

An Integrated Soft Computing Approach for Predicting Biological Activity of Potential HIV-1 Protease Inhibitors

Răzvan Andonie*, Levente Fabry-Asztalos[†], Sarah Abdul-Wahid*, Catharine J. Collar[†], and Nicholas Salim[†]

*Computer Science Department

Central Washington University, Ellensburg, USA

Email: {andonie, abdulwahids}@cwu.edu

[†]Department of Chemistry

Central Washington University, Ellensburg, USA

Email: {fabryl, collarc, salimn}@cwu.edu

Abstract—Using a neural network-fuzzy logic-genetic algorithm approach we generate an optimal predictor for biological activities of HIV-1 protease potential inhibitory compounds. We use genetic algorithms (GAs) in the two optimization stages. In the first stage, we generate an optimal subset of features. In the second stage, we optimize the architecture of the fuzzy neural network. The optimized network is trained and used for the prediction of biological activities of newly designed chemical compounds. Finally, we extract fuzzy IF/THEN rules. These rules map physico-chemical structure descriptors to predicted inhibitory values. The optimal subset of features, combined with the generated rules, can be used to analyze the influence of descriptors.

I. INTRODUCTION

Artificial neural networks are particularly well-suited for QSPR (Quantitative Structure-Property Relationship) and QSAR (Quantitative Structure-Activity Relationship) because of their ability to extract both linear and nonlinear information present in the mapping of physico-chemical descriptors to biological activity [1], [2]. Neural models have been used to predict properties of chemical compounds such as inhibition of HIV-1 reverse transcriptase [3], lipophilicity and aqueous solubility [4], intestinal absorption [5], and site of protease cleavage [6]. Several neural architectures were successful for such tasks: backpropagation [3], [7], [8], associative neural networks [4], probabilistic neural networks [9], generalized regression neural networks [10], radial basis function networks [6], [10], cascade correlation [11], neural networks trained via evolutionary algorithms [1], fuzzy ARTMAP [12]–[14], self-organizing maps [15], fuzzy neural networks [16], and support vector machines [17].

Choosing a molecular representation for efficient computer handling is often time-consuming, since there are so many kinds of molecular descriptors. DRAGON, for example, provides more than 1,600 molecular descriptors [18]. In QSPR/QSAR neural networks modeling, two different methods have been proposed for the definition of descriptors [11], [19]: *i)* using known physico-chemical properties; and *ii)* using topological indices, which code specific morphological properties of the molecular graph. Another option is to process directly the chemical graph structures [11], [19], [20].

The general approach for the application of neural networks to QSPR/QSAR is to represent the molecule by a vector of features which may be both topological indices and physico-chemical properties. The problem with this representation is that it may give only a partial description of the molecular structure, making the representation non-reversible. The number of vector features may be very large and many of them may be inter-correlated or redundant.

Using the complete set of descriptors may lead to overfitting if the number of descriptors is too large compared to the size of the training set [3]. Several authors have used feature selection techniques (principal component analysis, genetic algorithms, mutual information) for reducing the number of chemical descriptors [6], [10], [16], [21]–[23]. Feature selection is also important because of the high inter-correlation amongst features.

In previous work [2] we investigated the use of a fuzzy neural network (FNN) for biological activity (IC_{50}) prediction¹. Besides the flexibility to deal with noisy and incomplete data, our approach had the capability to easily expose the learned knowledge in the form of fuzzy IF/THEN rules.

In this paper, we integrate a two-stage GA optimization of our FNN. First, we select the subset of features that are useful to build a good predictor. Based on this optimal subset of features, we generate in the second phase the optimal FNN architecture. We use this approach to predict biological activities of newly designed chemical compounds. From the generated FNN model, we extract rules. The number of generated rules, equal to the number of hidden nodes in the FNN, is a function of the optimization parameters.

The importance of predicting biological affinity for potential HIV-1 protease inhibitors and the descriptors used are described in Section II, while the GA approach for feature selection is explained in Section III. The GA optimization of the FNN is addressed in Section IV. Section V describes our approach. Results obtained from the GA-optimized FNN with 30 descriptors (GA-FNN) and the GA-optimized FNN with an optimal subset of features (FS-GA-FNN) are compared in Section VI. Section VII concludes with closing

¹The IC_{50} value represents the concentration of a compound that is required to reduce enzyme activity by 50%.

TABLE I
MOLECULAR DESCRIPTORS

no.	Descriptor
D1	Total Number of Atoms
D2	Number of Bromine Atoms
D3	Number of Carbon Atoms
D4	Number of Chlorine Atoms
D5	Number of Fluorine Atoms
D6	Number of Hydrogen Atoms
D7	Number of Nitrogen Atoms
D8	Number of Oxygen Atoms
D9	Number of Sulfur Atoms
D10	Molecular Volume
D11	Index of Hydrogen Deficiency
D12	Total Number of Bonds
D13	Number of Single Bonds
D14	Number of Double Bonds
D15	Number of Triple Bonds
D16	Number of Aromatic Rings
D17	Number of Amide Bonds
D18	Molecular Weight
D19	Total Charge
D20	Bond Stretching Energy
D21	Angle Bending Energy
D22	Torsional Energy
D23	Out of Plane Bending Energy
D24	One to Four Van Der Waals Energy
D25	Van Der Waals Energy
D26	One to Four Electrostatic Energy
D27	Electrostatic Energy
D28	Van Der Waals Electrostatic Pairs
D29	One to Four Van Der Waals Electrostatic Pairs
D30	Scaled Van Der Waals Electrostatic Pairs

remarks.

II. PREDICTION OF IC_{50} FOR HIV-1 PROTEASE POTENTIAL INHIBITORS

Extensive research has resulted in many HIV-1 protease inhibitors [24]–[27]. Current treatments for HIV/AIDS consist of co-administering a protease inhibitor with two reverse transcriptase inhibitors (usually referred to as combination therapy). This therapy is effective in reducing viremia to very low levels; however, in 30-50% of patients it is ineffective due to resistance development often caused by viral mutations. Due to poor resistance and bioavailability² profiles, as well as toxicity associated with these therapies, there is an urgent need for the development of more efficient drugs.

Computer aided design is an essential component of the modern drug discovery process. Techniques used include molecular modeling, neuro-fuzzy systems, genetic algorithms, and statistical analyses. Using these techniques, *de novo* structures with steric and chemical complementarity to an enzyme active site are generated and their biological

²Bioavailability is the rate at which the drug reaches the systemic circulation.

activities are predicted [28]–[30]. This increases the efficiency of drug discovery by shortening the otherwise lengthy drug design, synthesis, and evaluation process. Therefore, new methods that aid in potential lead compound design and biological activity prediction are expected to be a great asset to drug discovery. The methodology described here (i.e. feature selection, prediction, and rule extraction) can assist in this direction.

In our study, 30 molecular descriptors were selected, based on their contribution to molecular entity [2]. These descriptors are displayed in Table I. SYBYL³ molecular modeling software was used to create the data files that provided descriptors. Descriptors were then normalized according to the formula $(x - y)/z$, where x is the actual value for the descriptor of the inhibitor of interest, y is the minimum descriptor value for the data set, and z is the range (maximum-minimum) value for the data set. As a result of normalization, descriptors for all the molecules in the data set fell in the range [0, 1].

The data set for training and testing consisted of 151 known compounds with experimentally determined IC_{50} values [31]–[34]. A total of 26 novel compounds were designed via a "combinatorial" approach using only known compounds with low IC_{50} and high TI values⁴. Side chains were then modeled on all core structures, excluding the one from which they came from, into their original binding pockets.

III. GA-FEATURE SELECTION

Feature selection is a very important problem in QSPR/QSAR modeling because of *i*) the large number of features, *ii*) the relatively small number of data and *iii*) the high inter-correlation amongst features [21]. The objective of feature selection is three-fold: improving the performance of the predictors/classifiers, providing faster and more cost-effective predictions/classifications, and providing a better understanding of the underlying process that generated the data [35].

Feature selection algorithms can be classified into two categories based on whether or not feature selection is done independently of the learning algorithm used to construct the predictor. If feature selection is done independent of the learning algorithm, the technique is said to follow a *filter* approach. Otherwise it is said to follow a *wrapper* approach [36]. While the filter approach is generally computationally more efficient than the wrapper approach, its major drawback is that an optimal selection of features may not be independent of the inductive and representational biases of the learning algorithm to be used to construct the predictor. The wrapper approach, on the other hand involves the computational overhead of evaluating candidate feature subsets by executing a selected learning algorithm on the data set represented using each feature subset under consideration [37].

³<http://www.tripos.com/>

⁴The TI is the ratio of the concentration of a compound that gives undesirable effects to that which gives desirable effects.

Feature subset selection, in the context of FNN prediction, presents an instance of a multi-criteria optimization problem. The multiple criteria to be optimized, described by a fitness function f , include the accuracy of prediction and the compactness of the FNN model (the number of hidden nodes). Evolutionary algorithms offer an attractive approach to multi-criteria optimization problems. Since the early 1990's, evolutionary computation - in particular GAs - have been increasingly used in a variety of global optimization problems in chemistry [38]. GAs have been applied both for filter and wrapper feature selection in computational chemistry [21], [37], [39].

In our approach, we use the GA feature selection idea introduced by Siedlecki and Sklansky [40]. In their work, a GA is used to find an optimal binary vector, where each bit is associated with a feature. If the i -th bit of this vector equals 1, then the i -th feature is allowed to participate in prediction; if the bit is 0, then the corresponding feature does not participate. Each resulting subset of features is evaluated according to the fitness function f on a set of testing data, using a FNN and cross-validation.

The method we use can be thought as a hybrid filter-wrapper method, since during feature selection we perform a partial learning of the FNN model. In the second stage, the FNN will be further fine-tuned by another GA. Our strategy follows the principles described by Chakraborty and Pal (neuro-fuzzy learning + feature selection + subsequent network pruning) [41].

IV. THE GA-OPTIMIZED FUZZY NEURAL NETWORK MODEL

The FNN was implemented according to a modification of the Min-Max Fuzzy Inference Network (MMFIN) described in [42]. This modification consists of replacing the $d - y$ error of the MMFIN learning algorithm, with the symmetric relative error, $(d - y)/(d + y)$. This incorporates the fact that small differences of low IC_{50} values are chemically more significant than the same amount of difference of high IC_{50} values. Note that all d and y are positive in our FNN. The network consists of three layers:

- 1) The vector of input values.
- 2) The hidden layer, each node of which represents a prototype of the input vectors; the number of these neurons is determined by the training algorithm, with the maximum number possible being the number of ligands used for training.
- 3) The output layer, consisting of one neuron which predicts the IC_{50} value for the molecule.

The first and second layers are connected by fuzzy membership functions. The fuzzy membership value of each input molecular descriptor is calculated, in relation to previously learned input vectors, and stored in a matrix. The first dimension represents the descriptors; the second dimension represents the hidden layer neurons. Thus, for each molecule passing through the network, the membership value of each descriptor in each of the previously learned prototypes is determined.

The second and third layers are connected by a weight matrix, w_2 . The first dimension of this weight matrix represents the hidden layer neurons; the second dimension represents the output neuron. The output of the hidden layer is the minimum (fuzzy AND) of the membership function input. This hidden layer output is then multiplied by w_2 . The resultant matrix serves as input to the third layer of the network. The output of the third layer of the network is the maximum (fuzzy OR) of this input, and is the computed IC_{50} value of the input vector of descriptors.

During the training process, the final output is compared to the target value, which is the known IC_{50} value of the training ligand. If the output differs from this known value by more than the established error tolerance, the membership functions and w_2 weight matrix are adjusted and this ligand must be "re-learned." If it is not possible to "re-learn" this molecule so that it falls within the ranges of the already established prototypes (hidden layer neurons), then a new neuron is added to the hidden layer. Training continues until all training molecules are learned and produce output values within the acceptable error tolerance. During training, the calculated output is compared to the target value plus or minus the error tolerance. A high error tolerance thus results in a very compact structure, which has a low predictive ability, but generalizes well. A low error tolerance results in a network with as many nodes as input vectors and is overfitted. Therefore, the generalization degree of the network is controlled by the error tolerance. In this manner, the FNN quickly learns all training patterns.

Since the FNN architecture is determined only by the error tolerance, the challenge is how to find the best value for this parameter. The goal of optimizing the structure of the FNN with a specified subset of descriptors is to produce the most compact network possible, which still maintains a high prediction and generalization ability. These optimization criteria are measured by the fitness function f (the function mentioned in Section III), which incorporates both the degree of compactness (the number of hidden nodes) and the predictive ability of the FNN. In our approach, we use a GA to optimize the parameter of the FNN. The objective of the GA optimization is to find the optimum balance between the extremes of a highly compact network which is unable to predict accurately, and a very loose, overfitted network which is not able to generalize. In our implementation, we use the same GA in two different instances: to determine the optimal subset of features and to determine the optimal error tolerance for the FNN.

V. A MORE DETAILED DESCRIPTION OF OUR METHOD

The prediction accuracy of the model can be expressed by several statistics [43]. We have computed statistics directly on the predicted IC_{50} values, and not on their log transform, since this proved to give better results in the case of FNN prediction [2].

A commonly used measure is the RMSE (Root Mean Squared Error). Considering the observation in Section IV, a

relative error [44] seems to be also appropriate. We used the Symmetric Mean Absolute Percentage Error (sMAPE) [43]:

$$200/K * \sum |d - y|/(d + y),$$

where K is the number of training samples. We also used r , the Pearson product-moment coefficient of correlation [45]; for an accurate prediction, one would expect r as close to 1 as possible.

Our GA uses two types of chromosomes:

- An fChromosome, in which each gene is either a 0 or a 1; this chromosome defines the subset of features with which to train the FNN.
- An eChromosome, in which each gene is an integer; all genes are in the range [0, 9], except the first gene which is [8, 9]. This chromosome is converted to a float which is used as the error tolerance to train the FNN. The resultant error tolerance therefore falls within the range [0.8, 1.0). The minimum value of 0.8 was selected following numerous sequential optimizations; it was noted that when the error tolerance is less than 0.8, the number of hidden nodes is too high, resulting in overfitting.

We use the following crossover/mutation operators:

- *Crossover*: The population consists of 70 individuals. Select thirty-four individuals for crossover using the roulette wheel method. For each pair, randomly determine the point of crossover. Exchange the genes beyond the crossover point.
- *Mutation*: Select seven individuals using the roulette wheel method. For each individual to be mutated, randomly determine the point of mutation. Perform point mutation: *i*) in the case of Feature Selection, in which each gene is either a 0 or a 1, flip the gene; *ii*) in the case of Error Tolerance optimization, in which each gene is an integer, randomly select a different integer, within the correct range, for the mutated gene.

Our method can be described by the following four stages:

1) *Feature Selection Optimization*: An initial population of fChromosomes is instantiated. If a previous optimization of the process has been conducted, the fittest individual from that prior run is used to seed this initial population. To obtain an estimate of the fitness of each member of this population, 200 eChromosomes are randomly generated for each fChromosome. Ten-fold cross-validation is used to train and test the FNN with each eChromosome for each fChromosome. During this cross-validation, the sMAPE, RMSE, and Pearson's r are calculated, and the average number of hidden nodes (aveM) determined. The fitness is evaluated according to the following formula:

$$f = \frac{1}{sMAPE + 2 \cdot aveM} + 0.01 \cdot r \quad (1)$$

The fitness value ultimately assigned to each fChromosome is the fittest value obtained from evaluating all 200 eChromosomes with this particular subset of descriptors.

Following optimization of the subset of features, the FNN trained with this fittest fChromosome is fine-tuned by using a second GA to fully optimize the error tolerance.

2) *Error Tolerance Optimization*: The subset of features used is defined by the fittest fChromosome obtained from the feature selection optimization.

An initial population of eChromosomes is generated. It is seeded by the eChromosome obtained during the estimate of the fittest eChromosome during the feature selection optimization. The leave-one-out (LOO) cross-validation method is used to calculate the sMAPE, RMSE, and Pearson's r , and to determine the aveM. The fitness value of each eChromosome is calculated according to eq. (1).

3) *Inferring IC_{50} values of novel compounds*: The optimized FNN trained with the optimized subset of descriptors is used to infer IC_{50} values of the novel inhibitors.

4) *Rule Extraction*: Following optimization of the FNN trained with the optimum subset of descriptors, the fuzzy IF/THEN rules are extracted from each of the hidden nodes. Each hidden node corresponds to a learned prototype as well as to a fuzzy IF/THEN rule. This rule is accessible at the end of the learning phase.

VI. METHODOLOGY AND RESULTS

From a theoretical point of view, no single prediction method can be designated as the "best". The "best" model will have to be related to the purpose of predicting — its value for improving decision making and the needs of the person using the prediction [44]. Comparing different prediction models is a difficult problem [46] and we must judge the appropriateness of whichever measure we use by how effectively it provides information about future predictions. When estimating the quality of the prediction model, both the generalization ability and the prediction accuracy play an important role. One is interested not only in how accurately the model approximates the learning data, but also how the model generalizes on new data. We used LOO and ten-fold cross-validations to estimate the model's generalization capability with respect to the training set. LOO cross-validation is especially useful when dealing with relatively small training sets [46].

We generated two models. One is GA-optimized, but without feature selection (GA-FNN). The second is GA-optimized and has in addition a GA-optimized subset of features (FS-GA-FNN). The GA-FNN model uses all 30 molecular descriptors in Table I. The FS-GA-FNN model uses only the following 13 molecular descriptors from Table I: (D6-D9, D12-D13, D20, D25-30).

The GA for feature selection in the FS-FNN model is initiated with a population of randomly generated individuals. The fitness of each individual is evaluated using ten-fold cross-validation. The chromosomes are fChromosomes. The program ran for 60 generations. However, the global fittest individual was identified during the second generation.

The GA used to further optimize the architecture of the FS-FNN is set up similarly to that used for feature selection, except that individuals are evaluated by using LOO

TABLE II
 IC_{50} PREDICTION PERFORMANCE ANALYSIS USING LOO

	GA-FNN	FS-GA-FNN
RMSE	1216	968
sMAPE	99.9	83.6
Pearson's r	0.77	0.89
Number of Hidden Nodes	15	10

cross-validation, in order to obtain more accurate statistical analysis. In this case, the chromosomes, which are eChromosomes, define the level of error tolerance used to train the network, rather than the subset of descriptors. This phase of optimization runs for 19 generations.

To assess the effect of feature selection optimization, this FS-GA-FNN is compared to the GA-FNN optimized by the GA in exactly the same way. The statistics for both are displayed in Table II. The LOO cross-validation results, with and without feature selection, are shown in Figures 2 and 3, respectively. Our results indicate that using an optimal subset of features results in better predictions with respect to all three statistics, and generates fewer hidden nodes.

The predictions for the novel inhibitors are made by the FS-GA-FNN (see Table III). We extracted a fuzzy IF/THEN rule from each hidden node of the FS-GA-FNN. Details can be found in [2]. One of the rules is shown in Figure 1. These rules can be used to analyze the contribution of each descriptor in predicting IC_{50} values. It is important to note that we do not have rule proliferation, since the number of generated rules is controlled by the fitness function of the GA according to eq. (1).

The enormous search space makes some of the experiments run for hours or days. This happens especially during feature selection. There are faster, but less accurate, feature selection techniques that could be used. We preferred this slower solution because *i*) it provides nearly the global optimum and *ii*) the training sets in computational chemistry are usually small, because of the high cost required to obtain them. In drug design, obtaining an accurate prediction is more important than execution time since the savings obtained in actual synthesis of compounds are substantial.

The FNN learning algorithm stops after it has learned completely the whole training sequence. For "conflicting" cases, i.e. identical training vectors with different target values, the algorithm does not converge. Reducing the set of features increases chances for conflicting cases. There are several possibilities to override this aspect. The simplest solution is to assign very low f values to the conflicting cases. This does not affect the learning process, since it is independent of the statistic aspect of the training sequence: the frequency of the input vectors does not count.

VII. CONCLUSIONS

Essentially, our approach consists of GA optimizations of the first two FNN layers. The optimal subset of chemical descriptors, as well as the extracted rules are valuable material for further chemical interpretations. Feature selection

aids in determining physico-chemical properties that are most important for obtaining lead chemical compounds with high biological affinities toward therapeutically important enzymes. Our feature selection + prediction + rule extraction method could greatly improve the efficiency of the drug design process, and consequently reduce the amount of failures in the last stages of drug development.

ACKNOWLEDGMENTS

The authors thank the Science Honors Program, the Graduate Studies and Research Office, the College of the Sciences, and the Office of Undergraduate Research at Central Washington University for partial support.

REFERENCES

- [1] D. Weekes and G. B. Fogel, "Evolutionary optimization, backpropagation, and data preparation issues in QSAR modeling of HIV inhibition by HEPT derivatives," *BioSystems*, vol. 72, pp. 149–158, 2003.
- [2] R. Andonie, L. Fabry-Asztalos, C. Collar, S. Abdul-Wahid, and N. Salim, "Neuro-fuzzy prediction of biological activity and rule extraction for HIV-1 protease inhibitors," in *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'05)*, 2005, pp. 113–120.
- [3] I. V. Tetko, V. Y. Tanchuk, and A. I. Luiik, "Evaluation of potential HIV-1 reverse transcriptase inhibitors by artificial neural networks," in *Proceedings of the Seventh Annual IEEE Symposium on Computer-Based Medical Systems*, 1994, pp. 311–316.
- [4] I. V. Tetko and V. Y. Tanchuk, "Application of associative neural networks for prediction of lipophilicity in ALOPS 2.1 program," *J. Chem. Inf. Comput. Sci.*, vol. 42, pp. 1136–1145, 2002.
- [5] T. Niwa, "Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures," *J. Chem. Inf. Comput. Sci.*, vol. 43, pp. 113–119, 2003.
- [6] Z. R. Yang and R. Thomson, "Bio-basis function neural network for prediction of protease cleavage sites in proteins," *IEEE Transactions on Neural Networks*, vol. 16, pp. 263–274, 2005.
- [7] I. V. Tetko, A. I. Luiik, and G. I. Poda, "Application of neural networks in structure-activity relationships of a small number of molecules," *J. Med. Chem.*, vol. 36, pp. 811–814, 1993.
- [8] J. Devillers, "Designing molecules with specific properties from inter-communicating hybrid systems," *J. Chem. Inf. Comput. Sci.*, vol. 36, pp. 1061–1066, 1996.
- [9] T. Niwa, "Prediction of biological targets using probabilistic neural networks and atom-type descriptors," *J. Med. Chem.*, vol. 47, pp. 2645–2650, 2004.
- [10] P. Potocnik, I. Grabec, M. Setinc, and J. Levec, "Hybrid modeling of kinetics for methanol synthesis," in *Soft Computing Approaches in Chemistry*, H. Cartwright and L. M. Sztandera, Eds. Heidelberg: Springer-Verlag, 2000.
- [11] A. M. Bianucci, A. Micheli, A. Sperduti, and A. Starita, "Application of cascade correlation networks for structures to chemistry," *Applied Intelligence*, vol. 12, pp. 117–147, 2000.
- [12] D. Yaffe, Y. Cohen, G. Espinosa, A. Arenas, and F. Giralt, "Fuzzy ARTMAP and back-propagation neural networks based quantitative structure-property (QSPRs) for octanol-water partition coefficient of organic compounds," *J. Chem. Inf. Sci.*, vol. 42, pp. 162–183, 2002.
- [13] G. Espinosa, D. Yaffe, A. Arenas, C. Y., and G. F., "A fuzzy ARTMAP-based qualitative structure-property relationship (QSPR) for predicting physical properties of organic compounds," *J. Chem. Inf. Sci.*, vol. 40, pp. 2757–2766, 2001.
- [14] G. Espinosa, A. Arenas, and F. Giralt, "An integrated som-fuzzy ARTMAP neural system for the evaluation of toxicity," *J. Chem. Inf. Sci.*, vol. 42, pp. 343–359, 2002.
- [15] S. Draghici and R. B. Potter, "Predicting HIV drug resistance with neural networks," *Bioinformatics*, vol. 19, pp. 98–107, 2003.
- [16] J. Paetz, "Evolutionary optimization of interval rules for drug design," in *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBC'04)*, 2004, pp. 238–243.

- [17] X. J. Yao, A. Panaye, J. P. Doucet, R. S. Zhang, H. F. Chen, M. C. Liu, Z. D. Hu, and B. T. Fan, "Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression," *J. Chem. Inf. Comput. Sci.*, vol. 44, pp. 1257–1266, 2004.
- [18] R. Todeschini, V. Consonni, R. Mannhold, H. Kubinyi, and H. Timmerman, *Handbook of Molecular Descriptors*. Wiley-VCH, 2000.
- [19] A. M. Bianucci, A. Micheli, A. Sperduti, and A. Starita, "A novel approach to QSPR/QSAR based on neural networks for structures," in *Soft Computing Approaches in Chemistry*, H. Cartwright and L. M. Sztandera, Eds. Heidelberg: Springer-Verlag, 2000.
- [20] A. Micheli, F. Portera, and A. Sperduti, "A preliminary experimental comparison of recursive neural networks and a tree kernel method for qsar/qspr regression tasks," *Neurocomputing*, in press.
- [21] M. Ozdemir, M. J. Embrechts, F. Arciniegas, C. M. Breneman, L. Lockwood, and K. P. Bennett, "Feature selection for in-silico drug design using genetic algorithms and neural networks," in *Proceedings of the IEEE Mountain Workshop on Soft Computing in Industrial Applications*, 2001, pp. 53–57.
- [22] J. Gasteiger, A. Teckentrup, L. Terflath, and S. Spycher, "Neural networks as data mining tools in drug design," *J. Phys. Org. Chem.*, vol. 16, pp. 232–245, 2003.
- [23] C. D. Neagu, E. Benfenati, G. Gini, P. Mazzatorta, and A. Roncaglioni, "Neuro-fuzzy knowledge representation for toxicity prediction of organic compounds," in *Proceedings of the 15th European Conference on Artificial Intelligence, ECAI'2002*. IOS Press, 2002, pp. 498–502.
- [24] A. Wlodawer, "Structure-based inhibitors of HIV-1 protease," *Annu. Rev. Biochem.*, vol. 62, pp. 543–585, 1993.
- [25] A. Wlodawer and J. Vondrasek, "Inhibitors of HIV-1 protease," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 27, pp. 249–284, 1998.
- [26] D. Leung, G. Abbenante, and D. P. Fairlie, "Protease inhibitors: Current status and future prospects," *J. Med. Chem.*, vol. 43, pp. 305–341, 2000.
- [27] D. H. Rich, "Comprehensive medicinal chemistry," *Chem. Rev.*, vol. 2, pp. 391–441, 1990.
- [28] A. C. Nair, P. Jayatilike, X. Wang, S. Miertus, and W. J. Welsh, "Computational studies on tetrahydropyrimidine-2-one HIV-1 protease inhibitors: improving three-dimensional quantitative structure-activity relationship comparative molecular field analysis models by inclusion of calculated inhibitor- and receptor-based properties," *J. Med. Chem.*, vol. 45, pp. 973–983, 2002.
- [29] R. S. Bohacek and C. J. McMartin, "Multiple highly diverse structures complementary to enzyme binding sites: Results of extensive application of a de novo design method incorporating combinatorial growth," *Am. Chem. Soc.*, vol. 116, pp. 5560–5571, 1994.
- [30] W. Jorgensen, "The many roles of computation in drug discovery," *Science*, vol. 303, pp. 1813–1818, 2004.
- [31] F. E. Boyer, J. V. Vara Prasad, J. M. Domagala, E. L. Ellsworth, C. Gajda, S. E. Hagen, L. J. Markoski, B. D. Tait, E. A. Lunney, A. Palovsky, D. Ferguson, N. Graham, T. Holler, D. Hupe, C. Nouhan, P. J. Tummino, A. Urumov, E. Zeikus, G. Zeikus, S. J. Gracheck, J. M. Sanders, S. VanderRoest, J. Brodfuehrer, K. Iyer, M. Sinz, and S. V. Gulnik, "5,6-dihydropyran-2-ones possessing various sulfonyl functionalities: potent nonpeptidic inhibitors of HIV protease," *J. Med. Chem.*, vol. 43, pp. 843–858, 2000.
- [32] M. P. Glenn, L. K. Pattenden, R. C. Reid, D. P. Tyssen, J. D. Tyndall, C. J. Birch, and D. P. Fairlie, "Beta-strand mimicking macrocyclic amino acids: templates for protease inhibitors with antiviral activity," *J. Med. Chem.*, vol. 45, pp. 371–381, 2002.
- [33] D. Scholz, A. Billich, B. Charpiot, P. Ettmayer, P. Lehr, B. Rosenwirth, E. Schreiner, and H. Gstach, "Inhibitors of HIV-1 proteinase containing 2-heterosubstituted 4-amino-3-hydroxy-5-phenylpentanoic acid: synthesis, enzyme inhibition, and antiviral activity," *J. Med. Chem.*, vol. 37, pp. 3079–3089, 1994.
- [34] S. E. Hagen, J. V. Prasad, F. E. Boyer, J. M. Domagala, E. L. Ellsworth, C. Gajda, H. W. Hamilton, L. J. Markoski, B. A. Steinbaugh, B. D. Tait, E. A. Lunney, P. J. Tummino, D. Ferguson, D. Hupe, C. Nouhan, S. J. Gracheck, J. M. Saunders, and S. VanderRoest, "Synthesis of 5,6-dihydro-4-hydroxy-2-pyrones as HIV-1 protease inhibitors: the profound effect of polarity on antiviral activity," *J. Med. Chem.*, vol. 40, pp. 3707–3711, 1997.
- [35] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
- [36] G. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Machine Learning: Proceedings of The Eleventh International Conference*, Morgan Kaufman, 1994, pp. 121–129.
- [37] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intelligent Systems*, vol. 13, pp. 44–49, 1998.
- [38] R. Johnston, Ed., *Applications of Evolutionary Computation in Chemistry*. Springer-Verlag, Berlin, 2004.
- [39] M. L. Raymer, P. W. F. E. D. Goodman, L. A. Kuhn, and J. A. K., "Dimensionality reduction using genetic algorithms," *IEEE Trans. on Evolutionary Computation*, vol. 4, pp. 164–171, 2000.
- [40] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognition Letters*, vol. 10, pp. 335–347, 1989.
- [41] D. Chakraborty and N. Pal, "A neuro-fuzzy scheme for simultaneous feature selection and fuzzy rule-based classification," *IEEE Trans. on Neural Networks*, vol. 15, pp. 110–123, 2004.
- [42] L. Y. Cai and H. K. Kwan, "Fuzzy classifications using fuzzy inference networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 28, pp. 334–347, 1998.
- [43] R. J. Hyndman and A. B. Hoehler, "Another look at measures of forecast accuracy," Monash University, Department of Econometrics and Business Statistics, Tech. Rep. 13/05, May 2005.
- [44] S. Makridakis, "Accuracy measures: theoretical and practical concerns," *International Journal of Forecasting*, vol. 9, pp. 527–529, 1993.
- [45] F. E. Fischer, *Fundamental Statistical Concepts*. Canfield Press, 1999.
- [46] D. M. Hawkins, "The problem of overfitting," *J. Chem. Inf. Sci.*, vol. 44, pp. 1–12, 2004.

IF...	LOW	LOW-MEDIUM	MEDIUM	MEDIUM-HIGH	HIGH
Number of Hydrogen Atoms					
Number of Nitrogen Atoms					
Number of Oxygen Atoms					
Number of Sulfur Atoms					
Total Number Bonds					
Number of Single Bonds					
Bond Stretching Energy					
Van der Waals Energy					
One to Four Electrostatic Energy					
Electrostatic Energy					
Van der Waals Electrostatic Pairs					
One to Four Van der Waals Electrostatic Pairs					
Scaled Van der Waals Electrostatic Pairs					
THEN: IC_{50} Value					

Fig. 1. Extracted rule.

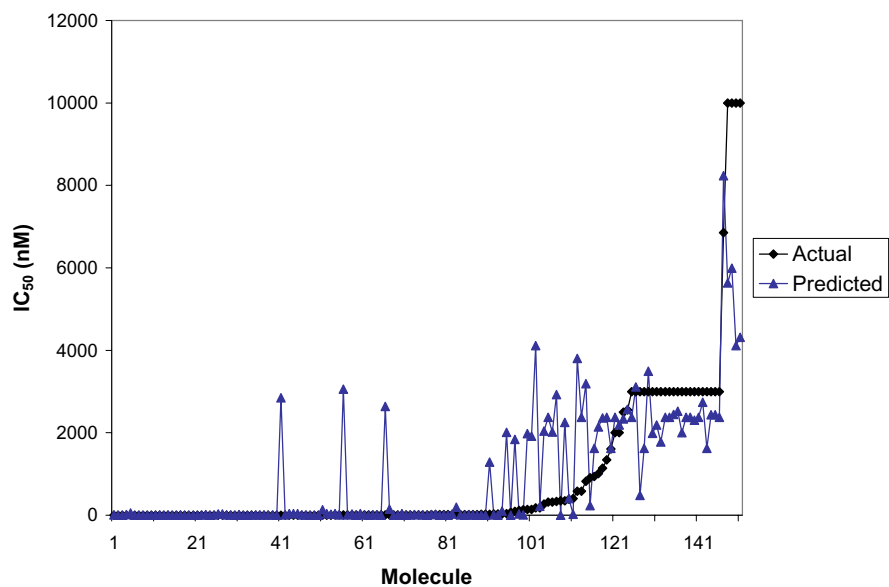


Fig. 2. LOO cross-validation results using all features.

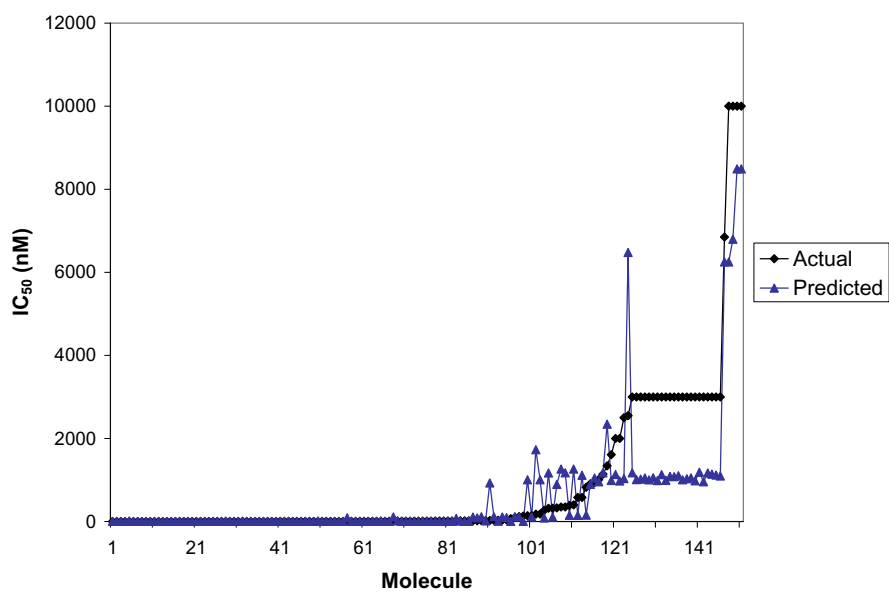
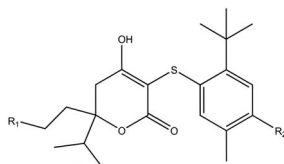
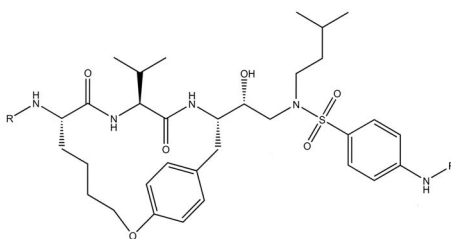


Fig. 3. LOO cross-validation results using the optimal subset of features.

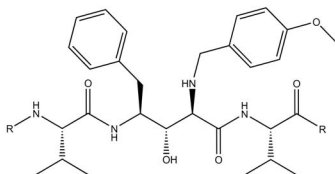
TABLE III
 NOVEL INHIBITORY STRUCTURES WITH THEIR CORRESPONDING PREDICTED IC_{50} VALUES USING THE OPTIMAL SUBSET OF FEATURES AND THE OPTIMIZED FNN



no.	R_1	R_2	Predicted IC_{50} (nM)
1	OCH_3	<i>Cbz</i>	3.5
2	<i>Cbz</i>	OCH_3	3.5
3	<i>Cbz</i>	COH	3.5
4	<i>Cbz</i>	NH_2	4.81
5	<i>Cbz</i>	$OSO_2Ph(4 - CN)$	106.02
6	OCH_3	COH	0.68
7	OCH_3	NH_2	0.65
8	OCH_3	$OSO_2Ph(4 - CN)$	4.67
9	$Ph - OH$	COH	0.71
10	<i>Furan</i>	NH_2	0.72
11	<i>Furan</i>	$OSO_2Ph(4 - CN)$	4.67
12	<i>Thiophene</i>	NH_2	0.73
13	<i>Thiophene</i>	$OSO_2Ph(4 - CN)$	11.99
14	4 - methylthiazole	NH_2	0.71
15	4 - methylthiazole	$OSO_2Ph(4 - CN)$	11.68



no.	R	Predicted IC_{50} (nM)
16	OCH_3	130.46
17	COH	125.27
18	<i>Cbz</i>	7000.0
19	<i>Furan</i>	127.57
20	<i>Thiophene</i>	59.0
21	4 - methylthiazole	59.0



no.	R	Predicted IC_{50} (nM)
22	OCH_3	90.67
23	COH	73.26
24	<i>Cbz</i>	837.5
25	<i>Furan</i>	87.53
26	<i>Thiophene</i>	73.79