features the value is less than 7.0); in twenty two sets the *VIF* of 2 to 4 features in each set is more than 5.0 but below 10.0 (for most of the features the value is less than 7.0); in most of the remaining twenty one sets the *VIF* of one to three features in each set is more than 10.0. In the stepwise regression procedure, from these hundred sets only one set has explained the variance in the activity of training set compounds to the extent of 57.4 percent (r^2=0.574). All the remaining sets have explained the variance in the activity far below this level. The Y-residuals from the regression of additional trial sets have shown correlations (r) in the range of 0.41 to 0.72 with those from selection sets. Here, only six trial sets have shown correlation (r) more than 0.70 with the selection sets. Interestingly, in ANN (Table 4) all these hundred feature sets have led to statistically worthwhile predictive models for HIV-1 RT inhibition. From the training set perspective, the maximum and minimum r^2 values from the additional trial sets are found to be 0.946 and 0.765, respectively. Out of the hundred trial sets, the ANN models from more than 90 sets have shown a training set r^2 value greater than or equal to 0.80. From these results, the training set r^2 values of 39 sets are very good ($r^2 \geq 0.90$). For these hundred trial sets, from the test set point of view, the maximum and minimum r^2 values are 0.918 and 0.707, respectively. Also, the ANN models from 77 trial sets have shown test set r^2 values more than or equal to 0.80 of which 23 models have test set r^2 values more than or equal to 0.90. Collectively, the ANN models from these hundred sets have shown average training, validation and test set r^2 values as 0.873(sd = 0.048), 0.857(sd = 0.052) and 0.841(sd = 0.056), respectively. The analysis of Y-residuals from ANN models of these hundred trial sets and the selection sets suggested that the former ones are largely not related to the later ones. The Y-residuals from only one additional trial set (ATS067) has shown a correlation of 0.54 (r^2=0.29) with that of CPMLR set. However, among the additional trial sets, the Y-residuals from one set (ATS045) had shown correlation with the Y-residuals of five other sets (ATS021, ATS046, ATS059, ATS067, ATS089) (r^2 is 0.50 to 0.55). Among these six sets, the features of ATS021 and ATS067 have shown multicollinearity. A closer examination revealed that the features of ATS045 are correlated moderately to high order with the features of ATS021, ATS046, ATS059, ATS067 and ATS089. The Y-residuals of ATS021, ATS045, ATS046, ATS059 and ATS089 have shown correlation with the Y-residuals of CPMLR set to a lower order (r^2 is 0.20 to 0.27). As the prediction of Y in ANN is complex process, it is very difficult to pinpoint the exact nature of

relations leading to these intercorrelations. Barring these, a vast majority of the remaining intercorrelations between the Y-residuals are nominal. Figure 4 shows the distribution of ANN training and test set r^2 values of the additional trial sets against their *ID* values. This distribution does not suggest any specific pattern except for the abundance of statistically acceptable models in the whole trial sets (descriptors) space.

In a further experiment, a meaningless data of 100 columns of random numbers, each with 184 entries distributed between +1 and -1, was generated to verify the scope of such numbers in ANN to model the HIV inhibitory activity similar to those ones from the trial sets of thiazolidinones. This artificial dataset of meaningless numbers has been created using an in-house written random-number generation program. In ANN, the random number subsets, each with 11 columns of meaningless numbers, in the input (in place of molecular features of thiazolidines) did not lead to any worthwhile model for HIV inhibitory activity (data given in supporting information). This has clearly indicated that features of trial sets are embedded with information content relevant the activity and they are not meaningless numbers.

In thiazolidinones, also the influence of "parsimony" effect of input features on the outcome of ANN statistics has been investigated with appropriate selection sets from the feature selection approaches as well as with the arbitrary sets having matching number of features (Table 6). There is hardly any multicollinearity among the features of "parsimony" selection sets and corresponding arbitrary sets. Moreover, none of the Y-residuals of these sets are significantly intercorrelated. Table 6 has suggested that exclusion of one or more descriptors under "parsimony" criteria from a feature set leads to lowering the significance of the ANN statistics of the model from altered set. But the ensuing statistics of such models may still qualify the contents of the concerned sets as predictive features of the models. This trend has been noticed both with selection sets and with arbitrary feature sets (Table 6). Also, the Y-residuals (ANN) of "parsimony" selection sets and corresponding arbitrary sets are largely not related to each other (maximum r^2 is 0.25). These results have clearly suggested in favor of the possibility of discovering an ANN model of equal or comparable quality from both selection sets and arbitrary sets even after the due adjustment for the "parsimony" effect.

Since the abovementioned ANN results from the feature sets of thiazolidinones are appeared to be unusual, the investigation of 'selected' and 'arbitrary' feature sets has been extended to more databases. The two databases, anilinoquinolines and piperazinoquinolines (Table 1) are part of this exploration. Each of these databases is investigated with one selection set and three arbitrary sets. Also, the feature sets of each database are fairly independent and well separated from one another in the descriptor space in terms of *ID*, *ED* and *DS* values. In multicollinearity test of feature sets of these databases, in one set (AQS4) the *VIF* of two features is more than 5.0 (5.43 and 8.70); for all other sets, *VIF* is far less than 3.0. In linear regression, the selection sets of anilinoquinolines and piperazinoquinolines have respectively explained 71.2 (r^2=0.712) and 67.0 (r^2=0.67) per cent variance in their corresponding activities. In the arbitrary sets of these databases no such relation could be found with the activity. Also, none of the Y-residuals of these sets are markedly intercorrelated. The ANN network parameters for these feature sets are shown in Table 4. Here as well, for both the databases, the statistics from the ANN models of arbitrary sets are comparable to the statistics obtained from the corresponding selection set

(Table 7). Also, the ANN Y-residuals from selection and arbitrary sets are independent with respect to each other. We have examined two more databases namely, flavones as aldose reductase inhibitors [37] and solubility estimation of drugs/ drug like molecules [38]. The results obtained from them too are in conformity with that of other databases reported herein. They are shown in supporting information.

In any data, the feature subsets identified through selection procedures hold well-ordered (regimented) information towards the Y. The arbitrarily formed feature subsets also hold information corresponding to the Y, but this information may not be a regimented one in terms of selection procedures. Also, often several of objects' definitions show different degrees of overlap with each other. It is well-known that in different classes of compounds the molar refractivity, molecular volume and hydrophobicity show overlap to different degree. In the composite data matrix of each database used in this investigation, many a descriptors have shown different degrees of intercorrelations, from very good to nominal values, with the other descriptors present in the data. For example in thiazolidinones, the descriptors MOR02m, GGI10, MOR29e and IC1 of GA-set (selection set) are partly correlated (r is between 0.34 and 0.64) with the descriptors MOR30m, Sv, E2s, Hy of Set-4 (arbitrary set). This study has involved many a subsets which are seemingly far less interrelated with each other than the just mentioned ones. The arbitrary sets of this investigation may be viewed as unregimented ones with varied information corresponding to Y and at the same time not amenable to linear modeling approaches. ANN being a powerful non-linear modeling procedure, it has led this unregimented information of arbitrary subsets to evolve into predictive models. The investigation reveals how different feature subsets have performed in ANN in accounting for the variance in Y.

All these results have prompted us to state that prior feature selection may not be essential for the development of an ANN model. Nevertheless, the initial rationales from a selection set offer insight into the structural requirements of the phenomenon (activity) under investigation. For example, in CPMLR selection set of thiazolidinones, the regression coefficient of AECC (average eccentricity; topological index) suggested in favor of decreasing eccentricity in the molecule for the activity; the descriptor GGI10 (topological charge index of order 10; topological charge index) suggested the influence of charge on vertex pairs at topological distance equal to ten on the activity. Similarly, the influence of other selection sets' descriptors on the activity can be reasoned. Even though several arbitrary sets have led to equally good ANN models, it is difficult to understand how these features can be adjusted to modulate the biological activity. Except for this, the performance of arbitrary sets in ANN can not be negated. In this context, descriptors from features selection approaches offer initial rationales for the activity prior to their incorporation in more complex nonlinear models. For this reason, the application of 'selected' input features in deriving ANN model(s) is more a desirable option for reasonable outputs.

4 Conclusions

The ANN outcomes from objectively selected and arbitrarily chosen input features has been investigated using the molecular descriptors of three databases namely thiazolidinones, anilinoquinolines and piperazinoquinolines in conjunction with their respective biological activities as dependent variables. The feature subsets analyzed in the investigation have shown

fair independence to some degree of multicollinearity. Regardless of the origin of input feature sets, a majority of ANN models from each database have well explained the activity and qualified themselves to be sound predictive models. These results position ANN as a powerful tool to identify the patterns in the data.

The importance of feature selection for any modeling study can not be understated. Especially in ANN, in the absence of any prior rationale for incorporating a set of features in the input, the physical meaning remains obscure. Nevertheless, the lack of a valid selection approach for a feature set can not reverse its predictive performance in ANN. As a corollary, the results clearly suggested that prior feature selection may not be essential to feed input variables to ANN to develop a model. However, from the diagnostic perspective it is difficult to understand how the features of arbitrary set can be adjusted to modulate the biological activity. In contrast, the prior rationales observed for the features of selection sets provide direction for the modification of the chemical space to modulate the activity. In this context, application of features from selection procedures to ANN input is more a desirable option for meaningful outputs in terms of rationale behind the input variables.

This report is neither a proposal for a modeling approach nor an argument in favor of unregimented sets for modeling a phenomenon. It emphasizes the innate ability of ANN to navigate the feature sets to develop complex network of relations connecting them in order to estimate the target variable. Elaborated results are given in the supporting information. It neither negates the importance of feature selection in ANN nor refutes the role of ANN in modeling approaches. A vast literature is available on feature generation, selection and procedural aspects of ANNs. Only selected works are cited to limit the list of references.

Acknowledgments

Authors thank Dr. M. Abbas, Department of Biometry & Statistics, CDRI, for helpful discussions, and the anonymous referees for valuable comments in the revision of manuscript. SD and VM gratefully acknowledge the ICMR and CSIR respectively, for the research fellowships. CDRI Communication No. 7596.

Supporting Information

Databases of thiazolidinones, anilinoquinolines, piperazinoquinolines, flavones and various drug/ drug-like moleculs; feature sets; *ID*, *ED*, *DS* and *VIF* values; linear regression and ANN statistics; activity/ property predictions in ANN.

References

[1] S. C. Basak, D. K. Harriss, V. R. Magnuson, POLLY, University of Minnesota: Duluth, MN, **1988**.

[2] A.R. Katritzky, V. Lobnov, M. Karelson, *CODESSA* (*Comprehensive Descriptors for Structural and Statistical Analysis*) University of Florida, Gainesville, FL **1994**, www.codessa–pro.com.

[3] Molconn–Z, EduSoft LC, a Virginia Corporation, Ashland, VA 23005 USA. **2002**, www.edusoft–lc.com.

[4] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON software (version 3.0-**2003**), Milano, Italy, and references cited therein. http://www.talete.mi.it./

[5] Z. R. Li, L. Y. Han, Y. Xue, C. W. Yap, H. Li, L. Jiang, Y. Z. Chen, *Biotechnol. Bioeng.* **2007**, *97*, 389-396.

[6] J. H. Wikel, E. R. Dow, *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645-651.

[7] H. Kubinyi, *Quant. Struct.–Act. Relat*. **1994**, *13*, 285–294.

[8] D. Rogers, A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854-866.

[9] S.-S. So, M. Karplus, *J. Med. Chem.* **1996**, *39*, 1521-1530.

[10] a) I. V. Tetko, A. E. P. Villa, D. J. Livingstone, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794-803. b) D. J. Livingstone, D.W. Salt, *Rev. Comput. Chem.* **2005**, *21*, 287-348.

[11] C. L.; Waller, M. P. Bradley, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 345–355.

[12] Y. S. Prabhakar, *QSAR Comb. Sci.* **2003**, *22*, 583–595.

[13] Q. Shen, J.–H. Jiang, J.–C. Tao, G.–L. Shen, R.–Q. Yu, *J. Chem. Inf. Model.* **2005**, *45*, 1024–1029.

[14] a) F. Rosenblatt, *Psychol. Rev.* **1958**, *65*, 386–408. b) J. J. Hopfield, *Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 2554–2558. c) G. Reibnegger, G. Weiss, G. Werner-Felmayer, G. Judmaier, H. Wachter, *Proc. Nati. Acad. Sci. U.S.A.* **1991**, *88*, 11426-11430. d) J. Gasteiger, J. Zupan, *Angew. Chem. Intl. Ed. Engl.* **1993**, *32*, 503-527.

[15] F. Marini, R. Bucci, A. L. Magrì, A. D. Magrì, *Microchemical Journal* **2008**, *88*, 178–185.

[16] a) D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Nature*, **1986**, *323(6088)*, 533–536. b) C.M. Bishop, *Neural Networks for Pattern Recognition*; Clarendon Press: Oxford, **1995**; pp 116-163 and 253-294. c) D. Graupe, *Principles of Artificial Neural Networks*, 2nd ed. World Scientific Publishing Co.: Singapore; **2007**; pp 59-111.

[17] a) D.T. Manallack, D. D. Ellis, D. J. Livingstone, *J. Med. Chem.* **1994**, *37*, 3758-3767. b) I. V. Tetko, D. J. Livingstone, A. I. Luik, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826–833. c) D. T. Manallack, D. J. Livingstone, *Eur. J. Med. Chem*. **1999**, *34*, 195-208. d) I. Kovesdi, M. F. Dominguez–Rodriguez, L. Orfi, G. Naray–Szabo, A. Varro, J. G. Papp, P. Matyus, *Med. Res. Rev.* **1999**, *19*, 249–269. e) B. Hemmateenejad, M. Akhond, R. Miri, M. Shamsipur, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1328–1334. f) M. Jalali-Heravi, A. Kyani, *J. Chem. Inf. Comput.*

Sci. **2004**, *44*, 1328-1335. g) L. Yang, P. Wang, Y. Jiang, Chen, J. *J. Chem. Inf. Model.* **2005**, *45*, 1804–1811. h) J. R. Votano, M. Parham, L. M. Hall, L. H. Hall, L. B. Kier, S. Oloff, A. Tropsha, *J. Med. Chem.* **2006**, *49*, 7169-7181. i) O. Deeb, B. Hemmateenejad, *Chem. Biol. Drug Des.* **2007**, *70*, 19–29. j) M. Goodarzi, M. P. Freitas, *QSAR Comb. Sci*. **2008**, *27*, 1092-1098.

[18] a) A. Rao, J. Balzarini, A. Carbone, A. Chimirri, E. D. Clercq, A. M. Monforte, P. Monforte, C. Pannecouque, M. Zappalá, *Il Farmaco* **2002**, *57*, 747-751. b) M. L. Barreca, J. Balzarini, A. Chimirri, E. D. Clercq, L. D. Luca, H. D. Höltje, M. Höltje, A. M. Monforte, P. Monforte, C. Pannecouque, A. Rao, M. Zappalà, *J. Med. Chem.* **2002**, *45*, 5410-5413. c) A. Rao, A. Carbone, A. Chimirri, E. D. Clercq, A. M. Monforte, P. Monforte, C. Pannecouque, M. Zappalá, *Il Farmaco* **2003**, *58*, 115-120. d) A. Rao, J. Balzarini, A. Carbone, A. Chimirri, E. D. Clercq, A. M. Monforte, P. Monforte, C. Pannecouque, M. Zappalá, *Il Farmaco* **2004**, *59*, 33–39. e) A. Rao, J. Balzarini, A. Carbone, A. Chimirri, E. D. Clercq, A. M. Monforte, P. Monforte, C. Pannecouque, M. Zappalá, *Antiviral Research* **2004**, *63*, 79–84.

[19] a) R. K. Rawal, Y. S. Prabhakar, S. B. Katti, E. D. Clercq, *Bioorg. Med. Chem*. **2005**, *13*, 6771–6776. b) R. K. Rawal, R. Tripathi, S. B. Katti, C. Pannecouque, E. D. Clercq, *Bioorg. Med. Chem.* **2007**, *15*, 1725–1731. c) R. K. Rawal, R. Tripathi, S. B. Katti, C. Pannecouque, E. D. Clercq, *Bioorg. Med. Chem.* **2007**, *15*, 3134–3142. d) R. K. Rawal, R. Tripathi, S. B. Katti, C. Pannecouque, E. D. Clercq, *Medicinal Chemistry*, **2007**, *3*, 355-363. e) R. K. Rawal, R. Tripathi, S. B. Katti, C. Pannecouque, E. D. Clercq, *Eur. J. Med. Chem.* **2008**, *43*, 2800-2806. f) R. K. Rawal, R. Tripathi, S. Kulkarni, R. Paranjape, S. B. Katti, C. Pannecouque, E. D. Clercq, *Chem. Biol. Drug Des.* **2008**, *72*, 147-154.

[20] R. Pauwels, J. Balzarini, M. Baba, R. Snoeck, D. Schols, P. Herdewijin, J. Desmyter, E. D. Clercq, *J. Virol. Meth*. **1988**, *20*, 309-321.

[21] (a) S. Delarue, S. Girault, L. Maes, M. D-Fontaine, M. Labaeıd, P. Grellier, C. Sergheraert.. *J. Med. Chem*. **2001**, *44*, 2827-2833. (b) S. D-Cochin, E. Paunescu, L. Maes, E. Mouray, C. Sergheraert, P. Grellier, P. Melnyk. *Euro. J. Med. Chem.* **2008**, *43*, 252-260. (c) E. D-Charvet, S. Delarue, C. Biot, B. Schwo¨bel, C. C. Boehme, A. Mu¨ ssigbrodt, L. Maes, C. Sergheraert, P. Grellier, R. H. Schirmer, K. Becker. *J.Med.Chem.* **2001**, *44*, 4268-4276

[22] S. Deshpande, V. R. Solomon, S. B. Katti, Y. S. Prabhakar. *J. Enzym. Inhbn. Med. Chem* **2009**, *24(1)*, 94-104 and references sited there in.

[23] a) D. W. Marquardt, *J. Soc. Ind. Appl. Math.* **1963**, *11*, 431-441. b) M. T. Hagan, M. B. Menhaj, *IEEE Trans. Neural Networks* **1994**, *5*, 989-993.

[24] SYBYL 7.3., Tripos Inc., 1699 South Hanley Road, Suite 303, St. Louis, MO 63144, USA.

[25] F. Nicoló, G. Bruno, R. Scopelliti, S. Grasso, A. Rao, M. Zappalá, *Acta Cryst.* **2001**, *C57*, 572-574.

[26] A. Semeniuk, A. Niedospial, J. K.-Tluscik, W. Nitek, B. J. Oleksyn. *J. Mol. Struct*. **2008**, *875*, 32–41.

[27] M. Clark, R. D. III. Cramer, N. V. Opdenbosch, *J. Comput. Chem.* **1989**, *10*, 982-1012.

[28] MOE: The Molecular Operating Environment from Chemical Computing Group Inc., 1255 University St., Suite 1600, Montreal, Quebec, Canada H3B 3X3. http://www.chemcomp.com.

[29] SYSTAT, Version 7.0: SPSS Inc., 444 North Michigan Avenue, Chicago, IL 60611.

[30] R. Todeschini, V. Consonni, M. Pavan, MOBYDIGS software (Version 1.2) for Windows, Talete Srl, Milan, Italy, **2002**. http://www.talete.mi.it/mobydigs.htm.

[31] M. Pavan, A. Mauri, R. Todeschini, *Anal. Bioanal. Chem.* **2004**, *380*, 430-444.

[32] a) Y. S. Prabhakar, V. R. Solomon, R. K. Rawal, M. K. Gupta, S. B. Katti, *QSAR Comb. Sc.* **2004**, *23*, 234-244. b) Y. S. Prabhakar, R. K. Rawal, M. K. Gupta, V. R. Solomon, S. B. Katti, *Comb. Chem. High-Throughput Scr.* **2005**, *5*, 431-437. c) M. Saquib, M. K. Gupta, R. Sagar, Y. S. Prabhakar, A. K. Shaw, R. Kumar, P. R. Maulik, A. N. Gaikwad, S. Sinha, A. K. Srivastava, V. Chaturvedi, R. Srivastava, B. S. Srivastava, *J. Med. Chem.* **2007**, *50*, 2942-2950.

[33] S.-S. So, M. Karplus, *J. Med. Chem.* **1997**, *40*, 4347-4359.

[34] J. D. Curto, J. C. Pinto, New multicollinearity indicators in linear regression models. *Int. Statist. ReV*. **2007**, *75*, 114–121.

[35] a) R. A. Johnson, D. W. Wichern, Applied multi Variate statistical analysis; Prentice-Hall: New York, 1992. b) M. H. Kutner, J. Nachtsheim, J. Neter, Applied Linear Regression Models; Fourth Edition, McGraw-Hill/Irwin, New York 2004.

[36] MATLAB, Version 7.6: <http://www.mathworks.com/products/matlab/>.

[37] Y. S. Prabhakar, M. K. Gupta, N. Roy, Y. Venkateswarlu, *J. Chem. Inf. Model.* **2006**, *46,* 86-92.

[38] A. Llinàs, R. C. Glen, J. M. Goodman, *J. Chem. Inf. Model.*, **2008**, *48*, 1289–1303.

Legend for Figures

Figure 1. General structures of (a) thiazolidinones, (b) anilinoquinolines and (c) piperazinoquinolines.

Figure 2. The architecture of back propagation function neural network diagram showing the input, hidden and output layers. In this, x is input vector, c_j is back propagation function center, d_j is back propagation function width, h_j is j^{th} output of hidden node, y_k is k^{th} output node, w_{kj} is weight connection between the k^{th} output node and the j^{th} hidden layer node, b_k is k^{th} bias and e_k is error due to output y_k.

Figure 3. The plots of observed versus predicted activities of thiazolidinones' training (◇), validation (○) and test (△) sets from BP-ANN. The solid line passing through the data points correspond to the least-squares fit of training set. The dashed line passing through the origin, making an angle of 45° with the axis, bisects the plot area.

Figure 4. Distribution of ANN training and test set r^2 values of thiazolidinones from the additional trial sets against their *ID* values.

Table 1. Composition of training, validation and test sets and the activity distribution therein.

Sample sets	Sample size			Activity spread					
	Train	Valid	Test	Train		Valid		Test	
				High	Low	High	Low	High	Low
Thiazolidinones	92	48	48	7.77	3.45	7.36	4.02	7.52	4.24
Anilinoquinolines	38	18	18	8.33	5.83	8.30	6.33	8.43	6.51
Piperazinoquinolines	22	11	11	9.05	6.58	8.13	6.67	8.30	7.20

Table 2. The names of descriptors in each selection and some trial sets of thiazolidinones.[a]

CPMLR	SR	GA	Set-1	Set-2	Set-3	Set-4
AECC[b] (Topo)	nCL (Const)	MAXDP (Topo)	Mw (Const)	J (Topo)	MATS3v (2D-Auto)	Sv (Const)
GGI10	Pol[b]	IC1[b]	ARR	IC1 (Topo)	MATS2e	nS[b]

(Galvez)	(Topo)	(Topo)	(Emp)		(2D-Auto)	(Const)
GATS5p (2D-Auto)	T(O..Br) (Topo)	GGI10 (Galvez)	MLOGP (Prop)	MATS3e (2D-Auto)	MATS4p (2D-Auto)	nCrH2[b] (Funct)
RDF075m[c] (RDF)	MAXDP (Topo)	MATS7e[b] (2D-Auto)	MSD (Topo)	MATS4e[b] (2D-Auto)	GATS6e (2D-Auto)	Hy (Emp)
Mor29u (3D-Mors)	RDF075m[b] (RDF)	DISPv (Geo)	JhetZ[b] (Topo)	GATS3v (2D-Auto)	L1u[b] (Whim)	X1sol (Topo)
Mor02m (3D-Mors)	Mor29u (3D-Mors)	G(O..S) (Geo)	X0Av (Topo)	RDF030m[b] (RDF)	E1v[b] (Whim)	MATS5v (2D-Auto)
Mor13m (3D-Mors)	Mor26p (3D-Mors)	RDF020p (RDF)	ATS4m (2D-Auto)	Mor07p[b] (3D-Mors)	L2p (Whim)	Mor27u (3D-Mors)
Mor18m (3D-Mors)	L3v[b] (Whim)	Mor02m (3D-Mors)	MATS2m (2D-Auto)	Mor32p (3D-Mors)	E2p (Whim)	Mor30m (3D-Mors)
Mor32m (3D-Mors)	E3m[b] (Whim)	Mor10e (3D-Mors)	Mor02u (3D-Mors)	E3u (Whim)	L3s (Whim)	Mor21e (3D-Mors)
Mor29v[b] (3D-Mors)	E1s (Whim)	Mor29e[b] (3D-Mors)	P2m[b] (Whim)	L3m (Whim)	Tm (Whim)	G1u (Whim)
E3e (Whim)	E3e (Whim)	E3e (Whim)	R8v+[b] (Get)	Ks (Whim)	Tp[b] (Whim)	E2s[b] (Whim)

[a] Descriptor (descriptor class).The 0D-2D descriptors are shown in gray shade. The descriptor definitions are arranged under the respective descriptor classes.**(Const)**, Constitutional: nCL, number of Chlorine atoms; Mw, molecular weight; Sv, sum of atomic van der Waals volumes (scaled on Carbon atom); nS, number of Sulfur atoms. **(Topo)**, Topological: AECC, average eccentricity; Pol, polarity; T(O..Br), sum of topological distances between O..Br; MAXDP, maximal electrotopological positive variation; IC1, information content index (neighborhood symmetry of 1-order); MSD, mean square distance index (Balaban); JhetZ, Balaban-type index from Z weighted distance matrix (Barysz matrix); X0Av, average valence connectivity index chi-0; J, Balaban distance connectivity index; X1sol, solvation connectivity index chi-1. **(Galvez)**, Galvez Topological charge: GGI10, topological charge index of order 10 topological charge indices. **(2D-Auto)**, 2D autocorrelations: ATS*kw*, MATS*kw*, and GATS*kw*, are, respectively, Broto-Moreau autocorrelation, Moran autocorrelation and Geary autocorrelation of '*k*' path length (*lag*) weighted by atomic property '*w*' ('*w*' is weighing atomic property where, (u)-unweighted case, (m)-atomic mass, (v)-van der Waals volume, (e)-Sanderson atomic electronegativity and, (p)-atomic polarizability). **(Funct)**, functional groups: nCrH2, number of ring secondary C(sp3); **(Emp)**, Empirical: ARR, aromatic ratio; Hy, hydrophilic factor. **(Prop)**, Molecular properties: MLOGP, Moriguchi octanol-water partition coeff. (logP). **(Geo)**, Geometrical: DISPv, d COMMA2 value / weighted by atomic van der Waals volumes; G(O..S), sum of geometrical distances between O..S. **(RDF)**, Radial Distribution Function: RDF*Rw*, RDF - in a spherical volume of radius R (R is 1 to 15.5Å with an increment of 0.5Å and expressed in integer as 010, 015, 020,…..,155) weighted by atomic property 'w'. **(3D-Mors)**, 3D-MoRSE: Mors*w*, scattered electron intensity in direction '*s*' (*s* is 0 to 31) / weighted by atomic property '*w*'. **(WHIM)**, Weighted Holistic Invariant Molecular descriptors): L*kw* / E*kw* / P*kw* / G*kw*, k^{th} component (*k* is1 to 3) 'L (size) / E (accessibility) / P (shape) / G (symmetry)' directional WHIM index / weighted by atomic property '*w*'; T*w*, T total size index / weighted by atomic property '*w*'; K*w*, K global shape index /

weighted by atomic property '*w*'. **(Get)**, GEometry, Topology, and Atom-Weights AssemblY: R8v+, R maximal autocorrelation of lag 8 / weighted by atomic van der Waals volumes. Also see ref 4.

[b] not present in the parsimony sets (Table 6)

[c] not present in Set-5 but present in Set-8 (Table 6)

Table 3. Some characteristic correlations of anti-HIV activity with the selection and trial set descriptors of thiazolidinones in linear mode.

Feature Set	r^2 min[a]	r^2 max[a]	r^2 (des)[b]	r^2_{Yrand}(max)[c]
CPMLR	0.026	0.340	0.710 (9)	0.092(0.215)
SR	0.019	0.293	0.784 (11)	0.122(0.270)
GA	0.021	0.340	0.757 (11)	0.117(0.260)
Set-1	0.046	0.301	0.419 (3)	0.024(0.089)
Set-2	0.162	0.276	0.487 (5)	0.048(0.208)
Set-3	0.040	0.145	0.214 (2)	0.021(0.180)
Set-4	9.0E-6	0.0005	0.000 (0)	0.020(0.104)[d]

[a] Single descriptor versus activity.

[b] Multiple correlation coefficient of the descriptors with activity in stepwise regression (number of descriptors entered in equation from each set).

[c] Mean correlation coefficient of the regression model in Y-scrambling study from 100 simulations (max, maximum correlation observed in Y-scrambling); percentage chance correlation is zero for all described models.

[d] For the purpose of Y-scrambling, first two descriptors of Set-4 (Table 2) were considered.

Table 4. The ANN architecture and its parameters.[a]

Sample sets	Nodes		
	Input	Hidden	output
Thiazolidinones	11+1[b]	7	1
Anilinoquinolines	5+1[b]	5	1
Piperazinoquinolines	3+1[b]	3	1

[a] Optimization algorithm is Levenberg-Marquardt; transfer function is sigmoid; Iterations 10 to 25.

[b] Bias

Table 5. Goodness of fit of training, validation and test compounds' activity with the selection and trial sets (sets with fixed number of features) of thiazolidinones in BP-ANN models.

Feature Set	μ[a]	α	r^2			*RMSEP*			*RSEP*(%)			*MAE*(%)			r^2_{Yrand}(max)
			Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	
CPMLR	0.27	0.41	0.911	0.900	0.903	0.336	0.244	0.228	5.786	4.003	3.709	5.14	6.107	5.986	0.105(0.240)
SR	0.24	0.33	0.922	0.918	0.913	0.298	0.226	0.217	5.133	3.704	3.528	4.826	6.494	6.586	0.168(0.353)
GA	0.23	0.40	0.943	0.934	0.931	0.261	0.199	0.184	4.493	3.272	2.996	4.731	5.887	5.875	0.137(0.314)
Set 1	0.24	0.36	0.921	0.912	0.917	0.324	0.231	0.208	5.588	3.796	3.395	5.131	5.961	5.924	0.129(0.276)
Set 2	0.30	0.41	0.923	0.903	0.910	0.301	0.242	0.216	5.180	3.969	3.518	5.082	6.265	6.145	0.176(0.321)
Set 3	0.31	0.41	0.983	0.974	0.976	0.141	0.123	0.118	2.427	2.017	1.926	3.471	4.705	4.363	0.139(0.287)
Set 4	0.35	0.41	0.937	0.914	0.902	0.273	0.284	0.272	4.703	4.636	4.427	4.692	6.932	6.737	0.129(0.276)

[a] μ, Learning rate; α, momentum; r^2, squared correlation coefficient; *RMSEP*, root mean square error of prediction; *RSEP*, relative standard error of prediction; *MAE*, mean absolute error. See equations 7-10. r^2_{Yrand}(max), mean correlation coefficient of the ANN model in Y-scrambling study from 100 simulations (max, maximum random correlation observed); percentage chance correlation is zero for all described models..

Table 6. Goodness of fit of training, validation and test compounds' activity with the features of "parsimony" sets in BP-ANN models.

Feature Set[a]	μ	α	r^2			*RMSEP*			*RSEP*(%)			*MAE*(%)			r^2_{Yrand}(max)
			Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	
Set-5 (8)	0.32	0.54	0.830	0.825	0.802	0.438	0.352	0.333	7.550	5.767	5.419	6.102	7.945	7.736	0.101(0.202)
Set-6 (7)	0.42	0.57	0.818	0.811	0.804	0.451	0.372	0.311	7.779	6.109	5.056	6.399	8.041	7.440	0.118(0.243)
Set-7 (7)	0.38	0.59	0.855	0.848	0.836	0.403	0.328	0.316	6.938	5.378	5.139	5.958	7.548	7.503	0.114(0.306)
Set-8 (9)	0.34	0.57	0.874	0.851	0.842	0.375	0.309	0.293	6.461	5.069	4.773	5.333	7.521	7.134	0.115(0.289)
Set-9 (8)	0.35	0.57	0.898	0.889	0.880	0.339	0.266	0.250	5.849	4.364	4.064	5.269	6.734	6.513	0.097(0.285)
Set-10 (8)	0.38	0.60	0.843	0.837	0.825	0.450	0.385	0.332	7.749	6.312	5.413	6.270	8.174	7.590	0.156(0.332)
Set-11 (8)	0.40	0.63	0.831	0.830	0.823	0.464	0.316	0.295	7.998	5.178	4.800	6.325	7.708	7.218	0.121(0.273)
Set-12 (8)	0.34	0.61	0.901	0.900	0.887	0.363	0.260	0.260	6.250	4.259	4.226	5.551	6.651	6.772	0.109(0.291)
Set-13 (8)	0.45	0.67	0.832	0.829	0.818	0.479	0.367	0.321	8.249	6.016	5.229	6.478	7.957	7.595	0.126(0.283)

[a] The number in the parentheses refers to number of features in the set. Sets 5 and 8 are subsets CPMLR-set, set 6 is subset to SR-set, set-9 is subset to GA-set, set-7 is from GA approach with descriptors namely PCWTe, BELp4,

Mor26m, Mor29u, BELe3, ASP and GATS5p, and sets 10 to 13 are respectively subsets to trail sets 1 to 4. See footnotes of Tables 2 and 5.

Table 7. Goodness of fit of training, validation and test compounds' activity with the selection and trial sets of anilinoquinolines and piperzinoquinolines in BP-ANN models.

Feature Set	μ[a]	α	r^2			*RMSEP*			*RSEP*%			*MAE*%			r^2_{Yrand}max
			Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	
Anilinoquinolines															
AQS1	0.63	0.68	0.96	0.93	0.90	0.15	0.18	0.22	2.07	2.35	2.83	5.54	9.11	10.37	0.12(0.24)
AQS2	0.59	0.63	0.91	0.89	0.87	0.23	0.21	0.27	3.08	2.76	3.56	6.91	9.35	10.64	0.10(0.32)
AQS3	0.66	0.69	0.85	0.82	0.79	0.28	0.28	0.29	3.88	3.75	3.84	7.37	11.01	11.40	0.10(0.21)
AQS4	0.60	0.73	0.89	0.86	0.84	0.25	0.25	0.28	3.37	3.28	3.68	6.93	10.75	10.68	0.15(0.22)
Piperzinoquinolines															
PQS1	0.61	0.67	0.94	0.92	0.91	0.14	0.14	0.14	1.79	1.85	1.79	6.80	10.29	9.99	0.10(0.25)
PQS2	0.68	0.70	0.81	0.80	0.79	0.26	0.19	0.17	3.38	2.52	2.23	9.31	11.65	10.00	0.07(0.16)
PQS3	0.58	0.64	0.81	0.80	0.79	0.30	0.26	0.17	3.84	3.37	2.16	9.53	13.23	10.31	0.09(0.19)
PQS4	0.54	0.69	0.82	0.81	0.80	0.27	0.22	0.14	3.48	2.87	1.83	9.78	12.23	10.42	0.11(0.25)

[a] See footnote of Table 5.