Neuro-fuzzy Prediction of Biological Activity and Rule Extraction for HIV-1 Protease Inhibitors

Răzvan Andonie*, Levente Fabry-Asztalos†, Catharine J. Collar†, Sarah Abdul-Wahid* 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*— A fuzzy neural network (FNN) and multiple linear regression (MLR) were used to predict biological activities of 26 newly designed HIV-1 protease potential inhibitory compounds. Molecular descriptors of 151 known inhibitors were used to train and test the FNN and to develop MLR models. The predictive ability of these two models was investigated and compared. We found the predictive ability of the FNN to be generally superior to that of MLR. The fuzzy IF/THEN rules were extracted from the trained network. These rules map chemical structure descriptors to predicted inhibitory values. The obtained rules can be used to analyze the influence of descriptors. Our results indicate that FNN and fuzzy IF/THEN rules are powerful modeling tools for QSAR studies.**

I. Introduction

The possibility of relating some significant aspects of molecular structures to any particular behavior of a selected class of chemical compounds offers a significant challenge in many fields of research such as the prediction of physico-chemical properties, chemical reactivity, or biological activity of molecules. In the case of predicting physical properties, we speak of Quantitative Structure-Property Relationship (QSPR). When predicting the biological activity of chemical compounds, we speak of Quantitative Structure-Activity Relationship (QSAR).

Neural models have been used to predict properties of chemical compounds such as inhibition of HIV-1 reverse transcriptase [1], lipophilicity and aqueous solubility [2], intestinal absorption [3], and site of protease cleavage [4]. Several neural architectures were successful for such tasks; among these are backpropagation [1], [5], [6], associative neural networks [2], probabilistic neural networks [7], generalized regression neural networks [8], radial basis function networks [4], [8], cascade correlation [9], neural networks trained via evolutionary algorithms [10], fuzzy ARTMAP [11]–[13], and support vector machines [14].

A major problem in QSPR/QSAR modeling is represented by the need to find a set of complete and relevant molecular descriptors. A good structure representation should have different code for each 3D structure (uniqueness) and the same number of variables for all structures. It should also be reversible, translational and rotational invariant. Fulfilling all these requirements is a difficult task which has not yet been solved satisfactorily [15], [16]. 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 [17].

In QSPR/QSAR neural networks modeling, two different methods have been proposed for the definition of descriptors [9], [18]: *i)* using known physico-chemical properties; and *ii)* using topological indices, which code specific morphological properties of the molecular graph.

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 it non-reversible. The number of vector features may be very large and many of them may be inter-correlated or redundant.

Another option is to process directly the chemical graph structures. This approach was followed by Bianucci *et al.* [9], [18], [19]. It is worth noting that their molecular representation is not for all possible chemical structures. A major problem for graph representations is the computational complexity of comparing two labeled graphs [20] (i.e., two chemical structures). Non-standard similarity measures have to be used [21].

The numerical encoding of chemical structures must be devised by experts. However, once done, the information can be stored in a database which can then be used easily by individuals possessing less expertise. There are several of these databases accessible on-line [22]. We have employed the `mol2` molecular file format. This format is based on a vector of topological indices. It is a unique, reversible, and general representation. These data can be fed easily into existent computational chemistry tools and into machine learning algorithms.

Our primary goal is to investigate the use of a feed-forward neuro-fuzzy model for biological activity (IC_{50})

0-7803-9387-2/05/$20.00 ©2005 IEEE

prediction[1]. The integration of neural and fuzzy systems leads to a symbolic relationship in which fuzzy systems provide a powerful framework for expert knowledge representation, while neural networks provide learning capabilities. Actually, there is some equivalence between fuzzy rule-based systems and neural networks [23]. Besides the flexibility to deal with noisy and incomplete data, many of these neural models have the capability to easily expose the learned knowledge in the form of fuzzy IF/THEN rules. The final goal would be to explain to a drug designer, in a human-comprehensible form, how the network arrives at a particular decision, and to provide him/her with insight into the influence of the input features on the predicted target. Few authors have addressed this problem in computational chemistry. Neagu *et al.* [24], [25] have used a FNN, trained by backpropagation, to extract fuzzy IF/THEN rules for toxicity prediction of organic compounds. Other neuro-fuzzy systems with rule extraction capability were employed in computational chemistry by Loukas [26], and by Cundari and Russo [27].

In Section II we describe the importance of predicting HIV-1 protease inhibitors and their biological affinity. Section III describes the chemical descriptors and data we have used, in the framework of SYBYL[2] software. In Section IV we refer to the FNN employed. Section V presents the results we have obtained in MLR and FNN prediction, and in rule extraction. Section VI concludes with closing remarks.

II. IC_{50} Prediction of HIV-1 Protease Inhibitors

Many debilitating disease states, such as HIV/AIDS, Alzheimer's, some cancers, and malaria, are caused by misfunction of endogenous enzymes, or invasion by foreign enzymes [28]–[31]. The current clinical approach for the therapy of HIV/AIDS utilizes the co-administration of two reverse transcriptase inhibitors with one protease inhibitor (usually referred to as combination therapy). Combination therapy reduces viremia to very low levels; however, in 30-50% of patients, antiviral therapy is ineffective due to resistance development. Furthermore, in many patients, side effects associated with these drugs pose serious problems. Due to current drug resistance and toxicity, there is an urgent need for the development of more efficient drugs with different resistance profiles, decreased toxicity and more than one type of inhibitory action.

An essential component of the drug discovery process is computer aided drug design, which results in the prediction of lead chemical compounds[3]. Using molecular modeling, neuro-fuzzy techniques, and mathematical and statistical programs, de novo structures with steric and chemical complementarity to an enzyme active site are generated [32]–[34]. As a result, the otherwise lengthy drug design, synthesis, and evaluation process is shortened and the efficiency of drug discovery is improved. Thus, new methods for designing enzyme inhibitors, and for predicting their properties, are expected to have great value in drug discovery.

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

[2]http://www.tripos.com/

[3]The lead is a prototype compound that has some desirable property that is likely to be therapeutically useful.

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

III. Data Description

Molecular descriptors were extracted from chemical structures using logical methodology and mathematical procedures. Data files (`mol2` and `txt`) were created using molecular modeling software SYBYL. We designed a software to extract 30 molecular descriptors representing HIV-1 protease inhibitors. The molecular descriptors used in this study are shown in Table I.

Selection of molecular descriptors was determined by contribution to molecular entity. The summation of atoms and bonds (D1 and D12) along with the number of each type of atom (D2-D9), bond (D13-D15), and functional group (D16 and D17) provided atomic descriptors for each inhibitory structure. It has been shown that there is a correlation between molecular weight and biological affinity of inhibitory compounds [35]–[38]. Thus, molecular weights (D18) were also included. Index of Hydrogen Deficiency (D11) was incorporated to describe the degrees of unsaturation for each molecule.

Molecular volume (D10) was calculated to represent the overall three-dimensional spaces that molecules occupy. Total charge (D19), molecular energy descriptors (D20-D27), and summations of molecular interactions (D28-D30) established the conformational molecular relationships for each inhibitory structure.

For each descriptor, the minimum value and maximum value of all molecules, (training/test molecules and novel inhibitors) were determined. These values were used to calculate the normalized value according to the formula:

$$(descriptor - min)/(max - min)$$

Since all molecules in our datasets were used to determine the minimum and maximum values, no descriptor value fell outside the range [0, 1].

We used a data set of 151 molecules with known IC_{50} values for training/validation, and a set of 26 newly designed compounds, which have not yet been synthesized, for prediction. The molecules with known IC_{50} values were previously synthesized, and their actual biological activities were determined in the laboratory [39]–[42]. The novel compounds were designed using statistically significant chemical and biological data. The basic structural features of these compounds are identical to training/testing molecules, while substituents (R groups) were determined from the most reoccurring substituents of molecules with low IC_{50} and high therapeutic index (TI) values[4]. The newly designed compounds were docked inside HIV-1 protease. Substituents of novel compounds bind into the same pockets within the active site of the protease as the substituents of the training/validation molecules.

IV. The Fuzzy Neural Network Model

The FNN was implemented according to the Min-Max Fuzzy Inference Network (MMFIN) described in [43]. It has a feed-forward architecture, consisting of three layers. Each layer has, respectively, the following types of fuzzy neurons: Transit, Minimum, and Maximum Neurons. Since we predict only one target (the IC_{50} value), we have used a single output node (a Maximum Neuron). At each iteration of the learning phase, the error $d-y$ (d is the target value and y is the actual output value of the network) is compared to E (the output error tolerance — a parameter of the training process). According to the result of this comparison, the fuzzy membership functions and the number of hidden nodes (the Minimum Neurons) are adjusted. The network architecture was adapted automatically during the learning procedure. The number of hidden nodes and the generalization degree of the network were controlled by E.

The error $d-y$ treats all errors of the same magnitude difference equally, regardless of the magnitude of the individual value being predicted. For example, $2-1$ and $1001-1000$ are equal from this point of view. In the case of predicting IC_{50} values, $2-1$ possesses greater chemical significance than $1001-1000$. This is because within the range of [0,..., 100], small differences are chemically more important than larger differences in the [100,..., 10000] range. For this reason, we replaced the $d-y$ error with the symmetric relative error (sRE) $(d-y)/(d+y)$ in the MMFIN learning algorithm from [43]. Note that all d and y are positive in our FNN.

Each hidden node corresponds to a learned prototype as well as to a fuzzy IF/THEN rule. This rule is easy to extract at the end of the learning phase. The MMFIN can learn quickly, in supervised mode, with all outputs of the training patterns within the error tolerance limits.

V. Results

A. Prediction

Several statistics can be used to compare the prediction accuracy [44]. A commonly used measure is the RMSE (Root Mean Squared Error). A relative error seems to be also appropriate, considering the observation in Section IV. According to [45], the Mean Absolute Percentage Error (MAPE) is a relative measure that incorporates the best characteristics among the various accuracy criteria. We used the Symmetric Mean Absolute Percentage Error (sMAPE) [44]:

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

where K is the number of training samples.

We also used r, the Pearson product-moment coefficient of correlation [46]. For an accurate prediction, one would expect r as close to 1 as possible.

1) MLR Prediction: One of the most familiar standard approaches to QSAR is based on MLR. However, this approach can capture only linear relationships between molecular characteristics and structural or functional features to be predicted. Generally, IC_{50} values violate multiple linear regression assumptions and better results are obtained when using a logarithmic transformation [47]. In our study, the reported IC_{50} values range from 0.1 to 10000. We measured the biological activity by $log(1+IC_{50})$ in order to obtain only positive values.

In MLR if the p-value[5] of a coefficient is less than the chosen-level, there is evidence of a significant relationship between independent and dependent variables. In our study, the independent variables are molecular descriptors and the dependent variables are the actual IC_{50} values. Molecular descriptors consisting of the highest p-values were eliminated, one at a time, from the regression model. The result of this process is the optimal MLR model at a 95% confidence interval. We used the data obtained from "leave-one-out" cross-validation (CV) to compute the RMSE, sMAPE, and Pearson's statistics.

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

[5]The p-value measures consistency by calculating the probability of observing the results from the data sample, assuming the null hypothesis is true. The commonly used level is 0.05.

TABLE II

IC_{50} PREDICTION PERFORMANCE ANALYSIS USING CROSS-VALIDATION

	MLR	**sMAPE-FNN**
RMSE	2120.2	1245.7
sMAPE	81.9	91.1
r	0.62	0.76

2) FNN Prediction: We found that for FNN, better results were obtained when using the actual IC_{50} data, and not the corresponding log values. This is in accordance with the known capability of neural networks to recognize highly nonlinear relationships. It is an interesting observation, since other authors have used log values when comparing MLR and neural network prediction of IC_{50} values [48].

We used the FNN to generate an optimized prediction architecture called sMAPE-FNN. The optimal value for parameter E ($E = 0.85999995$) was obtained by minimizing the sMAPE, according to CV. Note that we have replaced the $d - y$ error in the MMFIN learning algorithm from [43] with sRE.

3) Comparison and Discussion: We have computed the RMSE, sMAPE, and r statistics for the MLR and sMAPE-FNN models, for all 151 molecules with known IC_{50} values, using CV. The results are shown in Table II.

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 [45]. Comparing different prediction models is a difficult problem [47] and we must judge the appropriateness of whichever measure we use by how effectively it provides information about future predictions.

Our results indicate that sMAPE-FNN outperforms MLR from the point of view of the RMSE and r statistics. The MLR sMAPE value is better than for the sMAPE-FNN. In order to make the two prediction models comparable, we have computed statistics on the predicted IC_{50} values, and not on their log transform.

The CV results for MLR and sMAPE-FNN predictions are shown in Figures 1 and 2. The predictions from sMAPE-FNN, especially in the lower range of IC_{50} values, are better than the MLR predictions. This is important from the point of view of medicinal chemists for the reason explained above. Another good aspect of sMAPE-FNN prediction is that, in case of a prediction error, it has a tendency to overpredict. This is important considering the high cost of synthesis and biological testing of newly designed molecules; a more "conservative" prediction is better.

The greater flexibility of the FNN to recognize nonlinear relationships is at the same time a curse: the prediction is more dependent on the individual observations, which adds random variability and can lead to worse predictions [47]. This can be seen in the lower range of the predicted IC_{50} values (Figure 2), where some visible outliers appear. This explains why the sMAPE value for the FNN is worse than for the MLR prediction: the sMAPE statistic is sensitive to outliers in the lower range of values.

We tried to use the sMAPE-FNN model with $log(1+IC_{50})$ values. The obtained statistics (RMSE = 2057.8, sMAPE = 111.5, and r = 0.34) are inferior to the ones obtained when using the actual IC_{50} values.

We have also predicted IC_{50} values for the 26 newly designed compounds. The results and chemical structures are shown in Table III. Based on these predictions, we have initiated synthesis on the most promising molecules.

B. Rule extraction

We have generated 10 fuzzy IF/THEN rules from the corresponding 10 hidden nodes of the trained FNN using a reduced set of descriptors. We illustrate with the following extracted rule:

IF:

Total Number of Atoms is *medium* AND
Molecular Volume is *medium* AND
Index of Hydrogen Deficiency is *medium* AND
Molecular Weight is *medium* AND
Total Charge is *low-medium* AND
Bond Stretching Energy is *medium* AND
Angle Bending Energy is *low-medium* AND
Torsional Energy is *medium-high* AND
Electrostatic Energy is *medium-high*

THEN:

IC_{50} is *low*

The [0, 1] descriptor range was divided into five equal partitions, labeled as follows:

1) *low*: [0.0, 0.2]
2) *low-medium*:]0.2, 0.4]
3) *medium*:]0.4, 0.6]
4) *medium-high*:]0.6, 0.8]
5) *high*:]0.8, 1.0]

The IC50 membership values were also partitioned:

1) *low*:]0, 20]
2) *low-medium*:]20, 50]
3) *medium*:]50, 100]
4) *medium-high*:]100, 500]
5) *high*: $>$ 500

These rules can be used to analyze the contribution of each descriptor in predicting IC_{50} values. Based on our analysis of the rules we found that molecules are expected to have a high IC_{50} value if: the number of atoms is medium-high, the molecular volume is medium-high, the Index of Hydrogen Deficiency is medium-high or high, the molecular weight is medium-high, the total charge is medium-high, the angle bending energy is low, the torsional energy is low-medium, and the electrostatic energy is low or low-medium. On the contrary, molecules are expected to have a low IC_{50} value if: the molecular weight is low-medium, the angle bending energy is medium, the torsional energy is low, medium or medium-high, and the electrostatic energy is medium-high. Bond stretching energy displayed medium or medium-high values for both

the high and low IC_{50} predictions, which suggests that it is the least contributing descriptor. From our observation we conclude that these trends play a role in determining IC_{50} values.

VI. Conclusions and Future Work

The experimental results generally show that the sMAPE-FNN performs better than the MLR model. This judgment is based on the accuracy measures we have selected and on the specificity of the IC_{50} prediction problem. Essentially, the sMAPE-FNN is a version of the known MMFIN model. We have optimized its architecture (the number of hidden nodes) by minimizing the sMAPE measure and using the "leave-one-out" cross-validation. This optimizes the generalization capability of the network.

Based on our sMAPE-FNN IC_{50} predictions for the novel compounds, we have initiated synthesis on the most promising molecules. We intend to apply the sMAPE model for predicting other potential enzyme inhibitors.

Results from rule extraction indicate potential for using FNN for the generation of fuzzy IF/THEN rules. Further investigation of this approach is needed in order to optimize the use of the generated information.

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] I. V. Tetko, V. Y. Tanchuk, and A. I. Luik, "Evaluation of potencial 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.

[2] 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.

[3] 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.

[4] 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.

[5] I. V. Tetko, A. I. Luik, and G. I. Poda, "Application of neural networks in tructure-activity relationships of a small number of molecules," *J. Med. Chem.*, vol. 36, pp. 811–814, 1993.

[6] J. Devillers, "Designing molecules with specific properties from intercommunicating hybrid systems," *J. Chem. Inf. Comput. Sci.*, vol. 36, pp. 1061–1066, 1996.

[7] T. Niwa, "Prediction of biological targets using probabilistic neural networks and atom-type descriptors," *J. Med. Chem.*, vol. 47, pp. 2645–2650, 2004.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] J. Zupan and M. Novic, "General type of a uniform and reversible representation of chemical structures," *Analytica Chimica Acta*, vol. 348, pp. 409–418, 1997.

[16] J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design: An Introduction*. Wiley-VCH, 1999.

[17] R. Todeschini, V. Consonni, R. Mannhold, H. Kubinyi, and H. Timmerman, *Handbook of Molecular Descriptors*. Wiley-VCH, 2000.

[18] 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.

[19] 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.

[20] H. Bunke and X. Jiang, "Graph matching and similarity," in *Intelligent Systems and Interfaces*, H. N. Teodorescu, D. Mlynek, A. Kandel, and H. J. Zimmermann, Eds. Kluwer Academic Publishers, 2000.

[21] B. Hammer and T. Villmann, "Classification using non-standard metrics," in *Proceedings of the European Symposium on Artificial Neural Networks (ESANN 2005)*, M. Verleysen, Ed., D-side publications, 2005.

[22] M. C. Hemmer and J. Gasteiger. (2000) Data mining in chemistry, Proceedings of the TERENA Networking Conference 2000. [Online]. Available: http://www.terena.nl/conferences/archive/tnc2000/proceedings/10B/10b5.html

[23] S. Mitra and Y. Hayashi, "Neuro-fuzzy rule generation: Survey in soft computing framework," *IEEE Transactions on Neural Networks*, vol. 11, pp. 748–768, 2000.

[24] C. D. Neagu, A. O. Aptula, and G. Gini, "Neural and neuro-fuzzy models of toxic action of phenols," in *Proceedings of the First International IEEE Symposium "Intelligent Systems"*, vol. 1, 2002, pp. 283–288.

[25] C. D. Neagu, E. Benfenati, G. Gini, P. Mazzatorts, 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.

[26] Y. L. Loukas, "Adaptive neuro-fuzzy inference system: An instant and architecture-free predictor for improved QSAR studies," *J. Med. Chem.*, vol. 44, pp. 2772–2783, 2001.

[27] T. R. Cundari and M. Russo, "Database mining using soft computing techniques. an integral neural network-fuzzy logic-genetic algorithm approach," *J. Chem. Inf. Sci.*, vol. 41, pp. 281–287, 2001.

[28] A. Wlodawer, "Structure-based inhibitors of HIV-1 protease," *Annu. Rev. Bichem.*, vol. 62, pp. 543–585, 1993.

[29] A. Wlodawer and J. Vondrasek, "Inhibitors of HIV-1 protease," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 27, pp. 249–284, 1998.

[30] D. Leung, G. Abbenante, and D. P. Fairlie, "Protease inhibitors: Current status and future prospects," *J. Med. Chem.*, vol. 43, pp. 305–341, 2000.

[31] D. H. Rich, "Comprehensive medicinal chemistry," *Chem. Rev.*, vol. 2, pp. 391–441, 1990.

[32] A. C. Nair, P. Jayatilleke, 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.

[33] 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.

[34] W. Jorgensen, "The many roles of computation in drug discovery," *Science*, vol. 303, pp. 1813–1818, 2004.
[35] Y. S. Hwang and J. Chmielewski, "Development of low molecular weight HIV-1 protease dimerization inhibitors," *J. Med Chem*, vol. 48, pp. 2239–2242, 2005.
[36] A. B. Smith 3rd, L. D. Cantin, A. Pasternak, and L. Guise-Zawacki, "Design, synthesis, and biological evaluation of monopyrrolinone-based HIV-1 protease inhibitors," *J. Med. Chem.*, vol. 46, pp. 1831–1844, 2003.
[37] J. R. Tagat, R. W. Steensma, S. W. McCombie, D. V. Nazareno, S. I. Lin, B. R. Neustadt, K. Cox, S. Xu, L. Wojcik, M. G. Murray, N. Vantuno, B. M. Baroudy, and J. M. Strizki, "Piperazine-based ccr5 antagonists as hiv-1 inhibitors. ii. discovery of 1-[(2,4-dimethyl-3-pyridinyl)carbonyl]-4- methyl-4-[3(s)-methyl-4-[1(s)-[4-(trifluoro- methyl)phenyl]ethyl]-1-piperazinyl]- piperidine n1-oxide (sch-350634), an orally bioavailable, potent ccr5 antagonist," *J. Med. Chem.*, vol. 44, pp. 3343–3346, 2001.
[38] J. L. Jenkins, M. Glick, and J. W. Davies, "A 3d similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes," *J. Med. Chem.*, vol. 47, pp. 6144–6159, 2004.
[39] 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.
[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] 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.
[45] S. Makridakis, "Accuracy measures: theoretical and practical concerns," *International Journal of Forecasting*, vol. 9, pp. 527–529, 1993.
[46] F. E. Fischer, *Fundamental Statistical Concepts*. Canfield Press, 1999.
[47] D. M. Hawkins, "The problem of overfitting," *J. Chem. Inf. Sci.*, vol. 44, pp. 1–12, 2004.
[48] G. W. Kauffman and P. C. Jurs, "Prediction of inhibition of the sodium ion-proton antiporter by benzoylguanidine derivatives from molecular structure," *J. Chem. Inf. Sci.*, vol. 40, pp. 753–761, 2000.

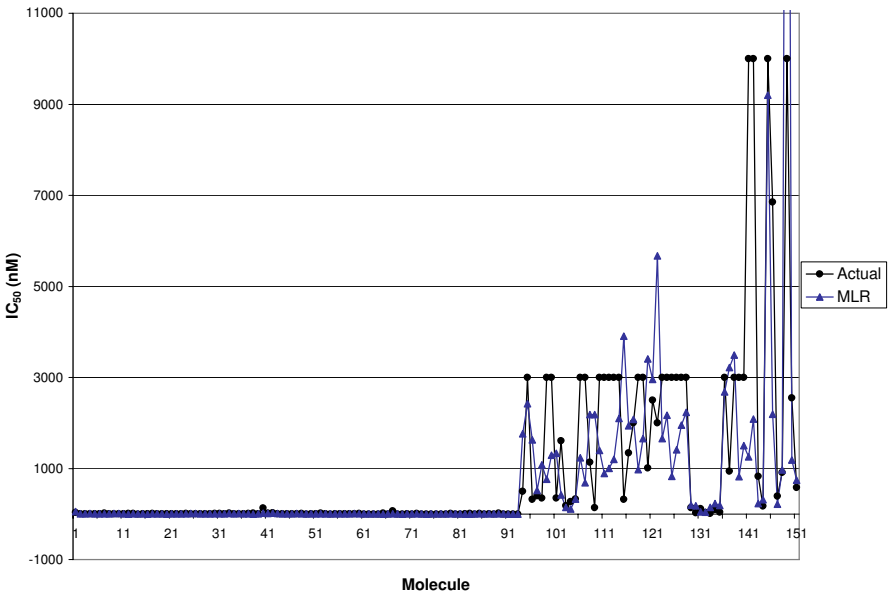


Fig. 1. Cross-validation results using MLR prediction.

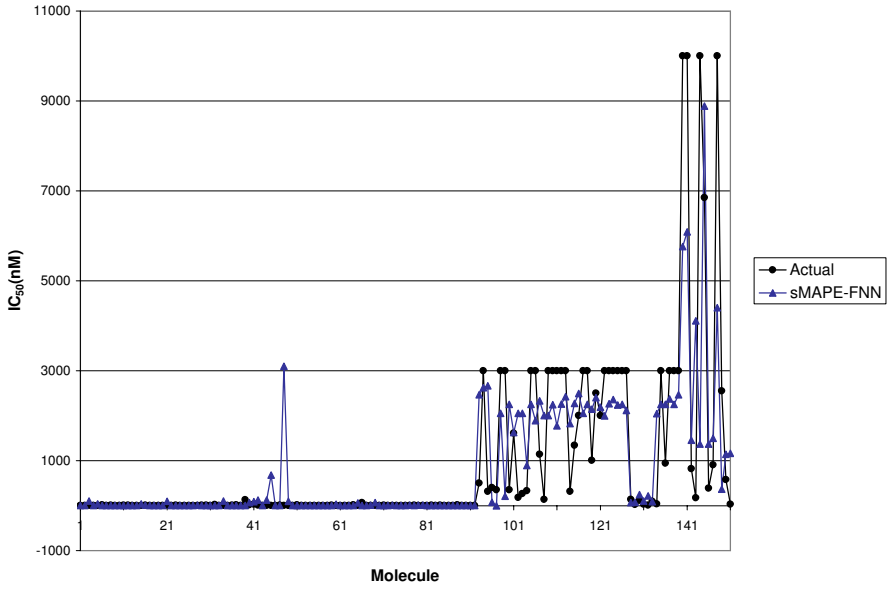
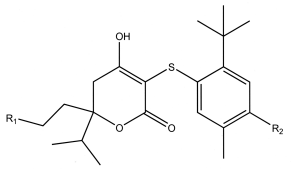


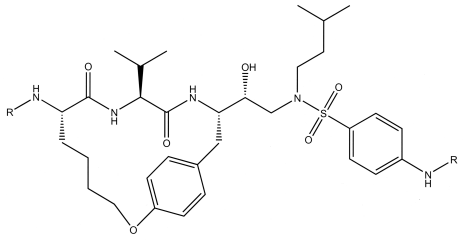
Fig. 2. Cross-validation results using sMAPE-FNN prediction.

TABLE III

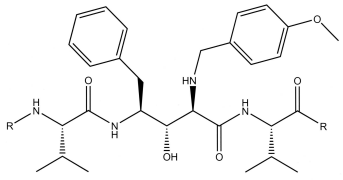
NOVEL INHIBITORY STRUCTURES WITH THEIR CORRESPONDING PREDICTED IC_{50} VALUES



no.	R_1	R_2	Predicted IC_{50} (nM)	
			MLR	sMAPE-FNN
1	OCH_3	Cbz	0.97	1.21
2	Cbz	OCH_3	-0.09	2.08
3	Cbz	COH	0.07	1.83
4	Cbz	NH_2	0.73	1.25
5	Cbz	$OSO_2Ph(4-CN)$	2.18	2.05
6	OCH_3	COH	-0.46	1.26
7	OCH_3	NH_2	-027	46.67
8	OCH_3	$OSO_2Ph(4-CN)$	1.36	2.05
9	$Ph-OH$	COH	-0.24	1.26
10	$Furan$	NH_2	-0.35	55.89
11	$Furan$	$OSO_2Ph(4-CN)$	0.73	2.05
12	$Thiophene$	NH_2	-0.19	1.34
13	$Thiophene$	$OSO_2Ph(4-CN)$	1.2	2.05
14	$4-methylthiazole$	NH_2	-0.35	1.30
15	$4-methylthiazole$	$OSO_2Ph(4-CN)$	0.49	2.05



no.	R	Predicted IC_{50} (nM)	
		MLR	sMAPE-FNN
16	OCH_3	165.6	5327.77
17	COH	321.01	821.69
18	Cbz	263.45	656.25
19	$Furan$	95.07	2283.33
20	$Thiophene$	99.89	2283.33
21	$4-methylthiazole$	71.73	2283.33



no.	R	Predicted IC_{50} (nM)	
		MLR	sMAPE-FNN
22	OCH_3	126.85	1500.00
23	COH	162.87	1500.00
24	Cbz	2668.3	1999.99
25	$Furan$	116.29	1409.55
26	$Thiophene$	143.49	999.99