Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program

Igor V. Tetko*,[†,‡] and Vsevolod Yu. Tanchuk[†]

Virtual Computational Chemistry Laboratory Project, Biomedical Department, IBPC, Ukrainian Academy of Sciences, Murmanskaya 1, Kyiv, 02094, Ukraine, and Laboratoire de Neuro-Heuristique, Institut de Physiologie, Rue du Bugnon 7, Lausanne, CH-1005, Switzerland

Received March 22, 2002

This article provides a systematic study of several important parameters of the Associative Neural Network (ASNN), such as the number of networks in the ensemble, distance measures, neighbor functions, selection of smoothing parameters, and strategies for the user-training feature of the algorithm. The performance of the different methods is assessed with several training/test sets used to predict lipophilicity of chemical compounds. The Spearman rank-order correlation coefficient and Parzen-window regression methods provide the best performance of the algorithm. If additional user data is available, an improved prediction of lipophilicity of chemicals up to 2−5 times can be calculated when the appropriate smoothing parameters for the neural network are selected. The detected best combinations of parameters and strategies are implemented in the ALOGPS 2.1 program that is publicly available at http://www.vcclab.org/lab/alogps.

INTRODUCTION

The Associative Neural Network (ASNN) represents a new challenge for development of physicochemical data prediction methods. This method is introduced as a combination of *k*-nearest neighbor and artificial neural network methods.[1−3] The new type of neural networks has memory that can coincide with the training set. Inclusion of new data to the ASNNs memory provides an extension of the Optimal Prediction Space[4] (OPS) of this method. This makes it possible to dramatically improve prediction performance of ASNN for the user's data without the need to retrain the neural network ensemble.

The ALOGPS 2.0 package[5] included programs to predict lipophilicity[6] and aqueous solubility[7] of chemical compounds. The program to predict lipophilicity was developed using the Efficient Partition Algorithm[8] and an ASNN approach. However, when developing ALOGPS 2.0 we did not systematically investigate parameters of the ASNN algorithm that could further improve the predictive ability of the neural networks. No study was also performed to detect the best strategies for the user-training feature of the algorithm. This current study compensates these drawbacks and carefully analyzes the ASNN parameters using several examples for prediction of lipophilicity of chemical compounds.

DATA SETS

PHYSPROP Data Set. The PHYSPROP database[9] used in the current study included 13 360 compounds (May 2000) with experimental values for lipophilicity (logP) from a diverse range of chemical compounds. Some of the compounds, namely metal-containing compounds (190), compounds that did not contain any carbons (11) and duplicates (251), such as L and D stereoisomers, were excluded from the analysis. The remaining set of 12 908 molecules was used to develop the lipophilicity program[6] that is a part of the ALOGPS 2.0 package. This set and several of its subsets, described below, were used to evaluate the predictive abilities of the differing ASNN methods.

2 × 6454 Sets. These were equal size sets selected randomly from the whole PHYSPROP data set. One set was used to train ASNN, and the second set was used to test the performance of the neural networks.

Star, Nova, and XLOGP Sets. The PHYSPROP database contained many compounds that were from the BioByte[10] Starlist. These molecules (9429), for which CLOGP v. 4.0 program provided some experimental values, were named as the star set. The remaining set of 3479 compounds was referred to as the nova set. The XLOGP set corresponds to the set of 1853 molecules used to develop XLOGP[11] program. It was previously shown that both XLOGP and star sets were not diverse enough to represent all molecules in the PHYSPROP database.[6]

Quinazolones. The last set of 18 quinazolones with experimental lipophilicity values[12] was used to demonstrate the user-training feature of the ACD/logP program.[13] This set was applied to study different strategies for the user-training feature of the ASNN program.

METHODS

This section provides a short description of ASNNs and the molecular parameters used in this study.

Associative Neural Network. Let us consider an ensemble of *M* neural networks

* Corresponding author phone: +49-89-3187-3575; fax: +49-89-3187-3585; e-mail: itetko@vcclab.org. Current address: Institute for Bioinformatics, GSF − Forschungszentrum für Umwelt und Gesundheit, GmbH, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany.

[†] Ukrainian Academy of Sciences.

[‡] Institut de Physiologie.

10.1021/ci025515j CCC: $22.00 © 2002 American Chemical Society
Published on Web 08/09/2002

$$[\mathbf{ANNE}]_M = \begin{bmatrix} ANN_1 \\ \vdots \\ ANN_j \\ \vdots \\ ANN_M \end{bmatrix} \tag{1}$$

The prediction of a case $\mathbf{x}_i$, $i = 1, ..., N$ can be represented by a vector of output values $\mathbf{z}_i = \{z^i_j\}^M_{j=1}$ where $j = 1, ..., M$ is the index of the network within the ensemble.

$$[\mathbf{x}_i] \cdot [\mathbf{ANNE}]_M = [\mathbf{z}_i] = \begin{bmatrix} z^i_1 \\ \vdots \\ z^i_j \\ \vdots \\ z^i_M \end{bmatrix} \tag{2}$$

A simple average

$$\bar{z}_i = \frac{1}{M} \sum_{k=1,M} z^i_k \tag{3}$$

is usually used to predict the test cases with a neural network ensemble.[14-16] This average neglects the predictions (variations) of individual neural networks. The ASNN uses the variance of individual networks to introduce a similarity measure of data case in the output space as

$$\xi_{ij} = ||\mathbf{z}^i, \mathbf{z}^j|| \tag{4}$$

where $||\cdot||$ denotes some proximity measure between vectors, $\mathbf{z}^i$ and $\mathbf{z}^j$. Equation 4 estimates similarity of cases in the output space or space of neural networks models. This measure is used to identify for each analyzed data case i a number of nearest neighbors $N_k(\mathbf{x}_i)$. The ASNN corrects the ensemble value $\bar{z}_i$ according to formula

$$\bar{z}'_i = \bar{z}_i + \sum_{j \in N_k(\mathbf{x}_i)} (y_j - \bar{z}_j) F(\xi_{ij}) \Big/ \sum_{j \in N_k(\mathbf{x}_i)} F(\xi_{ij}) \tag{5}$$

where $(y_j - \bar{z}_j)$ is the Artificial Neural Network Ensemble (ANNE) prediction error of the case j, $F()$ is some weight (kernel) function, and the summation is over the k-nearest neighbor cases determined using the similarity ξ_{ij}. The term $\sum_{j \in N_k(\mathbf{x}_i)} F(\xi_{ij})$ is used for normalization. This equation provides a correction for systematic bias of the neural network ensemble at the analyzed point i.[1-3]

The set of nearest neighbors, i.e., molecules, used in eq 5 represents ASNN memory. This set can be different from the training set used to develop ASNN, e.g. it can contain some private user's molecules that were not available to train and to optimize neural network weights. An option to extend ASNN memory using fresh data ***after*** neural network training, i.e., without changing neural network weights, represents a considerable advantage of the ASNN compared to the traditional neural network approach. This option makes it possible to dramatically improve the predictive ability of the ASNN for unseen data, such as private data of pharmaceutical firms, without the need to retrain the network.[2,3]

Neural Network Training. The neural networks used in the current study were trained using Early Stopping over Ensemble (ESE) method.[8,14,16] In this method initial training sets were randomly constructed with equal size learning and validation sets for each neural network in the ensemble. Thus, each neural network had its own learning and validation sets. The learning set was used to adjust neural network weights. The training was stopped when a minimum error for the validation set was calculated ("early stopping" point).[16] Following ensemble learning a simple average of all networks given by eq 3 was used to predict the test patterns. The updating of neural network weights was performed using Levenberg−Marquardt algorithm.[17] This is a second-order algorithm, i.e., both first- and second-order derivatives of the error function are required for the weight optimization. The Levenberg−Marquardt algorithm usually does not fall in local minima and provides the smallest errors for the fixed number of hidden neurons.[18] This algorithm was particularly useful for lipophilicity data analysis, as demonstrated in our previous study.[3] Indeed, the neural networks trained with this algorithm did not fall into local minima, as was the case when the training was performed using the first-order algorithm.[6] In addition, the Levenberg−Marquardt algorithm does not have adjustable parameters, such as learning rate or momentum, etc., and its results can be easily reproduced.

The number of hidden neurons was selected to be five for all studies. The neural networks trained with this number of hidden neurons calculated results for the PHYSPROP set that were similar to those from the lipophilicity program used in the ALOGPS 2.0 package.[6] However, in the previous study twice the number of hidden neurons were used for each network in the ensemble. Unless indicated otherwise 64 neural networks were used in each ensemble.

Molecular Parameters. The input indices for analysis of the PHYSPROP dataset and its subsets included the number of hydrogen and non-hydrogen atoms and 73 E-state indices[19-21] described in our previous study.[6] Some E-state indices had zero occurrences in the star and XLOGP sets; therefore, 70 and 61 input parameters were used for neural networks training, respectively.

ANALYZED PARAMETERS OF ASNN

This section explains the ASNN options, such as similarity measures, neighboring functions, number of neural networks in ensemble, and strategies to include new data in the program.

Similarity Measures. In our previous studies,[1,3] the Pearson's linear correlation coefficient

$$r(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (\mathrm{x}_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \tag{6}$$

was used as a measure of similarity ξ between data cases in the output space. However, r has rather poor statistic for deciding where an observed correlation is statistically significant. A use of this coefficient requires some assumptions on the distributions of x and y, i.e., that both variables have normal distribution.[17] However, in many cases, the distribution of the individual neural network predictions may not approximate a Gaussian distribution or be symmetrical and may be multimodal.[22]

The Spearman rank-order correlation coefficient (r_s) represents a nonparametric correlation coefficient[17] calculated using ranks of $z^i{}_k$ values. It is computed in a very similar way to the Pearson's linear correlation coefficient. For example, vectors of neural network responses $x_i = \{0.1, 0.2, 0.3, 0.05, 0.12\}$ and $y_i = \{0.25, 0.1, 0.3, 0.08, 0.2\}$ are replaced by their ranks $x'_i = \{2, 4, 5, 1, 3\}$, $y'_i = \{4, 2, 5, 1, 3\}$, and the same eq 6 is used to calculate this nonparametric correlation coefficient. The use of ranks instead of values provides a more resistant way to detect correlation. If a correlation is demonstrated to be present nonparametrically, its confidence levels can be easily estimated without any special assumptions on the distribution of analyzed variables.[17] Both Pearson's and Spearman nonparametric coefficients were used in the current study as a measure of similarity ξ_{ij}. Only positive correlations were considered for analysis. All negative correlations were considered to be equal to 0, i.e., no correlation.

One more investigated measure of similarity was based on the Euclidian distance

$$E(x, y) = \sqrt{\sum_i (x_i - y_i)2}$$

The meaning of Euclidian distance, $E(\mathbf{z}^i,\mathbf{z}^j)$, is opposite to the correlation coefficient, i.e., two identical vectors, $\mathbf{z}^i = \mathbf{z}^j$, have Pearson's correlation coefficient $R(\mathbf{z}^i,\mathbf{z}^j) = 1$ and Euclidian distance $E(\mathbf{z}^i,\mathbf{z}^j) = 0$. To apply the same approach for correlation coefficients and Euclidian distance, an inversed Euclidian distance introduced as

$$IE(\mathbf{z}^i, \mathbf{z}^j) = 1/(1 + E(\mathbf{z}^i, \mathbf{z}^j)) \qquad (7)$$

was used. Notice, that the inverse Euclidian distance is always positive.

Neighbor Functions. The KNN method used in our previous studies[1-3] was represented by function

$$F(\xi_{ij}) = 1 \qquad (8)$$

that provided a simple average over all k-nearest neighbors in eq 5. Thus, all nearest neighbors equally contributed to the error correction. Two distance-weighting functions

$$F(\xi_{ij}) = \xi_{ij}{}^2 \qquad (9)$$

and

$$F(\xi_{ij}) = \exp(-\sigma^2/\xi_{ij}{}^2) \qquad (10)$$

were also analyzed in the current study. Both functions weigh the contribution of each neighbor proportionally to its similarity ξ_{ij} to the analyzed case.[23] Thus, the nearest neighbor that is the most similar to the considered case (i.e., the data case with ξ_{ij} near to one) provides the largest contribution to the error correction. The function in eq 10 corresponds to the Parzen-window regression.[24] The large value of the adjustable parameter σ corresponds to a window concentrated near similarity equal to 1. The summation using eq 10 was restricted to the k-nearest neighbors, while it is done over all data cases in the original Parzen-window regression. This restriction used to have faster calculations. The number of nearest neighbors, k, and parameter σ for the Parzen-window regression represent smoothing parameters of the ASNN. These parameters were selected to minimize the ASNN error for the training set.

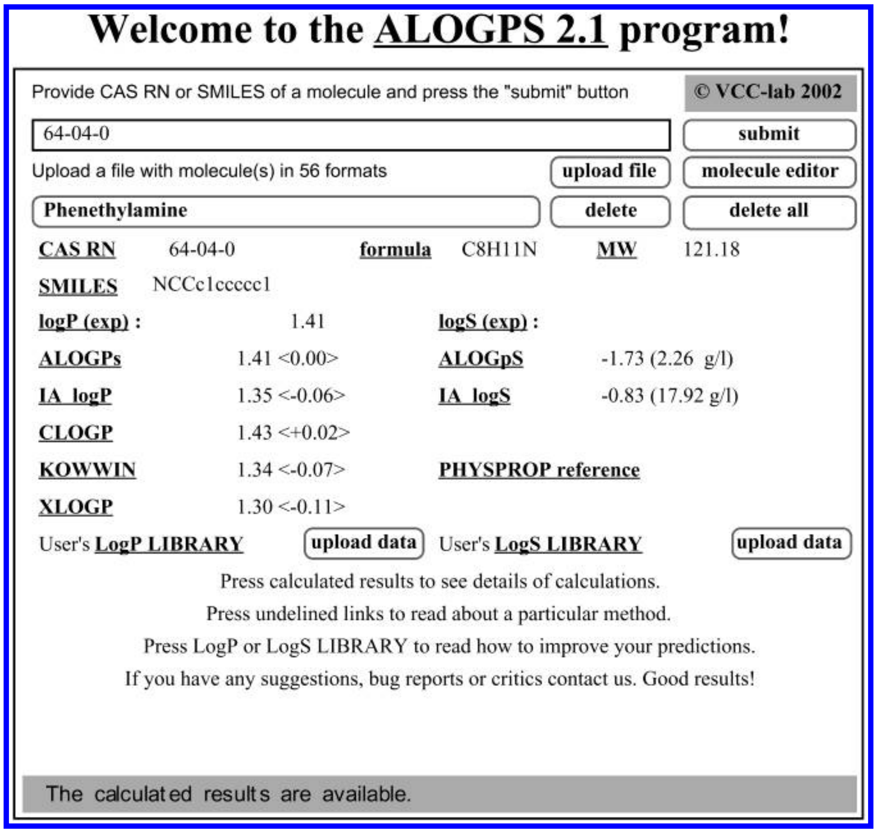
Number of Neural Networks in the Ensemble. The number of neural networks in the ensemble is important to decrease variability of ensemble prediction and to provide a more reliable estimation of the similarity values. However, a large number of networks in the ensemble slows down calculations and requires a larger number of bits to store network weights and calculated results. The number studied ranged from 4 to 256 networks.

Evaluation of Strategies To Extend OPS of Neural Networks. The predictive ability of ASNNs for the training sets was estimated using the leave-one-out (LOO) method. There were several possibilities to estimate the performance of ASNNs for the test sets. Such sets could be considered as the user's private data. In the first analyses the ASNNs just predicted the test sets, i.e., provided the blind test set prediction. These analyses may provide a poor prediction from the ASNN due to the diversity of the training and test sets, i.e., if molecules from the test set are outside of the OPS of neural networks.[6] Of course, if the user does not have any new experimental data, the blind prediction approach is the most appropriate way to test the neural network performance.

However, in many cases the user does have some reliable private data that are pertinent to his series of compounds. The user can include such data to the memory of ASNN (see eq 5) and can dramatically improve prediction of this method by extension of its OPS even with one molecule.[1-3]

There are several approaches to do such an extension. The main questions are how to select the nearest neighbor parameters (i.e. to use the parameters selected for the training set or to select new values according to the user's set) and how to use the available data (i.e., to use both training and user's data or to use only the user's data). The amount of data in the training set is usually considerably larger than the user's set. A selection of the ASNN parameters according to the joint set (i.e., combination of training and the user's sets) requires a long time, and it is biased in favor of the larger training set. Therefore our analyses were restricted to cases where smoothing parameters were selected either according to the training or to the user's set.

Another important question is how to test performance of different strategies to extend OPS. The possible solutions are to use the LOO method for the user's data or to use blind prediction of some parts of the user's data. Since each of these two approaches has different advantages and disadvantages, (e.g., a more complete use of the data in the LOO method compared to a more robust and appealing estimation of the predictive ability in the blind prediction), both strategies were used. In the blind prediction, 50% of test set molecules were left apart from the analysis and were predicted (blind prediction) using the selection of smoothing parameters and extension of the ASNN from the first 50% of molecules. The analysis was repeated using the second half of molecules for the blind prediction, and the calculated results were reported for the whole test set. This procedure is referred to blind_50 in the text of article. In the LOO analysis all the user's molecules were added to the memory of the ASNN, and the statistical parameters that provided a minimum LOO error in eq 5 were considered.



Welcome to the ALOGPS 2.1 program!

Provide CAS RN or SMILES of a molecule and press the "submit" button

© VCC-lab 2002

64-04-0

Upload a file with molecule(s) in 56 formats

Phenethylamine

CAS RN 64-04-0 formula C8H11N MW 121.18

SMILES NCCc1ccccc1

logP (exp) : 1.41 logS (exp) :

ALOGPs 1.41 <0.00> ALOGpS -1.73 (2.26 g/l)

IA_logP 1.35 <-0.06> IA_logS -0.83 (17.92 g/l)

CLOGP 1.43 <+0.02>

KOWWIN 1.34 <-0.07> PHYSPROP reference

XLOGP 1.30 <-0.11>

User's LogP LIBRARY User's LogS LIBRARY

Press calculated results to see details of calculations.

Press undelined links to read about a particular method.

Press LogP or LogS LIBRARY to read how to improve your predictions.

If you have any suggestions, bug reports or critics contact us. Good results!

The calculated results are available.

Figure 1. Front page of ALOGPS 2.1 available at the Virtual Computational Chemistry Laboratory site at http://www.vcclab.org/lab/alogps. The user can create its own library by using "upload data" buttons. A click on the calculated values open windows with detailed information about the used method. A comparison of different methods is also available for analysis of single compounds.

Table 1. Training and Test Sets of Neural Networks

name	no. of network params	training set	test set
ASNN1	75	12908 molecules, PHYSPROP set	
ASNN2	75	6454 molecules from the PHYSPROP set	6454 molecules from the PHYSPROP set
ASNN3	70	9429, star set	3479, nova set
ASNN4	61	1853, XLOGP training set	11055 molecules from PHYSPROP set

Thus, three protocols, namely blind, blind_50 and LOO results, were used to estimate the predictive ability of the ASNN for the user's set. The prediction ability of ANNE was also estimated using LOO results calculated for the training set.

Implementation. The program was developed in C/C++ language, and it is available online at http://www.vcclab.org/lab/alogps using Java interface (Figure 1). A standalone version of the program is available on request from the authors. If a user has their own experimental values, such data can be provided and automatically used to improve prediction ability of ALOGPS 2.1 program for the other user's molecules. The program calculates about 3000 molecules per minute using AMD 1 GHz computer.

RESULTS

Analysis of the Similarity Measures. Two protocols, namely blind and LOO results, were used to detect which similarity measure, namely Pearson's linear correlation coefficient, r, Spearman rank order coefficient, r_s, or inverse Euclidian distance provided the best performance for the neural networks. Four types of neural networks, ASNN1−ASNN4, were developed using different training sets as indicated in Table 1.

The use of an ASNN provided an improvement of results compared to the ANNE for all analyzed similarity measures (Table 2). The best performances of these methods were calculated with similarity measures according to Pearson's and Spearman rank correlation coefficients. The use of the inverse Euclidian distance provided the poorest results for the ASNNs, with the exception of the blind test set prediction from ASNN4. Both linear and nonparametric correlation coefficients provided similar results and could be selected for further program development. Although, the main advantages of Spearman rank correlation coefficient over Pearson's were the lower number of bits required to store the neural network responses and the faster speed of calculations. This is due to the fact that only ranks of neural networks are required to calculate Spearman rank correlation coefficient. For example, for an ensemble of $64 = 2^6$ networks only 6 bits are required to store the ranking of one network. On the contrary, float values are required to store neural network responses using correlation coefficient as a measure of distance. The float values occupy 32 or 64 bits, which is 5−10 times larger compared to the use of the rank correlations. The Spearman rank correlation coefficient is calculated using only integer arithmetic, while floating points operations are required to compute the Pearson's linear correlation coefficient. Since floating point arithmetic is

Table 2. Performance of ASNN for Different Similarity Measures

	training set LOO results			blind test set prediction[a]		
network, *k*	RMSE	MAE	outliers[b]	RMSE	MAE	outliers
			ANNE Analysis			
ANNE1	0.48 (0.44)[b]	0.36 (0.34)	129			
ANNE2	0.51 (0.46)	0.38 (0.36)	91	0.57 (0.48)	0.41 (0.38)	133
ANNE3	0.46 (0.42)	0.34 (0.33)	78	1.62 (0.59)	0.92 (0.47)	480
ANNE4	0.34 (0.33)	0.24 (0.24)	3	1.18 (0.59)	0.71 (0.47)	1058
			Pearson's Correlation Coefficient, *r*			
ASNN1 4	0.42 (0.38)	0.29 (0.28)	79			
ASNN2 7	0.44 (0.41)	0.32 (0.31)	44	0.51 (0.44)	0.36 (0.33)	91
ASNN3 5	0.38 (0.37)	0.28 (0.27)	41	1.70 (0.57)	0.93 (0.45)	461
ASNN4 35	0.30 (0.29)	0.22 (0.21)	3	1.17 (0.57)	0.69 (0.45)	1008
			Spearman Rank Correlation Coefficient, r_s			
ASNN1 5	0.42 (0.38)	0.29 (0.28)	86			
ASNN2 7	0.44 (0.41)	0.32 (0.31)	45	0.51 (0.44)	0.36 (0.34)	86
ASNN3 5	0.38 (0.36)	0.28 (0.27)	37	1.73 (0.57)	0.94 (0.45)	468
ASNN4 57	0.30 (0.29)	0.21 (0.21)	3	1.18 (0.58)	0.69 (0.46)	1024
			Inverse Euclidian Distance			
ASNN1 6	0.46 (0.43)	0.34 (0.33)	110			
ASNN2 7	0.45 (0.42)	0.33 (0.32)	41	0.51 (0.45)	0.38 (0.35)	92
ASNN3 8	0.42 (0.40)	0.32 (0.31)	48	1.75 (0.59)	0.96 (0.47)	465
ASNN4 35	0.32 (0.31)	0.22 (0.21)	3	1.17 (0.57)	0.68 (0.45)	953

[a] The test sets are described in Table 1. [b] The molecules with prediction error above ±1.5 log units were considered as outliers. Results calculated without outliers are indicated in parentheses. ANNE is an abbreviation of Artificial Neural Network Ensemble that uses a simple average (eq 3) to predict analyzed data case. The nonweighted average given by eq 8 was used to calculate ASNN results.

slower, the calculation of the Spearman rank correlation coefficient is considerably faster. The speed of calculation and required data storage are important options for programs that are going to be used for prediction of a large number of compounds. Therefore, the Spearman rank correlation coefficient was selected for further analyses reported in this study.

The number of the detected nearest neighbors was much larger for the neural networks trained with XLOGP data set, i.e., the average over a larger number of molecules was required to provide optimal performance of the ASNN4 network. This result indicated that data in the XLOGP set were more homogeneous, e.g. the molecules were more similar amid themselves, compared to the others sets. Indeed, the molecules in the XLOGP set had a significantly smaller number of non-hydrogen atoms (12.96 non-hydrogen atoms/molecule) compared to the PHYSPROP data set (16.63 non-hydrogen atoms/molecule).[6] Thus, this data set contained simpler molecules compared to the other investigated data sets.

Analysis of Averaging Functions. The second analysis was performed to compare the performance of ASNNs for different averaging schemes, namely nonweighted average and two weighted averages given by eqs 8−10.

The nonweighted average provided the poorest predictive ability and the highest number of outliers according to the LOO results for the training set and blind prediction of the test sets. The best performance was calculated using the Parzen-window regression. The number of nearest neighbors for the Parzen-window regression was almost an order of magnitude larger compared to the other averaging methods, except for ASNN4 that was developed using the XLOGP set, see Table 3. Notice that in addition to the number of nearest neighbors the Parzen-window regression is also controlled by a parameter σ. This parameter (σ) was calculated to be 0.49 for ASNN4 and around 1 for the other ASNNs. Thus, even if the number of nearest neighbors in the Parzen-window regression was of the same order for all analyzed neural networks, the averaging window for ASNN4 was much wider compared to neural networks developed with other training sets. Since the Parzen-window regression provided the best results, it was selected for further analyses reported in this article.

Analysis of the Number of Networks in the Ensemble. ASNN prediction increased with the number of neural networks in the ensemble up to 32−64 networks (Table 4). After this number the performance of all analyzed networks stabilized and did not change more than a few percent even with 256 neural networks per ensemble. On the contrary, the smoothing parameters changed with the increase of the number of neural networks per ensemble. The general tendency was a decrease of the number of nearest neighbors, *k*, and the value of parameter σ. This result suggested that neural network ensembles with larger number of networks were able to more precisely identify the nearest neighbors and thus a smaller number of them was required to get improved prediction ability of ASSN. On the other hand, the decrease of the parameter σ suggested that the selected neighbors provided a similar contribution to the error correction. Since the performance of neural networks was approximately the same for more than 64 neural networks, this number was selected for the analysis of the strategies to extend OPS of neural networks.

Analysis of the Strategies To Extend OPS of ASNNs. For this type of analysis, the set of 18 quinazolones was used as the test set of ASNN1. The first three sets of analysis (ANALYSIS 1−3, Table 5) were done using parameters *k* and σ selected according to the training sets. The performance of all neural networks was significantly improved when molecules from the test sets were added to the memory of ASNN (ca. ANALYSIS 2, 3, and 1, Table 5). Thus, an extension of the OPS provided an important improvement for the prediction of ASNNs. Both training and test set

Table 3. Performance of ASNN for Different Nearest Neighbor Functions

	training set LOO results			blind test set prediction[a]		
network, $k(\sigma)$	RMSE	MAE	outliers[b]	RMSE	MAE	outliers
		Nonweighted Average (Eq 8)				
ASNN1 5	0.42 (0.38)[b]	0.29 (0.28)	86			
ASNN2 7	0.44 (0.41)	0.32 (0.31)	45	0.51 (0.44)	0.36 (0.34)	86
ASNN3 5	0.38 (0.36)	0.28 (0.27)	37	1.73 (0.57)	0.94 (0.45)	468
ASNN4 57	0.30 (0.29)	0.21 (0.21)	3	1.18 (0.58)	0.69 (0.46)	1024
		Weighted Average (Eq 9)				
ASNN1 6	0.40 (0.37)	0.28 (0.27)	80			
ASNN2 10	0.42 (0.40)	0.31 (0.30)	40	0.49 (0.43)	0.35 (0.32)	86
ASNN3 8	0.37 (0.35)	0.26 (0.26)	36	1.76 (0.56)	0.94 (0.44)	460
ASNN4 56	0.29 (0.29)	0.20 (0.20)	3	1.17 (0.57)	0.68 (0.45)	999
		Parzen-Window Regression (Eq 10)				
ASNN1 96 (0.99)	0.39 (0.35)	0.27 (0.26)	76			
ASNN2 46 (0.84)	0.42 (0.39)	0.30 (0.29)	40	0.49 (0.42)	0.34 (0.31)	82
ASNN3 186 (0.95)	0.36 (0.34)	0.25 (0.25)	32	1.74 (0.56)	0.93 (0.44)	462
ASNN4 62 (0.49)	0.29 (0.28)	0.20 (0.20)	4	1.17 (0.57)	0.68 (0.45)	978

[a] The test sets are described in Table 1. [b] The molecules with prediction error above ±1.5 log units were considered as outliers. Results calculated without outliers are indicated in parentheses.

Table 4. Prediction Ability of ASNN as a Function of the Number of Neural Networks in the Ensemble

		training set LOO results			blind test set prediction[a]		
network	k (σ)	RMSE	MAE	outliers[b]	RMSE	MAE	outliers
			ASNN1				
4	0 (1.00)	0.52 (0.46)[b]	0.38 (0.36)	182			
8	199 (2.72)	0.50 (0.45)	0.37 (0.35)	139			
16	197 (9.19)	0.47 (0.43)	0.34 (0.32)	135			
32	198 (1.50)	0.41 (0.37)	0.29 (0.28)	90			
64	96 (0.99)	0.39 (0.35)	0.27 (0.26)	76			
128	19 (0.75)	0.39 (0.35)	0.27 (0.26)	77			
256	19 (0.66)	0.39 (0.35)	0.27 (0.26)	76			
			ASNN2				
4	153 (0.10)	0.57 (0.49)	0.42 (0.38)	136	0.62 (0.51)	0.45 (0.40)	168
8	133 (1.45)	0.52 (0.47)	0.38 (0.36)	88	0.59 (0.50)	0.43 (0.39)	129
16	144 (1.91)	0.47 (0.43)	0.34 (0.33)	55	0.55 (0.47)	0.39 (0.36)	107
32	89 (1.25)	0.44 (0.40)	0.31 (0.30)	46	0.51 (0.44)	0.35 (0.33)	89
64	46 (0.84)	0.42 (0.39)	0.30 (0.29)	40	0.49 (0.42)	0.34 (0.31)	82
128	35 (0.58)	0.41 (0.38)	0.29 (0.28)	35	0.48 (0.41)	0.33 (0.31)	83
256	17 (0.51)	0.41 (0.38)	0.29 (0.28)	37	0.48 (0.41)	0.33 (0.31)	86
			ASNN3				
4	107 (0.10)	0.47 (0.43)	0.35 (0.33)	85	1.63 (0.59)	0.95 (0.47)	529
8	178 (2.40)	0.44 (0.42)	0.33 (0.32)	55	2.09 (0.58)	1.09 (0.47)	519
16	194 (2.04)	0.41 (0.38)	0.30 (0.29)	43	1.94 (0.59)	1.03 (0.47)	483
32	142 (1.42)	0.37 (0.35)	0.27 (0.26)	37	2.03 (0.57)	1.04 (0.45)	464
64	186 (0.95)	0.36 (0.34)	0.25 (0.25)	32	1.74 (0.56)	0.93 (0.44)	462
128	41 (0.67)	0.35 (0.33)	0.25 (0.24)	31	1.69 (0.55)	0.91 (0.43)	460
256	18 (0.46)	0.35 (0.33)	0.25 (0.24)	29	1.63 (0.55)	0.90 (0.44)	461
			ASNN4				
4	199 (0.10)	0.38 (0.35)	0.27 (0.26)	11	1.24 (0.61)	0.77 (0.49)	1320
8	81 (1.31)	0.35 (0.34)	0.25 (0.25)	5	1.20 (0.60)	0.72 (0.47)	1123
16	175 (1.23)	0.33 (0.31)	0.23 (0.23)	4	1.18 (0.59)	0.70 (0.47)	1013
32	145 (0.85)	0.30 (0.29)	0.21 (0.21)	4	1.17 (0.58)	0.68 (0.45)	982
64	62 (0.49)	0.29 (0.28)	0.20 (0.20)	4	1.17 (0.57)	0.68 (0.45)	978
128	60 (0.50)	0.29 (0.28)	0.20 (0.20)	3	1.17 (0.58)	0.68 (0.45)	1006
256	59 (0.10)	0.28 (0.27)	0.20 (0.19)	3	1.17 (0.58)	0.68 (0.45)	1014

[a] The test sets are described in Table 1. [b] The molecules with prediction error above ±1.5 log units were considered as outliers. Results calculated without outliers are indicated in parentheses.

molecules were required for optimal performance of the neural networks. Indeed, if only the test set molecules were added to the memory of ASNN, the neural network performance was poorer compared to analyses where both training and test sets were used as the ASNNs memory (ca. ANALYSIS 3 and 2, Table 5). As was expected, the LOO results were better than those obtained from the blind_50 results.

The next two analyses (ANALYSIS 4 and 5, Table 5) were done with smoothing parameters selected using the test sets. This selection provided a significant improvement of ASNN prediction ability compared to the previous results (ca. ANALYSIS 4 and 5 and ANALYSIS 2 and 3, Table 5). The most dramatic change was in the number of outliers that decreased by about two times for both ASNN4 and ASNN3 networks. Thus the nearest neighbor parameters selected

Table 5. Prediction Performance of ASNN for the Test Sets Using Different Strategies To Extend Optional Prediction Space of Neural Networks

	test set LOO results[a]			test set blind_50 results[a]		
network, $k(\sigma)$	RMSE	MAE	outliers[b]	RMSE	MAE	outliers
ANALYSIS 1: ASNN Memory = Training Set[c]						
ASNN1 96 (0.99)	1.05 (0.66)[b]	0.84 (0.55)	4			
ASNN2 46 (0.84)	0.49 (0.42)	0.34 (0.31)	82			
ASNN3 186 (0.95)	1.74 (0.56)	0.93 (0.44)	462			
ASNN4 62 (0.49)	1.17 (0.57)	0.68 (0.45)	978			
ANALYSIS 2: ASNN Memory = Training + Test Sets[c]						
ASNN1 96 (0.99)	0.32	0.23	0	0.33	0.25	0
ASNN2 46 (0.84)	0.43 (0.39)	0.30 (0.29)	54	0.45 (0.40)	0.31 (0.29)	53
ASNN3 186 (0.95)	0.70 (0.50)	0.45 (0.38)	136	0.73 (0.52)	0.48 (0.40)	147
ASNN4 62 (0.49)	0.79 (0.53)	0.52 (0.41)	587	0.87 (0.55)	0.56 (0.43)	702
ANALYSIS 3: ASNN Memory = Test Set[c]						
ASNN1 96 (0.99)	0.32	0.24	0	0.42	0.31	0
ASNN2 46 (0.84)	0.44 (0.40)	0.31 (0.30)	56	0.46 (0.42)	0.33 (0.32)	62
ASNN3 186 (0.95)	0.74 (0.51)	0.48 (0.38)	179	0.80 (0.54)	0.52 (0.41)	210
ASNN4 62 (0.49)	0.80 (0.53)	0.52 (0.42)	598	0.86 (0.55)	0.57 (0.44)	717
ANALYSIS 4: ASNN Memory = Test Set[d]						
ASNN1 4 (2.18)	0.21	0.18	0	0.33	0.25	0
ASNN2 38 (1.25)	0.44 (0.40)	0.31 (0.29)	57	0.47 (0.42)	0.33 (0.32)	60
ASNN3 13 (7.13)	0.62 (0.49)	0.42 (0.37)	98	0.73 (0.52)	0.49 (0.40)	159
ASNN4 6 (1.39)	0.58 (0.42)	0.36 (0.31)	228	0.67 (0.46)	0.41 (0.35)	336
ANALYSIS 5: ASNN Memory = Training + Test Sets[d]						
ASNN1 4 (2.18)	0.21	0.18	0	0.27	0.20	0
ASNN2 38 (1.25)	0.43 (0.38)	0.29 (0.28)	51	0.45 (0.40)	0.32 (0.30)	57
ASNN3 13 (7.13)	0.63 (0.49)	0.41 (0.36)	94	0.66 (0.52)	0.45 (0.39)	106
ASNN4 6 (1.39)	0.57 (0.42)	0.35 (0.31)	220	0.65 (0.46)	0.40 (0.34)	322
ANALYSIS 6: ASNN Memory = Training Set[d]						
ASNN1 4 (2.18)	1.21 (0.76)	0.99 (0.65)	5			
ASNN2 38 (1.25)	0.46 (0.41)	0.32 (0.30)	75			
ASNN3 13 (7.13)	2.08 (0.58)	1.05 (0.45)	467			
ASNN4 6 (1.39)	1.19 (0.58)	0.69 (0.46)	994			

[a] The set of 18 quinazolones[12,13] was used as the test set for ASNN1, and the other test sets are described in Table 1. [b] The molecules with prediction error above ±1.5 log units were considered as outliers. Results calculated without outliers are indicated in parentheses. [c] Parameters were selected using training sets. [d] Parameters were selected using test sets.

according to the training sets were not optimal for the prediction of the test sets. The additional training of the nearest neighbor predictor made it possible to select optimal parameters and considerably improved the prediction of the ASNN. The number of nearest neighbors for networks ASNN1, ASNN3, and ASNN4 decreased by about 10 times compared to the number selected according to the training set. On the contrary, the parameter σ of these networks increased by over two times. Thus, Parzen window parameters selected for the test set corresponded to a narrower window compared to the parameters selected for the training set.

The smallest increase of ASNN prediction ability was observed for ASNN2, i.e., the network developed using 6454 molecules selected by chance from the PHYSPROP dataset. The diversity of compounds in both sets was very similar, and thus no significant additional gain could be obtained by tuning the Parzen-window regression parameters for this test set.

The interesting question was, did ASNN prediction abilities improve due to new molecules being added or did prediction abilities improve due to the changing of the Parzen-window regression parameters. This question was assessed using smoothing parameters that were selected with the test sets, but only molecules from the training sets were utilized as the ASNN memory (ANALYSIS 6, Table 5). The calculated results were the poorest of the all analyses; i.e., the improvement of the ASNNs was due to fresh data and not due to the change of the Parzen-window regression parameters.

The smoothing parameters selected using test sets were not optimal for the prediction of molecules from the training set too. For example, root-mean-squared error, RMSE = 0.43 (0.39), absolute mean error, MAE = 0.30 (0.27) and 144 outliers were calculated for ASNN1 network (LOO results) using parameters selected with the test set of quinazolones. These results were lower compared to the ASNN1 LOO results, RMSE = 0.39 (0.35), MAE = 0.27 (0.26) and 76 outliers, calculated using parameters optimized for the training set itself (Table 4). These findings suggest a heuristic strategy for the user's data analysis. At first, new smoothing parameters (k_{user}, σ_{user}) are selected according to the user's data. For each test molecule a number of nearest neighbors, k_{user}, is calculated. If at least one of the detected molecules is from the user's data, the prediction of activity of the molecule is done using the smoothing parameters selected according to the user data. However, if all nearest neighbors, k_{user}, are from the training set, the smoothing parameters selected according to this set are used.

Prediction performance for three out of four neural networks represent a smooth function of the Parzen's parameter σ that reaches a single minimum for the optimal value of this parameter (Figure 2). On the contrary, the same function for ASNN3 has additional local minimum, RMSE

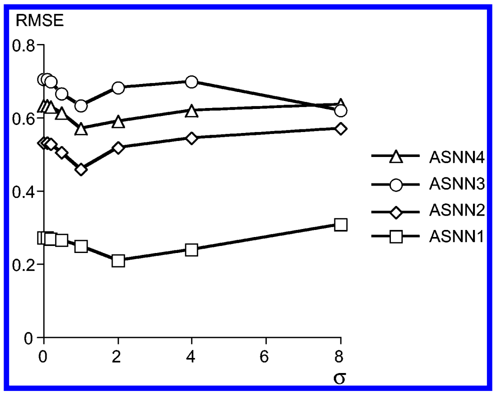


Figure 2. Test set root mean squared error as a function of parameter σ in Parzen-window regression. The number of nearest neighbors, *k*, for each network is indicated in Table 5 (analysis 4).

= 0.63, for $\sigma \approx 1$ and global minimum, RMSE = 0.62, for $\sigma = 7.13$. The presence of two minima could indicate that the nova set contains two different subsets requiring different smoothing parameters. The local minima near $\sigma \approx 1$ may correspond to a subset of compounds that require a smooth average over nearest neighbors using a wide Parzen window, while the second minimum, $\sigma = 7.13$, corresponds to the compounds that require narrow Parzen window. A use of the appropriate smoothing parameters for each such set would probably increase prediction performance of our method. Unfortunately, at the present moment it is not clear how such subsets could be separated.

The models developed using XLOGP, star, and PHYSPROP data sets could be considered as consecutive steps of the lipophilicity prediction program development using an increased amount of experimental data. The interesting question was to estimate how the increase in the amount of data influenced prediction ability of such programs. This study was performed using the set of 18 quinazolones[12,13] as the test set for all networks. Nine of the 18 analyzed quinazolones had general structure designated as (I), and nine had the general structure (II) as shown in Table 6. An increase in the number of molecules in the training sets provided some improvement for the calculated results of these molecules, shown in Table 6. However, even when the number of training samples in ASNN1 is about an order larger than in ASNN4, only a marginal improvement in the results is observed. Indeed, even for the best neural network, ASNN1, MAE was about 0.8 log units, and four outliers were calculated. On the contrary, the extension of the OPS of neural networks by including the new molecules to the ASNN memory significantly improved the prediction ability of the networks. Thus the impact of the user's additional data was more important for the enhancement of the ASNN predictions for this series of compounds compared to increasing the training set size.

DISCUSSION

This study provides the first systematic analyses of different parameters used in an ASNN. The main result of this study concerns strategies to apply to the ASNN for user's data. The calculated results suggest that a significant improvement of this method, up to several hundreds percent, can be obtained if Parzen-window parameters are selected according to the user's data. The trained neural network does provide a good approximation of the learning set, although only a minor improvement (about 5−20%) can be calculated for this set using the best nearest neighbor parameters detected from the learning set.

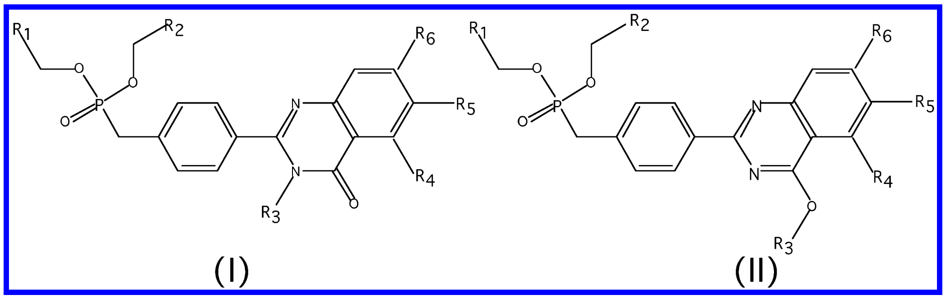
The ASNN predictions for the user's additional data, that may cover some regions of chemical space that were not available during the ASNN development, can be low if the training and test regions of chemical space differ too greatly. However selection of the nearest neighbor parameters for the user's data provides "on-the-fly" learning for the ASNN and dramatically improves its performance as it was demonstrated for LOO and blind_50 test protocols.

On the other hand, if the tested molecule does not have analogues in the user's set, there is no reason to use the smoothing parameters selected according to the user's data (ca. ANALYSIS 1 and 6, Table 5). Indeed, the smoothing parameters selected with the training set are optimal if no additional information about the user's molecules is used (ca. ANALYSIS 1 and 6, Table 5). The proposed approach provides a simple way to tune smoothing parameters of the algorithm depending on the analyzed molecule. This approach is implemented in the ALOGPS 2.1 program that is based on the ASNN1 neural network ensemble analyzed in this study. Selection of the smoothing parameters depending on the amount and quality of available data would probably further improve prediction performance of the ASNN, especially its user-training feature. For example, a use of such a method could provide a selection of smoothing parameters appropriate for each of two subsets of nova set compounds (Figure 2) for the ASNN3 neural network. The development of such an approach requires further theoretical study of the method.

The performance of a previous version of the ALOGPS 2.0 program ($n = 12908$, MAE = 0.29, RMSE = 0.39 with 131 outliers) favorable compares with other lipophilicity prediction programs, such as CLOGP,[25] KOWWIN,[26] and others that were developed using large databases of compounds.[6] A new version, ALOGPS 2.1, calculates even better results ($n = 12908$, MAE = 0.26, RMSE = 0.35 with 76 outliers). A recent comparison of several different approaches for theoretical physicochemical properties predictions (including our program) can be found elsewhere.[27] Of course, to a large degree a success of the ALOGPS is due to the use of the E-state indices[19,21] that are becoming very popular and an established parameter system in the quantitative structure−activity relationship studies.

ASNN3 developed using a star set of compounds provided a higher prediction error for the nova set ($n = 3479$ RMSE = 1.74) compared to the results of the ASNN4 for its test set ($n = 11055$ RMSE = 1.17). This result looks surprising since our previous analysis of quinazolones (Table 6) suggests that larger training sets generate models with higher prediction ability. The seeming conflict can be easily resolved using a more detailed analysis of the ASNN4 results. Actually the performance of the ASNN4 for prediction of molecules from the nova set, RMSE = 1.9, is even lower than that of the ASNN3 neural network. At the same time the performance of the ASNN4 for molecules from the star set is considerably better, RMSE = 0.79. Thus an apparently higher prediction performance of the ASNN4 compared to the ASNN3 is mainly due to a different composition of the both test sets.

Table 6. Prediction of the 18 Quinazolones Using Neural Networks Developed with Different Training Sets[b]



		blind prediction			LOO results			blind_50 prediction		
	training set, *n*	RMSE	MAE	outliers[a]	RMSE	MAE	outliers	RMSE	MAE	outliers
ASNN4	1853	1.74 (0.96)[a]	1.44 (0.83)	7	0.37	0.29	0	0.75 (0.61)	0.53 (0.45)	1
ASNN2	6454	1.18 (0.75)	0.96 (0.64)	4	0.29	0.24	0	0.72	0.56	0
ASNN3	9429	1.23 (0.72)	0.98 (0.60)	5	0.21	0.16	0	0.65	0.44	0
ASNN1	12908	1.05 (0.66)	0.84 (0.55)	4	0.21	0.18	0	0.27	0.20	0

[a] The molecules with prediction error above ±1.5 log units were considered as outliers. Results calculated without outliers are indicated in parentheses. [b] Notice, that almost an order increase in the number of molecules in the training set (1853 ⟹ 12 908) decreased RMSE for the test set molecules in less than two times, as it is demonstrated by blind prediction results. On the contrary, a use of molecules in the LIBRARY mode decreased the same error for the remaining molecules in 2−5 times, as it is demonstrated by LOO and blind_50 results.

The analysis of Table 6 suggested that a considerable improvement of the prediction performance of the ASNN method is calculated if the private user's molecules are used in the memory of neural networks. Indeed, the prediction errors for unseen molecules were decreased several times if even only nine molecules were added to the memory of the ASNN. Of course, this result does not mean that for each new series of compounds the user should experimentally measure 50% of molecules and use them to predict activity of the remaining substances. On the contrary, it clearly suggests that only a few molecules, e.g. 4−5 compounds per series, are sufficient to dramatically improve prediction ability of this method. The prediction ability of ASNN can be improved using an even smaller number of molecules. For example, if only two compounds with substituents R_1═R_2=Et, R_3=Me, R_4═R_5═R_6=H, representing the simplest versions of (I) and (II), are used as a memory of the ASNN1, the calculated results for the remaining compounds, RMSE = 0.45, MAE = 0.38, are also significantly improved compared to the blind prediction results (Table 6). Thus the use of just a single compound from each series increases prediction ability of the ALOGPS in about two times. A similar result calculated using a previous version of our program as well as a practical application of the ALOGPS to predict lipophilicity of 6100 in-house BASF compounds are analyzed in details elsewhere.[2,3]

In the current study both Pearson's linear correlation and Spearman rank correlation coefficients provided quite similar results. There were several aspects in favor of nonparametric rank correlation, i.e., faster speed of computation and lower number of bits required to store neural network results, mentioned previously in the article. Another important reason to prefer nonparametric over parametric correlation coefficient are provided by the recent discoveries in the neurophysiology.

The theories of brain coding suggest an importance of temporal coding for information processing in brain.[28−33] Analysis of speed of processing in the human visual system and particular features of this system (e.g., ability to recognize the same signals at different brightness of images) suggests an importance of the rank coding for information processing.[34] To this extent, each network could be considered as one spiking neuron, and the ASNN itself could be considered as an ensemble of spiking neurons. The use of the Spearman rank correlation coefficient, assuming that the rank of spiking neuron corresponds to its temporal delay (the first firing neuron has the highest rank), makes the ASNN approach comparable to the rank-coding model of Thorpe.[35] Notice, that usually only a few several spikes are required to initiate and to determine the neural network response. This feature of brain circuits provides both speed and accuracy of information processing in the brain. To this extent, the number of such spikes corresponds to the number of nearest neighbors used in the ASNN model.[35] Thus the proposed approach has profound biological background, and it is possible that this method will find many applications in different fields of science and technology.

ACKNOWLEDGMENT

This study was partially supported by Virtual Computational Chemistry Laboratory (INTAS-INFO 00-0363) grant. We thank Ben Tehan (Monash University, Australia) for careful reading of the manuscript and anonymous reviewers for their useful remarks.

REFERENCES AND NOTES

(1) Tetko, I. V. Associative Neural Network, CogPrints Archive, cog00001441, available as http://cogprints.soton.ac.uk/documents/disk0/00/00/14/41/index.html, 2001.

(2) Tetko, I. V. Associative Neural Network. *Neur. Proc. Lett.* **2002**, in press.

(3) Tetko, I. V. Neural Network Studies. 4. Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717−728.

(4) Gombar, V. K. Reliable Assessment of LogP of Compounds of Pharmaceutical Relevance. *SAR QSAR Environ. Res.* **1999**, *10*, 371−380.

(5) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Internet Software for the Calculation of the Lipophilicity and Aqueous Solubility of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 246−252.

(6) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. P. Prediction of *n*-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-state Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407−1421.

(7) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-state Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.

(8) Tetko, I. V.; Villa, A. E. P. Efficient Partition of Learning Data Sets for Neural Network Training. *Neural Networks* **1997**, *10*, 1361−1374.

(9) Syracuse Research Corporation. Physical/Chemical Property Database (PHYSPROP); SRC Environmental Science Center: Syracuse, NY.

(10) BioByte Corp., 201 W. Fourth Street, Claremont, CA.

(11) Wang, R.; Gao, Y.; Lai, L. Calculating Partition Coefficient by Atom-Additive Methodol. *Persp. Drug Discov. Design* **2000**, *19*, 47−66.

(12) Kurogi, Y.; Inoue, Y.; Tsutsumi, K.; Nakamura, S.; Nagao, K.; Yoshitsugu, H.; Tsuda, Y. Synthesis and Hypolipidemic Activities of Novel 2-[4-[diethoxyphosphoryl)methyl]phenyl]quinazolines and 4(3H)-quinazolinones. *J. Med. Chem.* **1996**, *39*, 1433−1437.

(13) ACD/LogP Prediction Software. An Overview and Comparison; ACD, Inc.: 133 Richmond St. W., Suite 605, Toronto ON M5H 2L3, Canada, 2001.

(14) Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural Network Studies. 1. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826−833.

(15) Tetko, I. V.; Luik, A. I.; Poda, G. I. Applications of Neural networks in Structure-Sctivity Relationships of a Small Number of Molecules. *J. Med. Chem.* **1993**, *36*, 811−814.

(16) Bishop, M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, 1995.

(17) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: New York, 1994; p 998.

(18) Shepherd, A. J. *Second-Order Methods for Neural Networks*; Springer-Verlag: London, 1997; p 145.

(19) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.

(20) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types − a Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.

(21) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: London, 1999.

(22) Lawrence, S.; Back, A. D.; Tsoi, A. C.; Giles, C. L. On the Distribution of Performance from Multiple Neural-Network Trials. *IEEE Trans. Neur. Netw.* **1997**, *8*, 1507−1517.

(23) Dudani, S. A. The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Trans. Sys. Man Cyber.* **1976**, *SMC-6*, 325−327.

(24) Härdle, W. *Smoothing Techniques with Implementation in S*; Springer-Verlag: New York, 1990.

(25) Leo, A. J.; Hoekman, D. Calculating log P(oct) with no Missing Fragments; The problem of Estimating New Interaction Parameters. *Persp. Drug Discov. Design* **2000**, *18*, 19−38.

(26) Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol−Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83−92.

(27) Livingstone, D. J. Theoretical Property Predictions. *Curr. Top. Med. Chem.* **2002**, in press.

(28) Abeles, M. *Corticotronics: Neural Circuits of the Cerebral Cortex*; Cambridge University Press: New York, 1991; p 280.

(29) Villa, A. E. P.; Tetko, I. V.; Hyland, B.; Najem, A. Spatiotemporal Activity Patterns of Rat Cortical Neurons Predict Responses in a Conditioned Task. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 1106−1111.

(30) Gerstner, W.; Kempter, R.; van Hemmen, J. L.; Wagner, H. A Neuronal Learning Rule for Sub-Millisecond Temporal Coding. *Nature* **1996**, *383*, 76−81.

(31) Tetko, I. V.; Villa, A. E. P. A Pattern Grouping Algorithm for Analysis of Spatiotemporal Patterns in Neuronal Spike Trains. 1. Detection of Repeated Patterns. *J. Neurosci. Methods* **2001**, *105*, 1−14.

(32) Tetko, I. V.; Villa, A. E. P. A Pattern Grouping Algorithm for Analysis of Spatiotemporal Patterns in Neuronal Spike Trains. 2. Application to Simultaneous Single Unit Recordings. *J. Neurosci. Methods* **2001**, *105*, 15−24.

(33) Hopfield, J. J.; Brody, C. D. What is a moment? Transient Synchrony as a Collective Mechanism for Spatiotemporal Integration. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 1282−1287.

(34) Thorpe, S.; Fize, D.; Marlot, C. Speed of Processing in the Human Visual System. *Nature* **1996**, *381*, 520−522.

(35) Gautrais, J.; Thorpe, S. Rate Coding Versus Temporal Order Coding: A Theoretical Approach. *Biosystems* **1998**, *48*, 57−65.

CI025515J