# Application of a Pruning Algorithm To Optimize Artificial Neural Networks for Pharmaceutical Fingerprinting

Igor V. Tetko*,†,‡

Department of Biomedical Applications, Institute of Bioorganic and Petroleum Chemistry, Murmanskaya, 1, Kiev-660, 253660 Ukraine

Alessandro E. P. Villa

Laboratoire de Neuro-heuristique, Institut de Physiologie, Université de Lausanne, Rue du Bugnon 7, Lausanne, CH-1005, Switzerland

Tatjana I. Aksenova

Institute of Applied System Analysis, Prospekt Peremogy, 37, 252056, Kiev, Ukraine

Walter L. Zielinski and James Brower

Division of Drug Analysis, U.S. Food and Drug Administration, St. Louis, Missouri 63101

Elizabeth R. Collantes§ and William J. Welsh*

Department of Chemistry and Center for Molecular Electronics, University of Missouri-St. Louis, St. Louis, Missouri 63121

Received December 4, 1997

The present study investigates an application of artificial neural networks (ANNs) for use in pharmaceutical fingerprinting. Several pruning algorithms were applied to decrease the dimension of the input parameter data set. A localized fingerprint region was identified within the original input parameter space from which a subset of input parameters was extracted leading to enhanced ANN performance. The present results confirm that ANNs can provide a fast, accurate, and consistent methodology applicable to pharmaceutical fingerprinting.

## INTRODUCTION

The discovery of fraudulent practices at a generic pharmaceutical firm in 1989 led the U.S. Food and Drug Administration (FDA) to conduct large-scale investigations into the pharmaceutical industry. It has been suggested that instances of fraud might be directly detectable from analytical "fingerprints" which could demonstrate sameness or differences between samples.[1−3] Consequently, recent initiatives have been directed toward an evaluation of computer-based methods that could be used to distinguish within- and between-batch product consistency, to examine the effects of process changes in the production of pharmaceutical products, and to determine whether a product marketed today is the same as that which was originally approved.[1]

It is well-established that information on the microscopic chemical composition of products provided by chromatographic trace organic impurity patterns represents an important component of the product fingerprint. A comparison of such patterns can often be used to reliably judge the *sameness* or *difference* between samples and to determine precursor and degradation profiles in the bulk drug.[4] While HPLC is extremely useful in testing bulk pharmaceutical products for impurities, it is not the cure-all for these detection efforts. One drawback is that HPLC trace impurity data are subject to concerns about repeatability and imprecision. Even HPLC columns that are nominally identical can exhibit variations in peak height and retention time for a given sample run under the same conditions. Because of these concerns, there is a considerable interest in developing data preprocessing methods and/or pattern recognition algorithms that can compensate for these experimental limitations.

Our previous studies[5,6] have already demonstrated that artificial neural networks (ANNs) represent a powerful method for pharmaceutical fingerprinting and can provide a better prediction ability compared to traditional chemometric techniques employed for classification. The present study describes a more detailed analysis of the use of ANNs for this purpose, in particular how ANNs can be applied to localize a fingerprinting region within the input space of parameters which, in turn, leads to improved prediction accuracy.

## EXPERIMENTAL SECTION

**HPLC Data.** The present study was conducted on the same HPLC data as previously investigated,[5] i.e., 253

* Corresponding authors.
† Present address: Institut de Physiologie, Université de Lausanne, CH-1005 Lausanne, Switzerland.
‡ E-mail: tetko@bioorganic.kiev.ua.
§ Present address: Ceregen, a Unit of the Monsanto Co., St. Louis, MO 63167.

S0095-2338(97)00439-3 CCC: $15.00 
Published on Web 06/06/1998

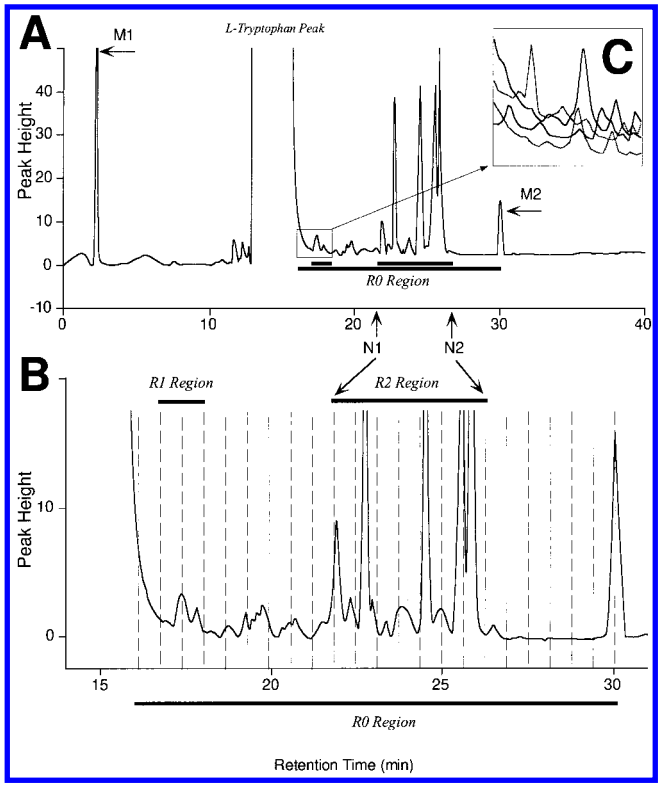


**Figure 1.** Example of the HPLC data used in the present analysis. (A) The location of the early (M1) and late (M2) markers are indicated by arrows. The thick line indicates initial fingerprinting region R0. (B) Illustration of the Windows preprocessing scheme which divides the region R0 into 22 time windows (dashed lines).[5] A subset of windows (regions R1 and R2) identified by the "error-based" pruning method are indicated by thick lines. (C) The instability of the HPLC signal in R1 (data are recorded for the same LT manufacturer) due to the presence in the nearby region associated with the LT peak manifold. The positions of new markers N1 and N2 bracket the beginning and end of the new fingerprinting region R2 that is proposed for manufacturer separation.

chromatographic profiles obtained on L-tryptophan (LT) drug substance from production lots of six different commercial LT manufacturers. It was of particular significance to analyze variations in the HPLC patterns of same-manufacturer samples due to differences in LT production lots, HPLC columns, and even run-days in the experimental design and to quantify what role these factors might play in hampering correct classifications. The experimental design, constructed to account for variations among manufacturers, production lots, HPLC columns, and between-day repeatability, was described earlier.[5] In this design, three to five replicate chromatograms were recorded for each combination of manufacturer, lot, column, and run day. Two markers, M1 and M2, were added to each sample to bracket the retention times of the peaks associated with the LT samples and to normalize the HPLC data as described elsewhere.[5] The region (R0) located between the LT peak manifold and the M2 peak marker (Figure 1) served as the source for initial data using the preprocessing scheme described below.

**Windows Preprocessing Scheme.** The Windows preprocessing scheme was designed[5] in an attempt to isolate the effects of lot-to-lot and column-to-column variations on the performance of the classifiers and to decrease the dimension of the initial input data set. In this scheme, the "fingerprint" region was divided sequentially into 22 time windows of equal width (Figure 1). Although no systematic study was carried out to optimize the number of windows for representing the HPLC data, the width of a single window was selected to be large enough to compensate for the aforementioned effects of column-to-column and lot-to-lot variations, as well as for baseline noise, between chromatograms of samples from the same manufacturer.[5] Each "time window" was analyzed to locate the highest peak $h_{max}$ within it. The resulting series of 22 $h_{max}$ values was then converted to a corresponding series of integer values (designated as H1−H22) according to a procedure described previously.[5] An additional series of 22 input entries (N1−N22) was obtained representing the number of nonnoise peaks in each of the 22 time windows taken in sequential order. To establish an acceptable criterion for *noise* vs *not noise* peak, a statistical analysis of variance was performed on 10 peak heights appearing after the marker M2 and thus presumed as noise. A peak was judged not noise if its magnitude was at least two standard deviations larger that the statistical mean of the 10 noise peaks (calculated at the 95% confidence level). Finally, two parameters were included to provide a cumulative statistic for the entire fingerprint region (i.e., the 22 time windows), viz., *AllPeaks* (corresponding to the total number of nonnoise peaks) and *HPeaks* (corresponding to the number of peaks having an $h_{max}$ value greater than the value of the M2 marker) (Figure 1). The complete set of these 46 parameters for each chromatogram served as the initial input for ANN training and validation. The parameters calculated in this way exhibited characteristically low pairwise correlation coefficients; i.e., $r^2 > 0.35$ (where $r^2$ is the square of the correlation coefficient) for only 8 pairs of parameters (noting that there are $46 \times 45/2 = 1035$ such combinations). In most cases, the highest correlations were found between parameters calculated for the same window, e.g. $r^2 = 0.42$, 0.72 for parameters N1 and H1, N22 and H22 in windows 1 and 22, respectively, or for the same parameter calculated for neighboring windows; e.g., $r^2 = 0.76$ for the number of nonnoise peaks N16 and N17 in adjacent windows 16 and 17.

**Training and Test Sets.** The 253 chromatograms in this classification study included 3−5 replicates for every combination of LT manufacturer, lot, and HPLC column. To remove any bias in the prediction ability of the ANN, as might be caused by placing even one replicate of a test-set chromatogram in the training set, a 6-fold cross-validation procedure was employed as previously indicated.[5,6] The chromatograms were partitioned into six separate combinations of training and test sets (i.e., runs 1−6) in such a way that (1) no chromatogram in the test set would encounter any of its replicates in the training set and (2) each unique combination of LT manufacturer, lot, and HPLC column was included in a test set just once. The sample size of the resulting data sets was 209−215 chromatograms for training and 44−38 chromatograms for testing. Further details on the data handling and processing of the chromatograms are given elsewhere.[5]

## BACKGROUND AND STUDY DESIGN

**Neural Network Architecture.** The ANNs employed in this study are fully connected feed-forward back-propagation networks with one hidden layer and bias neurons.[7] ANN training was accomplished using the SuperSAB algorithm.[8]

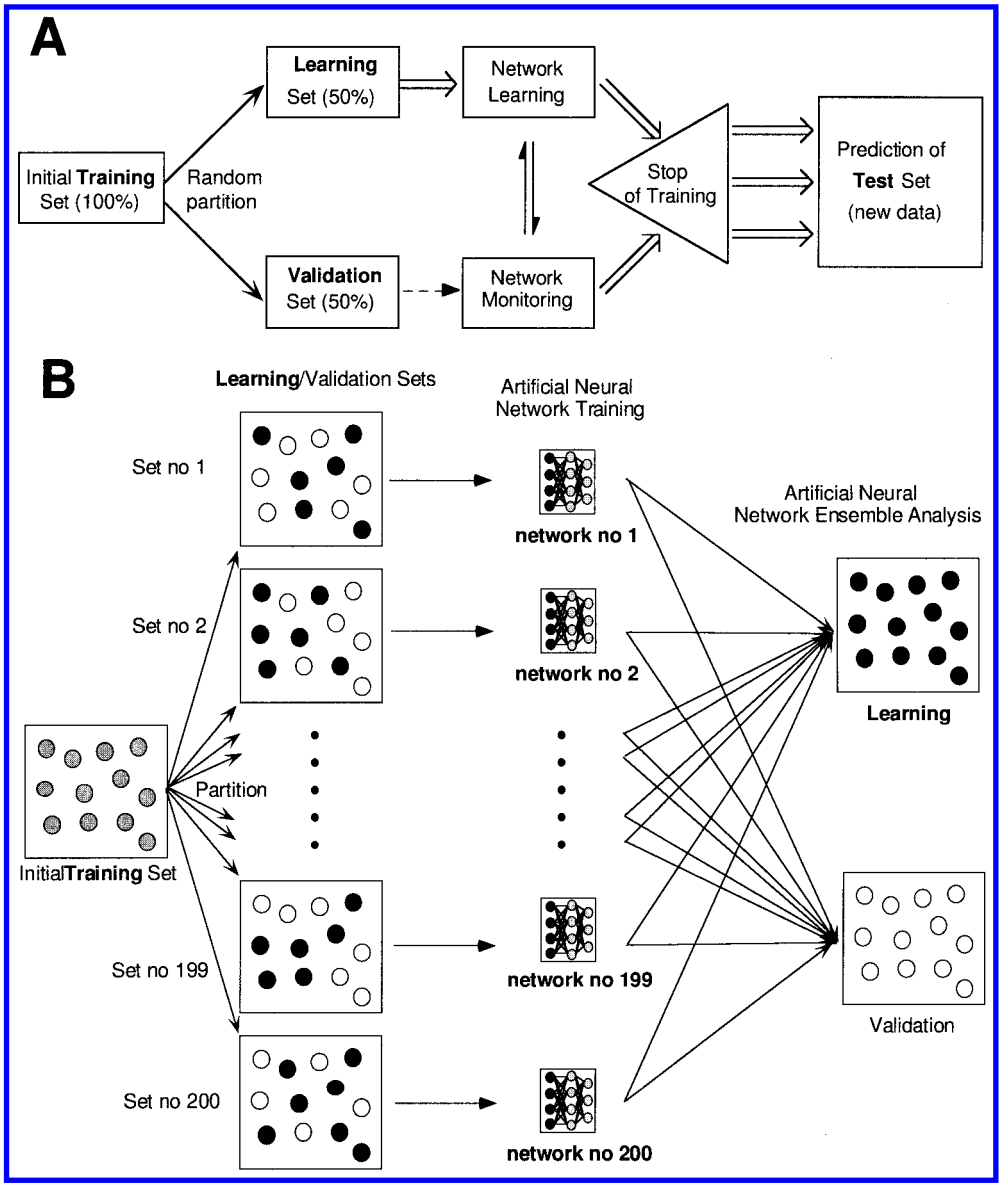


**Figure 2.** Partition procedure and process of neural network learning. (A) Functional scheme. (B) Illustrative example. The gray circles and left-most rectangle represent chromatograms (12 in total, each chromatogram is identified by its position in the rectangle) from the initial training set. Each chromatogram can belong only to a learning or validation set (designated by empty or black circles, respectively) of a given neural network. Because of random partitioning between the learning and validation sets, each chromatogram participates an equal number of times (on average) to both of these sets. After ensemble learning, results are calculated for each chromatogram, as if it belonged simultaneously to both learning and validation sets. The trained networks can be also be used to predict the test set (if available).

The logistic $f(x) = 1/(1 + e^{-x})$ activation function was used both for hidden and output nodes. The number of input nodes was conditioned by the number of used parameters depending on the pruning methods applied. The number of neurons in the hidden layer, set finally at 5, was optimized as indicated in the Results. Six output nodes (one for each LT manufacturer) were used for coding and prediction of the analyzed manufacturers. The output node with the highest numerical value was taken as the predicted LT manufacturer for a single network.

**Training and Classification Considerations.** A crucial feature of any classification method is its ability to generalize beyond the training set and to repeat the same predicted results regardless of the starting point. In our previous study,[5] the classification performance was evaluated using results from a single ANN. However, such interpretations are subject to chance correlation arising from the problem of variance. This type of prediction variance pertains more generally to a family of unstable classifiers (e.g., decision trees, ANNs, etc.). Other common classification methods, such as linear discriminant analysis (LDA) and K nearest neighbors (KNN), represent stable classifiers. The accuracy of a classifier can be described by the so-called fundamental decomposition expression[9]

$$\mathrm{PE}(C) = \mathrm{PE}(C^*) + \mathrm{Bias}(C) + \mathrm{Var}(C) \qquad (1)$$

which states that the misclassification rate PE($C$) of a classifier $C$ is given as a function of the minimum classification rate PE($C^*$) (which is determined by the optimal "Bayes classifier $C^*$" and does not depend on the type of classifier employed), the bias Bias($C$), and the variance Var($C$). All three terms are nonnegative values and contribute to the cumulative error of the classifier. In the typical case, a family of functions $\Omega$ (i.e., set of basis functions) is defined,

and classifier $C$ is selected as a function of $\Omega$ having a minimum misclassification rate over the training-set data. The balance between bias and variance depends on the size of $\Omega$ used to perform a classification. If the family of functions $\Omega$ is small, for instance, if $\Omega$ is the set of linear functions, and the analyzed classification problem is fairly nonlinear, then the bias can be large. However, the variance commonly increases while the bias diminishes as more functions are included in $\Omega$.[9] Stable classifiers are characterized by a small set of functions in $\Omega$ (e.g., the set of linear functions in LDA). Conversely, nonstable classifiers such as ANNs or decision trees are characterized by a large family of functions in $\Omega$. This difference explains why stable classifiers typically yield highly reproducible results (i.e., small variance) but are sometimes lacking in prediction ability (i.e., large bias). In contrast, ANNs and other nonstable classifiers are usually characterized by a low bias and high variance. In the present study, a detailed analysis of pharmaceutical fingerprinting using ANNs is carried out to address the problem of variance associated with nonstable classifiers and to estimate the statistical reliability of the predicted classifications.

A possible solution to this problem is addressed by using an aggregated classifier, i.e., an ensemble of ANNs, simply abbreviated ANNE.[10] Classifications from such an aggregated ensemble of classifiers can be accomplished by majority rule. This procedure can reduce the variance if a sufficient number of networks is aggregated. However, aggregation by itself does not reduce the bias of the classifiers and can sometimes increase it.[9]

Use of the so-called "early stopping" technique in combination with ensemble averaging ("early stopping" over an *ensemble* [ESE]) has been demonstrated to improve prediction ability of ANNs for nonlinear regression.[11] ESE was shown to reduce both the bias and variance of ANNs compared with either method used alone; employing this technique also avoids ANN overtraining, which is regarded as an important factor for variable selection.[12]

The following training scheme was implemented in the present study (Figure 2). Prior to each ANN analysis, the available chromatographic data—the initial training set—was partitioned randomly into learning and validation[13] data sets of equal size. An ensemble of $M = 200$ ANNs was trained after using a random-number generator to initialize the weights of the nodes. It is important to note that in some cases (e.g., for parameter pruning) the initial training set corresponded to a set of $N = 253$ samples and no test set was used. Otherwise, the initial training set consisted of a reduced set of $N = 209-215$ chromatograms (see Experimental Section), while the test set contained the remaining 38−44 chromatograms. The test sets never participated in training and were used only to evaluate the final prediction ability of the network after termination of ANNE training. Since partitioning of the chromatograms was done by chance, each chromatogram of the initial training set appeared in the learning and validation sets an equal number of times (100 ± 10) on average. In a given analysis by a single ANN, each chromatogram was included in either the learning or the validation set—but never simultaneously. Following ensemble learning, it was thus possible to estimate statistical parameters for each chromatogram from the initial training set in that it would belong to both the learning and validation sets. The overall size of the learning and validation sets for ANNE was therefore equal to that of the initial training set (see Figure 2 and refs 10 and 11 for more details of the ESE).

The root-mean-square error (rmse), calculated from analysis of the validation set, was used to monitor the performance of the ANN during the training phase. Training of a particular network was terminated at that epoch when the rmse started to increase ("early stopping point") rather than training to convergence for the learning set. Following training, the ANN weights were stored and then used to estimate various statistical parameters of the ANN classifier as well as its performance for the learning, validation, and test data (when available). Although as many as 2500 epochs were allowed for training, in all cases less than 300 epochs were required to reach the early stopping point[10] and to classify correctly the chromatograms from the learning data set.

**Prediction by Ensemble.** When implementing the ANNE, the LT manufacturer associated with a given chromatogram corresponded to that manufacturer predicted by the majority of the ANNs comprising the ensemble. Two criteria for determining the majority were employed: (1) simple *majority voting* (MV) and (2) *majority voting according a sign criterion* (MV95).[14] In the latter case, a prediction was considered "incorrect" if it was impossible to correctly classify the chromatogram according to LT manufacturer at the 95% level of confidence. Unless otherwise stated, each ensemble in this study was composed of 200 neural networks.

**Pruning.** Optimization of a set of input parameters has been shown to improve significantly the generalization ability of ANNs for nonlinear regression.[12,15] Use of a reduced set of parameters could provide additional advantages for interpretation of a classification study. In the present analysis, pruning was applied in order to determine which input parameters from the chromatographic data are the most significant for the pattern recognition of LT manufacturers by ANNs. Of five previously investigated pruning methods (designated A−E),[12] only A, B, and D were selected for this study.

The main principles of the pruning algorithm employed in this study are summarized as follows. A sensitivity $S_i$ of input variable $i$ was introduced by

$$S_i = \sum_{k \in \Omega_i} s_k \equiv \sum_{j=1}^{n_j} s_{ji} \qquad (2)$$

where $s_k$ is the sensitivity of the weight $w_k$, and the summation is carried over a set $\Omega_i$ of outgoing weights of the neuron $i$ or, using another order of weight numeration, $s_{ji}$ is the sensitivity of the weight $w_{ji}$ connecting the $i$th neuron to the $j$th neuron in the next layer.

**Method A.** The first method explores the idea that input neurons with bigger connection weights to the hidden and output layers play a more significant role than other input neurons. Therefore, the absolute magnitude of weight $w_k$

$$s_k = |w_k| \qquad (3)$$

is used as its sensitivity in eq 2.

**Method B.** The sensitivity in the second method is calculated by

$$S_i = \sum_{j=1}^{n_j} \left( \frac{w_{ji}}{\max_a |w_{ja}|} \right)^2 S_j \quad (4)$$

where $\max_a$ is taken over all weights ending at neuron $j$, $S_j$ is a sensitivity of the $j$th neuron in the upper layer, and $n_j$ is the number of hidden neurons. The sensitivities of the output layer neurons are set to 1.

**Method D.** Optimal brain damage[16] uses the second derivative of the error function with respect to the neuron weights to compute the sensitivities as follows:

$$s_k = (w_k)^2 (\partial^2 E / \partial w^2_k) \quad (5)$$

This weight sensitivity is used in eq 2.

To summarize, the sensitivity of a particular input node for methods A and B was estimated by analysis of the magnitudes of its outgoing weights (i.e., a "magnitude-based" method), while for method D it was measured by the change in network error resulting from elimination of some neuron weights (i.e., an "error-based" method). The sensitivity of input parameters was estimated at the early stopping points for all ANNs, followed by pruning of the least sensitive input. Pruning was terminated when the number of inputs was exhausted.

It should be noted that method C was excluded on the basis of its poor ability to generalize.[12] This method estimated the sensitivity of an input parameter by using the linear term in the Taylor series expansion of the error function. The linear term approaches zero if pruning is done on a well-trained function, and, most likely, this fact contaminated the prediction ability of this method. Method E, the other one excluded, is programmed according to the optimal brain surgeon (OBS) technique developed by Hassibi and Stork.[17] This method provided reliable estimation of the sensitivities of input variables as reported elsewhere.[12] However, it required a prohibitive amount of computing time (the speed of OBS decreases as a function of $Nz^2$, where $z$ is the number of weight connections and $N$ is the number of cases under study) to be applied for analysis of the current data set.

The computer codes for the ANN and pruning algorithms were programmed in ANSI C++. The calculations were performed at HP Workstation Cluster at the Swiss Center for Scientific Computing (CSCS).

## RESULTS

**Selection of ANN Architecture.** The original 46 input parameters, extracted from each chromatogram using the Windows preprocessor method,[5] comprised the initial input data set for subsequent analysis. A preliminary study was performed to determine the optimal number of hidden layer neurons for the present ANN analysis (Figure 3). The prediction accuracy for the test data sets was found to increase with addition of hidden neurons up to 5−6 neurons. Further addition of hidden neurons did not influence ANNE performance, which varied only slightly about a mean value. The number of hidden neurons was therefore fixed at 5 for all subsequent calculations. This number of hidden neurons yielded a good balance between the prediction ability with ANNE and the amount of computer time required.

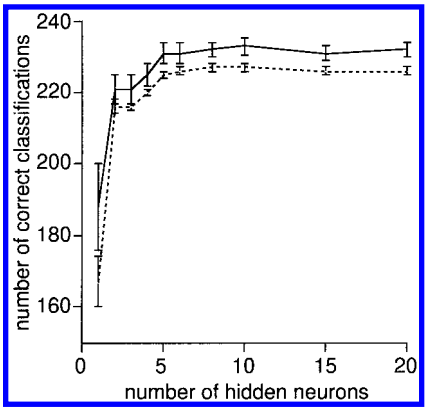


**Figure 3.** Prediction ability of an artificial neural network ensemble, calculated using 6-fold cross-validation procedure, as a function of the number of hidden neurons in the neural networks. The initial set of 46 parameters was used to train the ensemble. The solid line traces the number of chromatograms correctly predicted by majority voting (MV). The dashed line indicates only those chromatograms correctly predicted by MV with a level of significance of $p < 0.05$ (MV95). The error bars correspond to standard deviations estimated by analysis of 10 ensembles.

The ANNs trained to convergence exhibited, on average, lower prediction ability than those trained to an early stopping point. For example, the number of chromatograms correctly predicted for ANNs with 5 hidden neurons using early stopping and convergence were 225 and 222, respectively, by MV95 and 232 and 230, respectively, by MV. It is thus seen that the prediction ability of ANNs can be improved by implementing ESE as opposed to converged ANNs. Still, additional studies are required to validate the statistical validity of this improvement.

**Statistical Analysis of ANN Predictions.** The predicted classifications for chromatograms 1−4 (Table 1) from a test data set yielded a significant amount of variance, even though the global parameters of the respective ANNs were identical. It is possible that a single ANN could, by sheer chance, correctly classify every chromatogram in the test set. However, such an outcome would likely overestimate the true prediction ability of the ANN. The repeatability of making correct classifications was improved by adopting the ANNE approach with predictions based on MV. Nevertheless, there were some chromatograms (see, for example, chromatogram 2) where the predicted manufacturer might have easily been reversed by altering the number of ANNs in the ensemble. These chromatograms contributed significantly to the variance of the MV predictions. Using the alternative MV95 method, a classification for which the level of significance was $p > 0.05$ was regarded as incorrect. This technique reduced the variance of ANNE performance.

For almost every test-set chromatogram, the confidence level of the predictions improved as the number of ANNs increased in the ANNE analysis (Table 2, chromatogram 1). In some chromatograms (Table 2, chromatogram 2), however, it was impossible to make a significant prediction even when the ensemble contained as many as 10 000 ANNs. Such chromatograms were at the limit of class separation (e.g., the probability that chromatogram 2 belongs to manufacturers B or E is almost identical) and were responsible for the large variance in the predicted results according to the MV method.

**Optimization of Input Parameters.** To determine the most relevant set of features required for correct identification

**Table 1.** Example of ANNE Predictions for Test Chromatograms[a]

| | single ANN predictions | | | | | | | ANNE analysis | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| chromatogram | 1 | 2 | 3 | 4 | 5 | 6 | 7 | A | B | C | D | E | F | MV | MV95 | LTM | *p* |
| 1 | A | A | A | A | D | A | A | *166* | 0 | 10 | 24 | 0 | 0 | A | A | A | $<10^{-6}$ |
| 2 | B | B | C | E | B | E | F | 2 | *64* | 12 | 36 | 60 | 26 | B | ? | B | >0.3 |
| 3 | A | A | A | D | E | A | D | *130* | 0 | 5 | 53 | 5 | 7 | A | A | D | $<10^{-8}$ |
| 4 | E | D | E | F | E | A | E | 1 | 13 | 4 | 8 | *141* | 33 | E | E | E | $<10^{-17}$ |

[a] ANN, artificial neural network; ANNE, ANN ensemble; MV, majority voting; MV95, majority voting according to sign criterion (see details of calculations in the text); LTM, L-tryptophan manufacturer. The results are reported for training/test set partitioning of chromatograms in run 1 (see Experimental Section). Columns labeled 1−7 illustrate examples of single ANN predictions for test chromatograms. The ANN had 46 inputs, 5 hidden and 6 output neurons. The results were calculated at early stopping points. Columns labeled A−F count total numbers of ANN predictions for each manufacturer following ANNE analysis. The maximal counts calculated for manufacturers are italicized. The final ensemble predictions according to MV are shown in the next column. Column LTM indicates the actual LT manufacturer for each chromatogram. The last column shows the significance level (according to sign criterion) of the aggregated prediction. Note that chromatogram 2, for which the significance level is very low, is considered as incorrectly predicted by MV95.

**Table 2.** ANNE Results for Test Chromatograms in Which Dependency on the Number of Composite ANNs Was Analyzed[a]

| | chromatogram 1 | | | | | | | chromatogram 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| no. of ANNs | A | B | C | D | E | F | *p* | A | B | C | D | E | F | *p* |
| 10 | *7* | 0 | 1 | 2 | 0 | 0 | >0.04 | 1 | 2 | 0 | 2 | 5 | 0 | >0.1 |
| 50 | *39* | 0 | 3 | 8 | 0 | 0 | $<10^{-6}$ | 0 | 15 | 1 | 8 | 22 | 4 | >0.1 |
| 200 | *166* | 0 | 10 | 24 | 0 | 0 | $<10^{-28}$ | 2 | *64* | 12 | 36 | 60 | 26 | >0.3 |
| 500 | *411* | 1 | 32 | 54 | 0 | 2 | 0 | 2 | *166* | 34 | 86 | 157 | 55 | >0.3 |
| 2000 | *1,599* | 5 | 163 | 223 | 7 | 3 | 0 | 9 | 606 | 111 | 366 | *641* | 267 | >0.16 |
| 5000 | *4,008* | 12 | 374 | 582 | 18 | 6 | 0 | 19 | 1555 | 244 | 903 | *1635* | 644 | >0.08 |
| 10000 | *8,057* | 23 | 693 | 1173 | 37 | 17 | 0 | 45 | 3201 | 518 | 1,773 | *3314* | 1,149 | >0.08 |

[a] The results are reported for training/test set partitioning of chromatograms in run 1 (see Experimental Section). Columns count the total number times a specific manufacturer (A−F) was predicted by the ANNs composing the ANNE in question. The maximal counts calculated for a given ensemble (i.e., as selected by the MV approach) are italicized. The level of significance of the predictions increases, corresponding to the value of *p* decreasing, with the number of ANNs per ensemble for both chromatograms 1 and 2, but even a very large number of ANNs (10 000) was not adequate to obtain a statistically significant prediction for chromatogram 2. In fact, chromatogram 2 was predicted incorrectly by MV95 for every number of ANNs per ensemble. Note that two ANNEs, one with 200 ANNs and another with 500 ANNs, predicted the correct LT manufacturer, but this result can be attributed to chance correlation.

of each LT manufacturer, the entire data set of 253 chromatograms was taken as the initial training data set to enable optimization of the input parameters using various pruning methods. The prediction ability of an ANN was estimated on the basis of its performance for the validation data set only. It had been shown previously that the presence of replicate chromatograms in both the training and test sets overestimated the prediction ability of ANNs.[5]

During parameter pruning, the prediction ability decreased for only 1−3 chromatograms (depending on the selected pruning procedure) even if 28−35 out of the 46 input parameters were deleted from the input data set. Further pruning of parameters, however, rapidly decreased ANN performance (Figure 4). Those input sets that still, after pruning, accurately classified 250 out of the 253 chromatograms (based on MV95[18]) for the validation data set were chosen for further analysis. The sets of selected parameters were very similar in each case (Table 3). Two data sets identified by the magnitude-based pruning methods A and B were identical with the exception of one parameter. Also, half of the parameters selected by the error-based pruning method D coincided with those selected by methods A and B.

An estimation of a prediction ability of the pruned parameter sets was carried out using the 6-fold cross-validation procedure described in the Experimental Section. The best prediction ability was obtained using the parameter set extracted by pruning method D (Table 3).

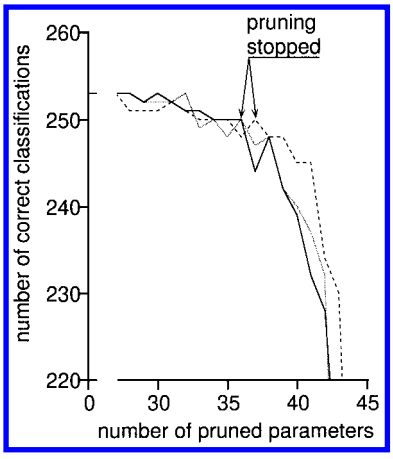


**Figure 4.** Prediction ability of an ANNE for the validation data set estimated by MV95, as a function of the number of parameters in the input data set. The dark and light solid lines and the dashed line correspond to pruning methods A, B, and D, respectively. The arrows correspond to the number of parameters at which the pruning was terminated.

One could argue that using the same data set for pruning and validation, as done here, can introduce a bias when estimating the prediction ability of a pruning method. This possibility was tested for the OBD-based method D which was applied to optimize the input parameters for each of the training data sets for runs 1−6. The optimized parameter sets were used to classify the test-set chromatograms in the corresponding run. The number of correct classifications

**Table 3.** Average and Standard Deviation of Prediction Accuracy of ANNs, Calculated Using 6-Fold Cross-Validation Procedure for Different Sets of Input Parameters

| no. | param sets | no. of params | description of param sets | prediction accuracy[a] |
|---|---|---|---|---|
| 1 | H1−H22, N1−N22, AllPeaks, HPeaks | 46 | initial set from fingerprint region R0 | 231 ± 3 (225 ± 2) |
| 2 | H2−H7, H16, N15, N18, N21, HPeaks | 11 | set selected by pruning method A | 214 ± 2 (211 ± 1) |
| 3 | H2−H7, H14, H16, N15, N21, HPeaks | 11 | set selected by pruning method B | 230 ± 2 (222 ± 1) |
| 4 | H2−H3, H12, H14−H16, N10, N11, N15, HPeaks | 10 | set selected by pruning method D | 238 ± 1 (235 ± 1) |
| 5 | H12, H14−H16, N10, N11, N15, HPeaks | 8 | reduced set of params from region R2 | 231 ± 2 (226 ± 1) |
| 6 | HPeaks | 1 | | 135 ± 0 (125 ± 0) |
| 7 | HPeaks, HR2Peaks | 2 | | 190 ± 1 (181 ± 1) |
| 8 | H12, H14−H16, N10, N11, N15, HPeaks, HR2Peaks | 9 | the same as set no. 5 + HR2Peaks | 240 ± 0 (240 ± 0) |

[a] Correct prediction of manufacturers over the total of 253 chromatograms evaluated. See description of parameters in Experimental Section. MV and MV95 (values in parentheses) were used to predict the LT manufacturer for each test chromatogram following ANNE calculation. Standard deviations of ensemble predictions were estimated by analysis of 10 ensembles. See text about detailed description of training/test protocol and training procedure of ANNs.

obtained was 230 out of 253 compared with 237 out of 253 (i.e., only 2% higher) with the parameters optimized for the entire data set. This result suggests that pruning methods are relatively insensitive to the number of chromatograms in training data sets.

A subset of windows from the aforementioned fingerprint region was identified by pruning method D as particularly useful for discriminating between the six LT manufacturers (Table 3). The first region (R1) consisted of windows 2−3 while the second region (R2) consisted of windows 10−16. A comparison of the HPLC data from these two regions reveals a high number of nonstationary peaks in the windows 2−3 (Figure 1). This particular region is located near the LT peak manifold. Small variations in position and magnitude of the LT peak can substantially influence the identification of the fingerprinting peaks by the Window preprocessor scheme. This factor explains why the input parameters extracted from windows 2−3 exhibit fluctuations that could bias the prediction ability of the classifier. As a result, it was deemed preferable to exclude this region from the data analysis.

Elimination of the input parameters from region R1 decreased the prediction ability of the ANN (Table 3). To compensate for this loss of information, a new parameter designated as HR2Peaks was introduced. This parameter counted the total number of very high peaks—peaks with height greater than that of marker M2—in region R2. This new parameter is analogous to parameter HPeaks which calculates the number of such peaks in the original fingerprinting region R0 (see Experimental Section). It should be noted that the parameter HPeaks was selected by all pruning methods, suggesting the importance of this parameter for the classification task at hand. In fact, using just this single parameter for ANN training provided a prediction accuracy of 135 out of the total 253 (53%) chromatograms using both MV and MV95 criteria (Table 3). Inasmuch as the region R2 is characterized by a vast number of such high peaks, these results indicate that the total count of "very high" peaks in that region is a critical factor in discriminating among the LT manufacturers. Indeed, even a simple combination of parameters HPeaks and HR2Peaks yielded a prediction accuracy of 75%, thus stressing the significance of both parameters for LT manufacturer separation.

Inclusion of parameter HR2Peaks into the optimized input parameter set increased the prediction ability of ANNE for our 6-fold cross-validation scheme. This parameter set exceeded all others considered in the present study in terms of classification performance. It is interesting to note that predictions obtained from this parameter set were found to be identical by both the MV and MV95 methods; i.e., the classifications of all test chromatograms were significant.

## DISCUSSION

The present work demonstrates that optimization of a set of input parameters by pruning can improve the prediction ability of the ANN classifier. Analysis of the input parameters selected by pruning methods is useful for interpreting the predicted results, for setting the boundaries of the fingerprinting region, and for introducing new parameters containing more information for making classifications more efficiently. The current approach can be extended for detection of localized fingerprinting regions and interpretation of results is neural network analysis of mass spectra[19] and infrared spectra[20,21] or for QSAR (quantitative structure−activity relationship) studies using, for example, "spectrumlike" representations of chemical structures.[22] The general idea of pruning can be also used for interpretation of more complex ANNs, such as the neural device proposed by Bashkin et al.[23]

There are certain advantages gained by decreasing the width of the fingerprint region. For example, the familiar drift of HPLC signals is known to increase with retention time due to the physical conditions of the chromatographic experiment (temperature, pressure, etc.)[24] Therefore, reducing the fingerprinting region to only a small retention-time span would tend to minimize this problem. Furthermore, two new markers N1 and N2 could be chosen that precisely bracket the beginning and end of the fingerprinting region, as shown in Figure 1. Using these markers to normalize the retention time of the chromatograms within this region would likely reduce the variance of the interpeak intervals more than normalization of the entire chromatographic region as done previously.[5] As a result, this refined procedure would enhance the inherent ability of the ANN classifier to generalize.

The new set of optimized parameters found in the present study contains only one parameter, HPeaks, that was extracted from the original fingerprinting region R0 by the Windows preprocessing scheme. This parameter counts the number of very high peaks (see Experimental Section). Analysis of the entire 253 chromatograms has shown that such peaks could not be found beyond the R2 region, i.e., after the marker N2. For this reason, no additional marker is required to bracket the data to extract this parameter.

Compared with other parameter optimization techniques, a significant advantage of the pruning algorithms employed here is their relative speed. In particular, pruning techniques based either on iterative removal of the least pertinent parameters or on exhaustive analysis of all possible combinations of the input parameters would require substantially more computer time than required for pruning method A, B, or D.

Comparison of pruning methods A, B, and D shows that input parameters selected with the error-based method D yielded the highest prediction ability compared to parameters selected with the magnitude-based methods A and B. Similar studies of ANN regression have indicated[12] that pruning of parameters by method B gave the best generalization ability, implying that the pruning method of choice is dependent on the specific nature of the problem under study. At the present time, it is impossible to predict a priori which pruning method is best suited for a particular task or for a given set of input data. As demonstrated here, the best course of action might be to select that pruning method giving the best results after evaluation of all available pruning methods.

One problem in applying pruning algorithms is the absence of any objective criteria for terminating parameter selection and optimization. Most regression studies monitor the predicted error as measured by the rmse for the validation set and terminate further parameter pruning when the predicted error reaches some defined minimum value. Of course, this procedure assumes that this predicted error provides a good estimation of the error for new test sets. In a similar fashion, the present analysis terminated pruning by monitoring the predicted classification error as measured by the number of classification errors for the validation data set. However, our earlier investigations have shown that classifications in the present application are made complicated by the presence of replicates for the same class which it turn can bias estimations of the prediction ability.[5] For example, the present analysis indicates that the classification error estimated using the validation set differed significantly from the error estimated using the training/test set protocol. This suggests that the choice of criteria for termination of pruning might itself impart bias.

Further investigation is required to develop training and estimation procedures that are insensitive to problems associated with the existence of replicates in training data sets. The efficient partition algorithm (EPA) offers a possible solution to this problem.[25,26] EPA selects the number of input parameter sets for ANN training in proportion to the complexity of the data set itself and, therefore, compensates in part for the aforementioned problem associated with replicates. Until recently, the development and application of this algorithm has been reserved to ANN regression.[25,26] Current efforts are aimed toward expanding the applicability of EPA to ANN classification.[27]

Confirming previous investigations,[14] the present study has demonstrated the utility of the ensemble aggregation approach for diminishing variance in classification studies due to chance correlation. An important corollary of this finding is that evaluation of ANN performance based on analysis of a single ANN is subject to the chance correlation problem which, in turn, can lead to overestimation of the true prediction ability and/or inadvertent bias. As expected, increasing the number of networks per ensemble enhances the confidence of results obtained from averaging across the ANNs. However, for some chromatograms it might not be possible to analyze an ensemble of ANNs that is large enough to render a classification with statistical confidence. An example is chromatogram 2 (Table 2) where classification at the 95% confidence level according to MV95 was not possible even by using an ensemble of 10 000 ANNs. Such chromatograms are located at the boundary of class separation in the feature space adopted here, making class discrimination virtually impossible. Classifications in such cases by ANNE are tantamount to random guesses. That is why the prediction ability of an ANNE can be overestimated (e.g., if ANNE correctly predicts such chromatograms by mere chance) unless the significance of every classification for each particular test chromatogram is tested. This criterion is satisfied by the MV95 approach but not by MV. For this reason, the MV95 approach is recommended in order to minimize the possibility of chance correlation when dealing with neural network applications. It is important to note that using the set of variables optimized by MV95 dramatically increased the prediction ability compared with the original analysis using the total set of 46 parameters, i.e., 240 (95%) versus 225 (89%) correctly predicted manufactures.

In addition to improving the prediction ability of ANNs, optimization of input parameters was also found to diminish the differences between MV and MV95 in terms of relative accuracy; i.e., all test chromatograms were predicted with a level of significance $p < 0.05$. This latter factor can be explained, from a theoretical point of view,[28] as being due to better separation of chromatograms from unlike LT manufacturers and better clustering of chromatograms from like LT manufacturers for the optimized (i.e., pruned) feature space over the original space of 46 input parameters.

## ACKNOWLEDGMENT

This study was partially supported by NATO HTECH.LG 972304, INTAS-Ukraine 95-0060, and the Swiss National Science Foundation FNRS 31-37723.93 grants. The overall project is supported in part by equipment grants from the Center for Molecular Electronics of the University of Missouri—St. Louis and by a contract with the FDA Division of Drug Analysis, St. Louis, MO, administered by Thomas P. Layloff. The authors wish to express their appreciation to Samuel W. Page of the FDA Center for Food Safety and Nutrition, Washington, DC, and Robert Hill of the Centers for Disease Control (CDC), Atlanta, GA, for providing the samples of the L-tryptophan bulk substance used in these studies. We also thank Brian Hyland and an anonymous reviewer for their helpful suggestions.

## REFERENCES AND NOTES

(1) Layloff, T. P. Scientific Fingerprinting: A Pharmaceutical Regulatory Tool. *Pharm. Technol.* **1991**, *15*, 146−148.

(2) Kirchhoefer, R. D. An FDA Laboratory Approach to Uncovering Potential Fraud in the Generic Drug Industry. *J. AOAC Int.* **1992**, *75*, 577−580.

(3) Haddad, W. *The Pink Sheet*; F-D-C Reports, Inc.: Chevy Chase, Md, 14 August 1989; p T&G 6.

(4) Inman, E. L.; Tenbarge, H. J. High-Low Chromatography: Estimating Impurities in HPLC Using a Pair of Sample Injections. *J. Chromatogr. Sci.* **1988**, *26*, 89−94.

(5) Welsh, W. J.; Lin, W.; Tersigni, S. H.; Collantes, E.; Duta, R.; Carey, M.; Zielinski, W. L.; Brower, J.; Spencer, J. A.; Layloff, T. P. Pharmaceutical Fingerprinting: Evaluation of Neural Networks and

Chemometric Techniques for Distinguishing among Same-Product Manufacturers. *Anal. Chem.* **1996**, *68*, 3473−3482.
(6) Collantes, E. R.; Duta, R.; Welsh, W. J.; Zielinski, W. L.; Brower, J. Preprocessing of HPLC Trace Impurity Patterns by Wavelet Packets for Pharmaceutical Fingerprinting Using Artificial Neural Networks. *Anal. Chem.* **1997**, *69*, 1392−1397.
(7) For example: Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH Publishers: New York, 1993.
(8) Tollenaere, T. SuperSAB: Fast Adaptive Back Propagation with Good Scaling Properties. *Neural Networks* **1990**, *3*, 561−573.
(9) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123−140.
(10) 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.
(11) Tetko, I. V.; Villa, A. E. P. An Enhancement of Generalization Ability in Cascade Correlation Algorithm by Avoidance of Overfitting/ Overtraining Problem. *Neural Process. Lett.* **1997**, *6*, 43−50.
(12) Tetko, I. V.; Villa, A. E. P.; Livingstone, D. J. Neural Network Studies. 2. Variable Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794−803.
(13) It should be noted that in our previous work[10] we referred to this set as "control set". However, the term "validation set" is more generally used in computer science literature (see for example: Leblanc, M.; Tibshirani, R. Combining Estimates in Regression and Classification. *J. Am. Statist. Assoc.* **1996**, *91*, 1641−1650) and will be used in the current study too.
(14) Tetko, I. V.; Luik, A. I.; Poda, G. I. Application of Neural Networks in Structure−Activity Relationships of a Small Number of Molecules. *J. Med. Chem.* **1993**, *36*, 811−814.
(15) Hosseini, M.; Maddalena, D. J.; Spence, I. Using Artificial Neural Networks To Classify the Activity of Capsaicin and Its Analogues. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1129−1137.
(16) LeCun, Y.; Denker, J. S.; Solla, S. A. Optimal Brain Damage. In *Advances in Neural Processing Systems 2 (NIPS*2);* Touretzky, D. S., Ed.; Morgan-Kaufmann: San Mateo, CA, 1990; pp 598−605.
(17) Hassibi, B.; Stork, D. Second-Order Derivatives for Network Pruning: Optimal Brain Surgeon. In *Advances in Neural Processing Systems 5 (NIPS*5)*; Hanson, S., Cowan, J., Giles, C., Eds.; Morgan-Kaufmann: San Mateo, CA, 1993; pp 164−171.
(18) These sets also correctly predicted 252 chromatograms according to MV.
(19) Eghbaldar, A.; Forrest, T. P.; Cabrol-Bass, D.; Cambon, A.; Guigonis, J.-M. Identification of Structural Features from Mass Spectroscopy Using a Neural Network Approach: Application to Thrimethylsilyl Derivatives Used for Medical Diagnosis. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 637−643.
(20) Klawun, C.; Wilkins, C. L. Joint Neural Network Interpretation of Infrared and Mass Spectra. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 249−257.
(21) Klawun, C.; Wilkins, C. L. Optimization of Functional Group Prediction from Infrared Spectra Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 249−257.
(22) Novic, M.; Nikolovska-Coleska, Z.; Solmajer, T. Quantitative Structure−Activity Relationship of Flavonoid p56lck Protein Tyrosine Kinase Inhibitors. A Neural Network Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 990−998.
(23) Bashkin, I. I.; Palulin, V. A.; Zefirov, N. S. A Neural Device for Searching Direct Correlations between Structures and Properties of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 715−721.
(24) Johnson, E. L.; Stevenson, R. *Basic Liquid Chromatography*; Varian Associates, Inc.: Palo Alto, CA, 1978.
(25) Tetko, I. V.; Villa, A. E. P. Efficient Partition of Learning Datasets for Neural Network Training. *Neural Networks* **1997**, *10*, 1361−1374.
(26) Tetko, I. V.; Villa, A. E. P. An Efficient Partition of Training Data Set Improves Speed and Accuracy of Cascade-Correlation Algorithm. *Neural Process. Lett.* **1997**, *6*, 51−59.
(27) Tetko, I. V.; Villa, A. E. P.; Welsh, W. J. Work in progress.
(28) Aivazyan, S. A.; Buchstaber, V. M.; Yenyukov, I. S.; Meshalkin, L. D. *Applied Statistics. Classification and Reduction of Dimensionality*; Finansy i statistika: Moscow, 1989.

CI970439J