Analysis of the Internal Representations Developed by Neural Networks for Structures Applied to Quantitative Structure−Activity Relationship Studies of Benzodiazepines

A. Micheli,[†] A. Sperduti,*[,‡] and A. Starita[§]

Dipartimento di Informatica, Università di Pisa, Corso Italia 40, 56125 Pisa, Italy

A. M. Bianucci[⊥]

Dipartimento di Scienze Farmaceutiche, Via Bonanno 6, 56126 Pisa, Italy

Received September 27, 1999

An application of recursive cascade correlation (CC) neural networks to quantitative structure−activity relationship (QSAR) studies is presented, with emphasis on the study of the internal representations developed by the neural networks. Recursive CC is a neural network model recently proposed for the processing of structured data. It allows the direct handling of chemical compounds as labeled ordered directed graphs, and constitutes a novel approach to QSAR. The adopted representation of molecular structure captures, in a quite general and flexible way, significant topological aspects and chemical functionalities for each specific class of molecules showing a particular chemical reactivity or biological activity. A class of 1,4-benzodiazepin-2-ones is analyzed by the proposed approach. It compares favorably versus the traditional QSAR treatment based on equations. To show the ability of the model in capturing most of the structural features that account for the biological activity, the internal representations developed by the networks are analyzed by principal component analysis. This analysis shows that the networks are able to discover relevant structural features just on the basis of the association between the molecular morphology and the target property (affinity).

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 big challenge in many fields of research, such as chemistry, biochemistry, pharmaceutical chemistry, etc. The assessment of such relationships represents the starting point for the prediction of required properties of new molecules. The ability of a model to predict specific properties of molecules allows the researchers to rationally design new compounds optimizing the requirement of both human and financial resources, so that the achievement of good predictive models constitutes a big task for either the basic or the applied research.

Many mathematical models were developed in the past with the aim of analyzing relationships between molecular structures and target properties such as chemical reactivity or biological activity. The earliest methods all imply a nondirect correlation of the molecular structure to the target property. In these models some physicochemical properties were currently used as molecular descriptors. They should be better classified as property−property or property−activity relationship models. The major problem in correlating some molecular properties (reflecting different structural aspects of molecules) to other kinds of properties (typically chemical reactivity or biological activity) is represented by the need to find a set of complete and relevant molecular descriptors.

The problem of identifying such proper descriptors, which initially had led to the use of physicochemical properties,[1−3] subsequently was faced by the use of a wide class of numerical descriptors, more specifically oriented to the representation of molecular geometry/shape and atom connectivities (topological indices).[4−7] Although these last methods use chemical graphs as versatile vehicles for representing structural information, the chemical graphs need to be encoded into the vectorial (or matricial) form required by the technique used to solve the regression problem. Of course, this encoding process is going to strip out structural information which may be relevant.

The mathematical and computational tools used in quantitative structure−activity relationship (QSAR) based drug design are quite different from each other and include equation-based models[1,2] and neural-network-based models.[8−10]

In summarizing the evolution toward the use of more direct representations of the molecular structures, we can mention models based on measurable or calculable physicochemical properties,[11−14] on topological indices,[15,16] or on matricial[17] graph representations, and finally a template-based approach.[18] This last model uses a neural network which partially mimics the chemical structures of the analyzed compounds by means of a common molecular template, staticly defined for all the compounds.

In this paper we show that neural networks for structures (or recursive neural networks[19]) allow a new approach to the analysis of QSAR. In fact, this class of networks can take the chemical graph directly as input and are able to learn the desired mapping (e.g., biological activity) on the basis

[†] E-mail: micheli@di.unipi.it. Phone: +39-050-887213. Fax: +39-050-887226.

[‡] E-mail: perso@di.unipi.it. Phone: +39-050-887213. Fax: +39-050-887226.

[§] E-mail: starita@di.unipi.it. Phone: +39-050-887213. Fax: +39-050-887226.

[⊥] E-mail: bianucci@farm.unipi.it.

10.1021/ci9903399 CCC: $20.00
Published on Web 11/10/2000

of a set of training examples. Specifically, given a structural representation of chemical graphs, a recursive network automatically encodes the structural information depending on the computational problem at hand, so that the numerical representation of molecular structures is not defined a priori by using a set of descriptors, but is learned as a result of the training process. The use of recursive neural networks therefore allows a target property to be directly correlated to the molecular structure of the compounds under analysis. Until now the model was applied, to evaluate its performance, to different quantitative structure−property relationship (QSPR) and QSAR tasks such as the analysis of a group of alkanes and a group of benzodiazepines.[22,23] The results outperformed those obtained by traditional equation-based approaches and were competitive with feed-forward neural networks using ad hoc chemical compound representations designed by QSPR/QSAR experts.

The focus of the present work consists of analyzing the internal representations of molecular structure developed by cascade correlation (CC) for structures trained on a specific QSAR task, and of interpreting such results as a function of the chemical meaning. This analysis is fundamental for a full assessment of the proposed methodology. In fact, if it is not possible to explain the good predictive performances of the proposed model by the development of internal representations which are directly correlated with the key features responsible for the physical or biological properties under examination, then it would be difficult to trust the model itself, since it would not be clear on which grounds the model generates its predictions.

Here, through a principal component analysis (PCA)[24] of the internal representations developed by the CC for structures network, we show that it is actually possible to demonstrate a strong correlation between the developed internal representations and structural features on which the studied biological activity hinges. This result confirms the ability of the model in capturing structural information which is relevant for the prediction task at hand.

The paper is organized as follows. Section II begins with the outline of the traditional QSAR approach and is followed by the introduction of the new QSAR approach based on recursive neural networks. A QSAR problem involving a class of benzodiazepines is explained in section III, with details on our representation of molecular structures given in section III.A. Simulation results are reported in section IV, where the PCA study of the internal representations is presented as well. The conclusions are drawn in section V.

II. TOWARD A NEW QSAR APPROACH: NEURAL NETWORKS FOR STRUCTURES

In this section we describe a new QSAR approach based on neural networks for processing data structures. First of all we briefly review the traditional way of performing QSAR studies. Then we suggest how the use of neural networks for processing data structures may help in reducing the burden of developing and selecting relevant structural features for molecular representation.

The aim of a QSAR study is to find an appropriate function $\mathcal{T}()$ which, given a structured representation of a molecule, predicts its biological activity, i.e.

$$\text{activity} = \mathcal{T}(\text{structure}) \tag{1}$$

The function $\mathcal{T}: I \rightarrow O$ is therefore a functional transduction from an input structured domain I, where molecules are represented, to an output domain O, such as the real number set. In eq 1 the term "structure" stresses the importance of the use of global information about molecular shape, atom connectivities, and chemical functionalities as understood in the QSAR studies.

The function $\mathcal{T}()$ is a complex object which can be described as the sequential solution to two main problems: (i) the *representation problem*, i.e., how to encode molecules through the extraction and selection of structural features; (ii) the *mapping problem*, i.e., the regression task usually performed by linear or nonlinear regression tools (e.g., equational modeling and feed-forward neural networks).

According to this view, $\mathcal{T}()$ can be decomposed as follows:

$$\mathcal{T}() = g(\tau()) \tag{2}$$

where $\tau()$ is the *encoding* function from the domain of the chemical compounds to the descriptor space, while g is the *mapping* function from the descriptor space to the biological activity space. This corresponds to the traditional QSAR approach, as summarized in Figure 1, where chemical features are represented by a suitable set of numerical descriptors (function τ), which are then used to predict the biological activity (function g). The representational problem is faced by using different approaches such as the definition and selection of physicochemical or geometrical and electronic properties, the calculation of topological indices, or an explicit vector-based representation of molecular connectivity. The question mark in the picture shown in Figure 1 stresses that the number and type of descriptors used to represent the chemical compound depend on the specific QSAR problem at hand. The exact number and type of descriptors used for a specific study are decided by an expert in the field.

In more detail, the encoding process requires the solution of two subtasks. The aim of the first one is to explicitly represent the relevant structural information carried by molecules, while the second one is to codify this structural information into a numerical representation. For example, when considering topological indices, first of all a molecule is represented by the molecular graph skeleton, and then invariant properties of the molecular graph skeleton are used to define and compute a numerical formula. Thus, the function τ can be understood as the following composition:

$$\tau() = \tau_E(\tau_R()) \tag{3}$$

where τ_R extracts a specific structure from the molecule (i.e., the solution to the first subtask) and τ_E computes a numerical value from the structure returned by τ_R (i.e., the solution to the second subtask). Examples of τ_E are the connectivity indices (χ) or the hydrophobic, electronic, polar, and steric properties.

In traditional QSAR, both τ_R and τ_E are defined a priori; i.e., they do not depend on the regression task. Therefore, they are designed through a very expensive trial and error approach to adapt them to the regression problem required by the QSAR study. So, even if the chemical graph is clearly

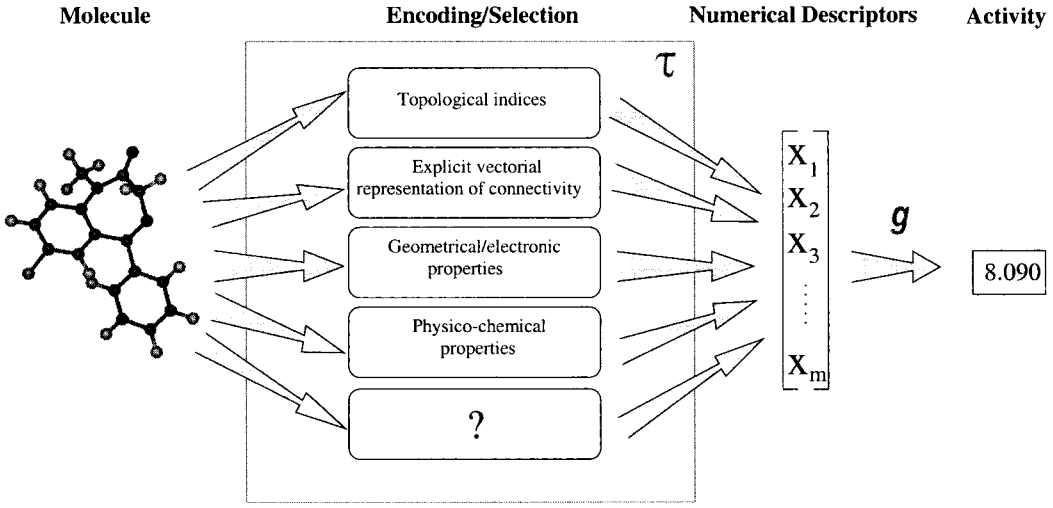

Figure 1. Outline of the traditional QSAR approach. Structural features of the molecule are represented through different numerical descriptors. The numerical descriptors can be obtained by using different approaches. Their number and type depend on the QSAR task at hand. The encoding process on the whole defines the τ function. A regression function (g) is then applied to the numerical descriptors to obtain the predicted biological activity.

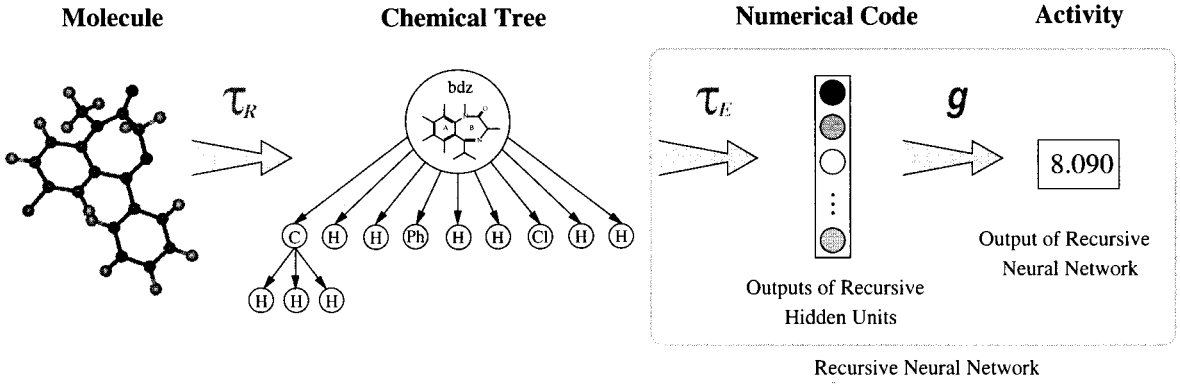

Figure 2. New QSAR scheme using recursive neural networks. The molecule, after a structural coding phase driven by ad hoc rules (τ_R), is directly processed by the recursive neural network through the adaptive encoding function τ_E. The internal representation developed by the recursive neural network is then used by the regression model implemented by the output part of the neural network (function g) to produce the final prediction (activity).

recognized as a flexible vehicle for the rich expression of chemical structural information, the problem of using it in a form amenable directly to QSAR analysis is still open.

In this paper we propose to realize the τ_E function through an adaptive mapping, thus allowing the automatic generation of numerical descriptors which are specific for the regression task to be solved. This can be done by using recursive neural networks,[19] which are able to take directly as input the graph generated by τ_R and to implement adaptively both τ_E and g.

To exemplify, in Figure 2, we show the outline of the proposed approach assuming that a given molecule is represented by τ_R as a labeled tree. [The definition of an appropriate function τ_R for the specific set of molecules studied in this paper is discussed in section III.A.] This tree-structured representation is then processed by a recursive neural network. The output of the recursive neural network constitutes the regression output, while the internal representations of the recursive neural network (i.e., the output of the hidden units) constitute the neural implementation of the numerical descriptors returned by τ_E. It must be stressed, at this point, that the recursive neural network does not need to take as input a fixed-size numerical vector for each input graph, as happens with standard neural networks typically used in QSAR studies, because it is able to treat variable-size representations of the input graph. Moreover, since the encoding function (τ_E) is learned by the neural network together with the mapping function (g), the resulting numerical code represents the "best" numerical coding of the input graph for the given QSAR task.

We may observe that the main difference between the traditional QSAR scheme shown in Figure 1 and the proposed new scheme reported in Figure 2 is due to the automatic definition of the τ_E function obtained by training the recursive neural network over the regression task. This implies that no a priori selection and/or extraction of features or properties by an expert is needed in the new scheme for τ_E.

To fully grasp the mathematical model underpinning recursive neural networks within the context outlined in Figure 2, it is crucial to understand how the encoding function, i.e., τ_E, is computed for each input graph.

For the sake of exposition, in the following we assume that τ_R returns labeled trees, where each label associated with each node of the tree is a symbol representing, for example, the atom type or a molecular group. Since τ_E will be realized by a recursive neural network, these symbols need to be

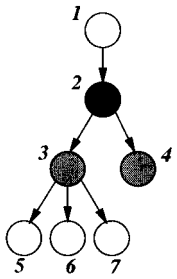

Figure 3. Coding process. A code is progressively generated for each node by using the code already produced for its descendants. Nodes colored with different gray levels are used to denote the time when the code of each node is used as state information for the current node: e.g., the code for node 2 is generated by using the codes generated for nodes 3 and 4 (in addition to the numerical label attached to node 2).

represented as numerical vectors. For example, a bipolar localist representation can be used to code (and to distinguish among) the types of chemical objects. In a bipolar localist representation each component of the vector is assigned to one entity and is equal to 1 if and only if the representation refers to that entity; otherwise it is set to -1. For example, assuming that the fluorine atom (F) is associated with the ith component and the clorine atom (Cl) is associated with the jth component, the fluorine atom is represented by the vector

$$[\underbrace{-1,-1,...,-1}_{i-1},1,-1,...,-1,-1]$$

while the clorine atom is represented by

$$[\underbrace{-1,-1,...,-1}_{j-1},1,-1,...,-1,-1]$$

The computation of τ_E is a progressive process which starts from the leaves of the input tree and terminates at the root of the tree, where a numerical code for the whole tree is generated. Specifically, this coding process starts at the leaf level by producing step by step a code for each visited leaf node and by storing these codes as state information for each corresponding leaf. Successively, the internal nodes are visited, from the frontier to the top of the tree. For each currently visited node its numerical label and the codes already computed for its children (stored in the state) are used to compute the code for the current node. Since this computation is performed in the same way for all the nodes in the tree, the generated codes are all constrained to be of the same size. Finally, the code computed for the root of the tree is used as the numerical code for the whole tree. The encoding function τ_E is therefore seen as a *state transition* function. Note that for leaf nodes the process starts with a null state because there is no previous information from descendants.

In Figure 3 we exemplify the above visit on an input tree where the labels are not explicitly represented. First, the leaves (nodes 4−7) are visited and the corresponding codes are generated. Then, node 3 is visited and a code for it is produced taking into account its label and the codes generated for its children, i.e., nodes 5−7. Successively, a code is computed for node 2 using the codes computed for (the subtrees rooted in the) nodes 3 and 4 and the label of node 2. Finally, the root node 1 is visited and the code for it, corresponding to the code for the whole tree, is generated. The different gray levels used to fill in the tree nodes convey information about the time when the code of each node is used as state information for the current node.

Note that the way the encoding function acts on a specific tree, such as the tree in Figure 3, is specified in terms of how the encoding function acts on the subtrees of each node. In this sense the encoding is "recursive". Moreover, the encoding is *stationary* and *causal*. Stationary means that the computation that produces the code is the same for all the nodes, while causal means that the computation of each code depends only on the current node and nodes descending from it.

Concerning the regression function g, it takes as input the code generated by τ_E for the root of each input tree and returns the desired value associated with the tree.

A. Recursive Neural Networks in QSAR. At this point we formally provide a proper instantiation of the input and output domains for the encoding and the output functions.

Let the structured input domain for τ_E, denoted by $\mathcal{G}$, be a set of labeled directed ordered acyclic graphs (DOAGs), as produced by the application of τ_R to the input data set of molecules $\mathcal{I}$. For a DOAG we mean a DAG where for each vertex a total order on the edges leaving from it is defined. Moreover, let us assume that $\mathcal{G}$ has for each node a bounded out-degree. Labels are tuples of variables and are attached to vertexes. Let IR^n denote the label space.

The descriptor (or code) space is chosen as IR^m, while the output space, for our purpose, is defined as $\mathcal{O} = IR$.

Finally, we define the encoding function as

$$\tau_E: \mathcal{G} \rightarrow IR^m \tag{4}$$

and the output function g as

$$g: IR^m \rightarrow IR \tag{5}$$

The use of a stationary and causal model for τ_E allows a uniform and quite simple neural realization for each step of τ_E to be chosen through the definition of a recursive neural network model. To process each node, the recursive neural network uses the information available at the current node: (i) the numerical label attached to the node; (ii) the numerical code for each subgraph of the node (state information).

As a result, if k is the maximum out-degree of DOAGs in $\mathcal{G}$, the recursive neural network, for each step of τ_E, gets input from the space

$$IR^n \times \underbrace{IR^m \times \cdots \times IR^m}_{\text{k times}}$$

and produces a code in IR^m.

Let us consider, for example, a recursive neural network with m hidden neurons. Given the current visited node, the output $\mathbf{x} \in IR^m$ of the hidden neurons (i.e., the code for the current node) is computed as follows:

$$\mathbf{x} = \mathbf{F}(\mathbf{W}\mathbf{l} + \sum_{j=1}^{k} \hat{\mathbf{W}}_j \mathbf{x}_{ch[j]} + \boldsymbol{\theta}) \tag{6}$$

where $\mathbf{l} \in IR^n$ is the label (external input) associated with the current node, $\mathbf{W} \in IR^{m \times n}$ is the weight matrix associated with the label space, $\hat{\mathbf{W}}_j \in IR^{m \times m}$ is the recursive weight

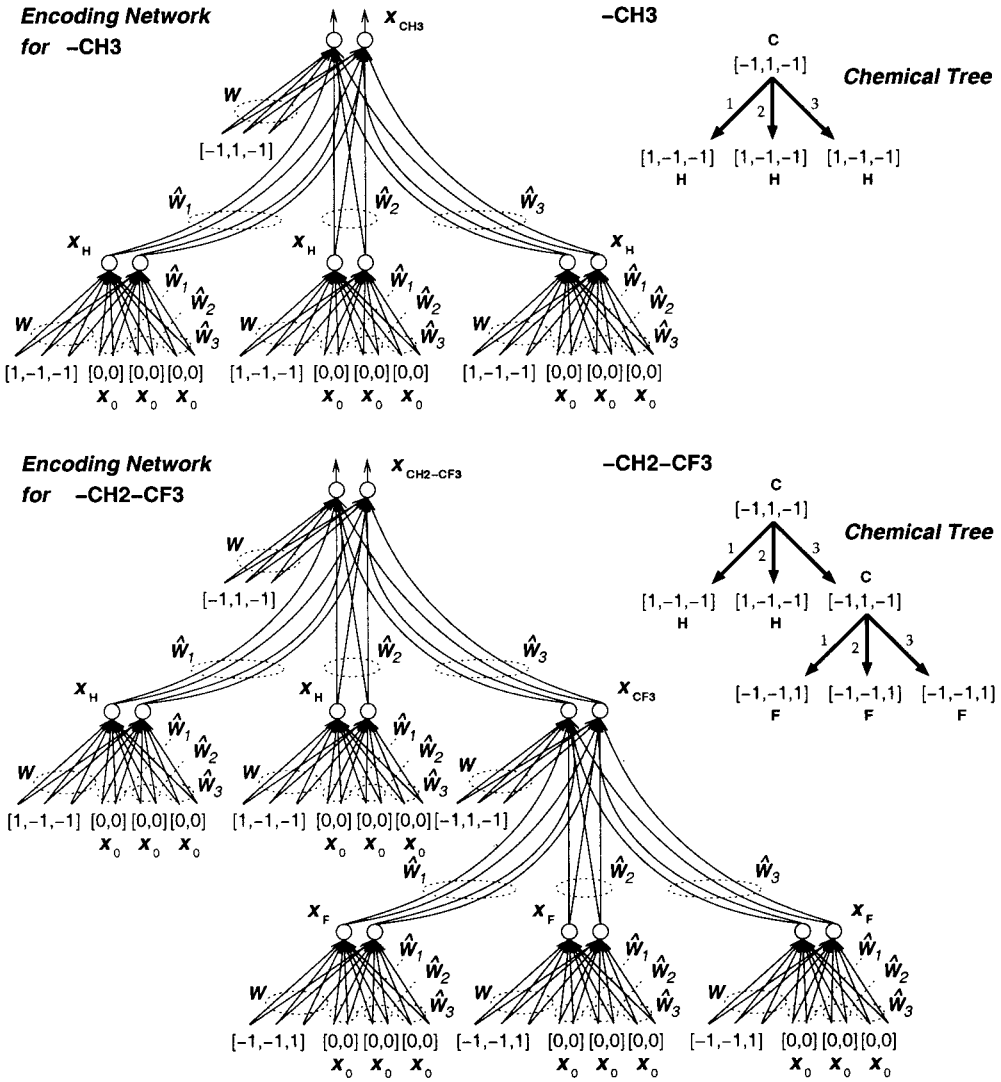

Figure 4. Examples of encoding networks (left side) for the chemical fragments $-CH_3$ and $-CH_2-CF_3$ with $n = 3$ and $m = 2$. The fragments are assumed to be represented by the chemical trees shown on the right side of the figure. The labels of the chemical trees represent the atom types: H is represented by [1, −1, −1], C by [−1, 1, −1], and F by [−1, −1, 1]. The encoding networks are obtained by replicating (unfolding) the recursive neurons for each node in the chemical trees (as shown by the multiple occurrences of the weight matrixes). Void subgraphs are encoded by the null vector $\mathbf{x}_0$. The output of each encoding network is the code computed for the corresponding chemical fragments (i.e., $\mathbf{x}_{CH_3}$ and $\mathbf{x}_{CH_2-CF_3}$, respectively).

matrix associated with the jth subgraph code, $\mathbf{x}_{ch[j]} \in IR^m$ is the code computed for the jth subgraph of the current node, $\boldsymbol{\theta} \in IR^m$ is the bias vector, and $\mathbf{F}(\mathbf{y})_i = f(\mathbf{y}_i)$, where $f(\cdot)$ is a sigmoidal nonlinear function.

Using eq 6, the recursive hidden neurons can realize each step of τ_E. Finally, in the simplest case, the output mapping function $g()$ is realized by a single standard neuron with m inputs.

The neural encoding process of an input graph can be represented graphically by replicating the same recursive neurons (through the input graph) and connecting these replica according to the topology of the input graph. We obtain in this way the so-called *encoding network*. Examples of encoding networks for $n = 3$ and $m = 2$ are shown in Figure 4. The examples involve two substituents ($-CH_3$ and $-CH_2-CF_3$) for the benzodiazepine class of molecules studied in this paper, where for the sake of simplicity, the labels shown here represent only the three different atoms involved in these examples (i.e., H is represented by [1, −1, −1], C by [−1, 1, −1], and F by [−1, −1, 1]).

The encoding network is a feedforward network that mimics the topology of the molecular graph. For each input graph a corresponding encoding network is built up. There is a correspondence between graph nodes and units of the encoding network; however, the template used to encode the molecular graph is not fixed a priori as happens in the template-based approach used in ref 18. Notice that the weight matrixes are shared by different encoding networks (see Figure 4), since the same recursive neurons are used to "visit" the nodes of different input graphs. This is a consequence of the use of a stationary model.

The neural network output for a given molecular graph is obtained by completing the corresponding encoding network with the neural realization of $g()$. Such a completed network is trained on the regression task. Thus, both the weights of the hidden recursive neurons and the weights of the output neuron (realizing $g()$) are trained simultaneously on the training set. As a result of this joint training, the encoding of the molecular graph is adaptive, since it is computed on the basis of the specific regression task.

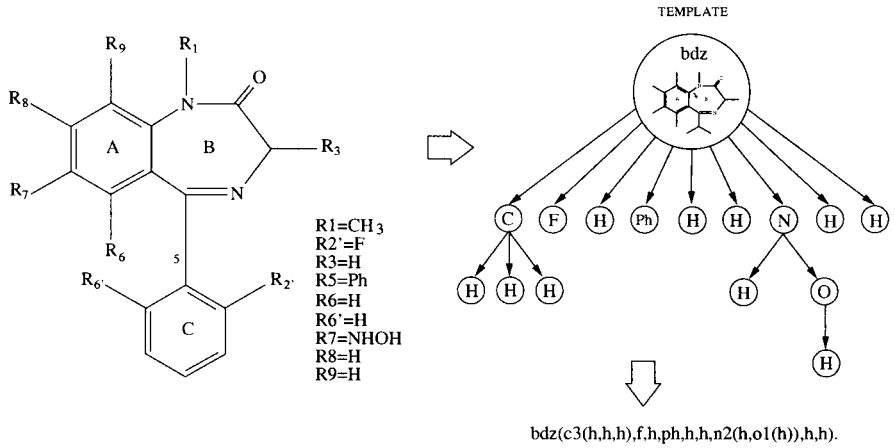

Figure 5. Example of a representation for a benzodiazepine.

There are different ways to realize the recursive neural network.[19] In the present work we choose to use a constructive approach that allows the training algorithm to progressively add the hidden recursive neurons during the training phase. The model is an (recursive) extension of cascade-correlation-based algorithms.[25,26] The built neural network has a hidden layer composed of recursive (hidden) units. The recursive hidden units compute the values of τ_E (in IR^m) for each input DOAG, as shown in Figure 4. The number of hidden units, i.e., the dimension m of the descriptor space, is automatically computed by the training algorithm, thus allowing an adaptive computation of the number and type of (numerical) descriptors needed for a specific QSAR task. In the CC for structures model, to realize the function g, we use a single standard linear output neuron. A complete description of the CC for structures algorithm and a formulation of the learning method and equations can be found in refs 19 and 20.

In summary, the hidden layer of a recursive network produces a numerical vectorial code (i.e., its internal representation) that represents the input molecular graph. In terms of QSAR studies, we can imagine that each hidden recursive neuron calculates an adaptive topological index on the basis of the information supplied to the model (i.e., the training set). The outputs of the hidden units are arranged into a vector of these topological indices and used as input for a linear regression model realized by the output unit (the $g()$ function), as shown in Figure 2. It is important to stress that these topological indices are automatically developed by the neural network, since they arise from the training process as a function of the relationship between structures and corresponding values of the target property. They are developed, for this reason, independently from the domain knowledge.

The advantage of this new approach is that it allows us to describe and to process a molecular graph in a way that considers both the graph topology (connectivity) and the atom types (or the chemical functionalities). The use of a neural network to realize the encoding and regression functions allows the production of a flexible prediction model. However, the use of a "black-box" approach to implement the encoding and the regression functions raises the following issues: (i) chemical meaningfulness of the numerical descriptors produced by the recursive neural network; (ii) relationship between the developed numerical codes and the qualitative aspects of the QSAR problem.

In the present work we try to address these issues by studying the internal representations developed by the recursive neural network trained on a specific family of benzodiazepines.

A complete answer to these issues would allow the extraction of the knowledge learned by the neural network, posing the basis for a full understanding by human experts of the model and therefore permitting the assessment of the model as a new tool for the rational design of new molecules.

III. QSAR TASK FOR BENZODIAZEPINES

Due to the strong therapeutic interest[27−31] and to the multiplicity of SAR studies[32,33] of this class of compounds, benzodiazepines were chosen as the starting application domain. At this stage, a group of 1,4-benzodiazepin-2-ones, previously studied by Hadjipavlou-Litina and Hansch[21] through traditional QSAR equations, was selected for testing our model, the evaluation of the method being the initial step of its application. The data set analyzed by Hadjipavlou-Litina and Hansch (see Table 2 of ref 21) is characterized by a good molecular diversity, and this last requirement makes it particularly significant for QSAR analysis. The total number of molecules analyzed was 77.

All the molecules present a common template consisting of the benzodiazepine nucleus (only in three compounds the A ring of the benzodiazepine nucleus consists of a thienyl instead of a phenyl group), and they differ from each other because of a large variety of substituents at the positions shown at the left side of Figure 5.

A. Molecular Structure Representation (Function τ_R by Rules). A specific type of representation of the molecular structure is required for the model presented here. The choice of the representation defines the function τ_R introduced in Figure 2. Since the functions τ_E and g are automatically developed by the model, in the new QSAR scheme the specification of function τ_R is the only one available for the designer's tuning.

Molecular structural formulas have already been treated in the literature as mathematical objects (graphs) according to chemical graph theory. In our case, a representation of molecular structures in terms of DOAGs is required. The candidate representation should contain detailed information about the shape of the compound, the atom types, the bond multiplicity, and the chemical functionalities, and finally it should retain a good similarity with the representations usually adopted in chemistry.

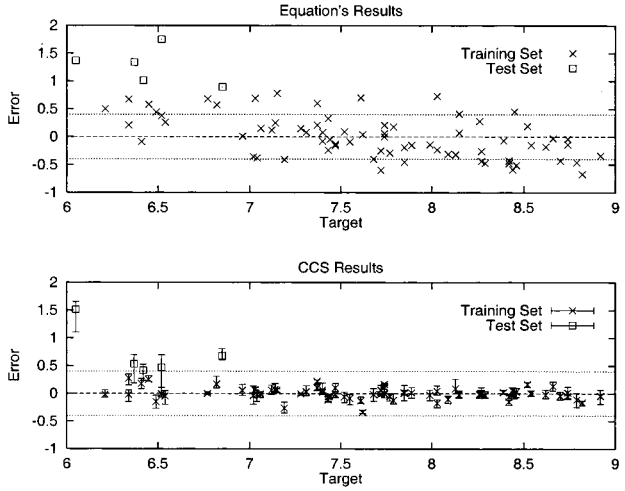

Figure 6. Residual error plot for the equation model proposed by Hadjipavlou-Litina and Hansch (top side) and for the CC for structures network (bottom side). Both models use the same training and test sets (data set I). Each point in the plots represents the average error, together with the deviation range (minimum and maximum values), as computed over six trials. The tolerance region is shown in the plots.

When the molecular structure is represented as a DOAG, the main representational problems which are encountered are (i) how to represent cycles, (ii) how to give a direction to edges, and (iii) how to define a total order over the edges.

An appropriate description of the molecular structures analyzed in this work is based on a labeled tree representation. The major atom group that repeats unchanged throughout the class of analyzed compounds (common template) constitutes the root of the tree. [An alternative representation, which the model was able to deal with, would have been to explicitly represent each atom in the major atom group. However, since this group is repeated for all the compounds, no additional information is conveyed by adopting this representation.] When other repeating atom groups do exist in all the analyzed molecules, single atoms, belonging to these groups, do not require to be explicitly represented. Each atom that requires to be explicitly represented or each repeating atom group corresponds to a node of the tree. Each bond that requires to be explicitly represented corresponds to an edge. A label is associated with each node. Here, these labels are just used to discriminate among different atoms (or atom groups) and do not contain any physicochemical information. The use of DOAGs for the molecular description implies the loss of only minor structural information. At the present level of development of the model, cycles are usually treated as repeating atom groups, for which a single label is

Table 1. Results Obtained for Benzodiazepines on Training Data Set I by Hadjipavlou-Litina and Hansch (HLH, First Row) and by a "Null Model" (Second Row) and on All the Training Data Sets by CC for Structures[a]

training set	mean no. of units (min−max)	mean abs error (min−max)	*R*	*S*
HLH		0.311	0.847	0.390
null model		0.580	0	0.702
data set I	29.75 (23−40)	0.090 (0.066−0.114)	0.99979	0.127
data set II	34.0 (27−38)	0.087 (0.080−0.102)	0.99982	0.117
data set III	19.7 (18−22)	0.087 (0.072−0.105)	0.99985	0.098
data set IV	16.5 (13−20)	0.099 (0.078−0.132)	0.99976	0.131

[a] The mean absolute error, the correlation coefficient (*R*), and the standard deviation of error (*S*) are reported.

Table 2. Results Obtained for Benzodiazepines on Test Data Set I by Hadjipavlou-Litina and Hansch (HLH, First Row) and by a "Null Model" (Second Row) and on All the Test Data Sets by CC for Structures[a]

test set	data no.	mean abs error (min−max)	mean max abs error (min−max)	S
HLH	5	1.272	1.750	1.307
null model	5	1.239	1.631	1.266
data set I	5	0.720 (0.611−0.792)	1.513 (1.106−1.654)	0.842
data set II	4	0.546 (0.444−0.653)	0.727 (0.523−0.973)	0.579
data set III	5	0.255 (0.206−0.325)	0.606 (0.433−0.712)	0.329
data set IV	4	0.379 (0.279−0.494)	0.746 (0.695−0.763)	0.460

[a] The mean absolute error, the mean of the maximum of the absolute error, and the standard deviation of error (*S*) are reported.

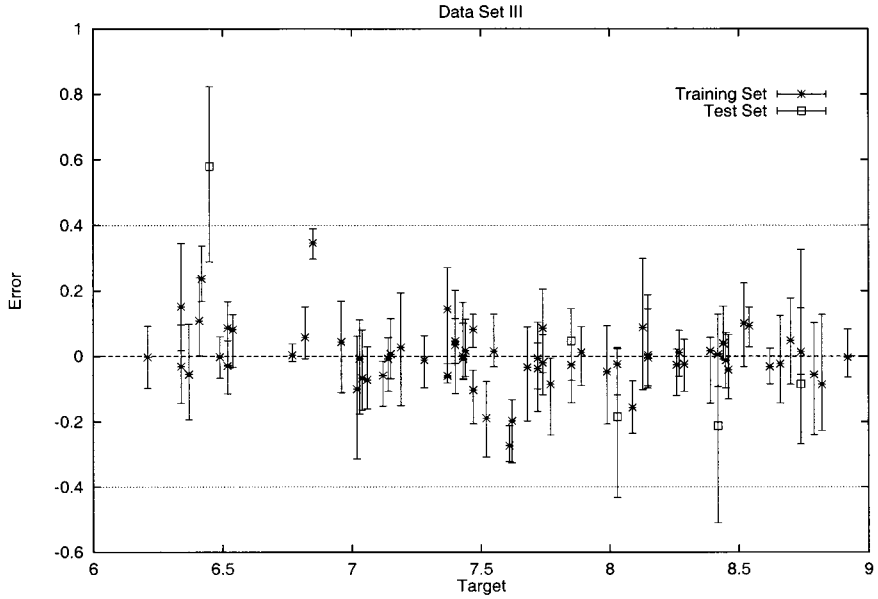

Figure 7. Residual error plot for the CC network using data set III. Each point in the plot represents the average error, together with the deviation range (minimum and maximum values), as computed over six trials. Note that the test data are spread across the input range.

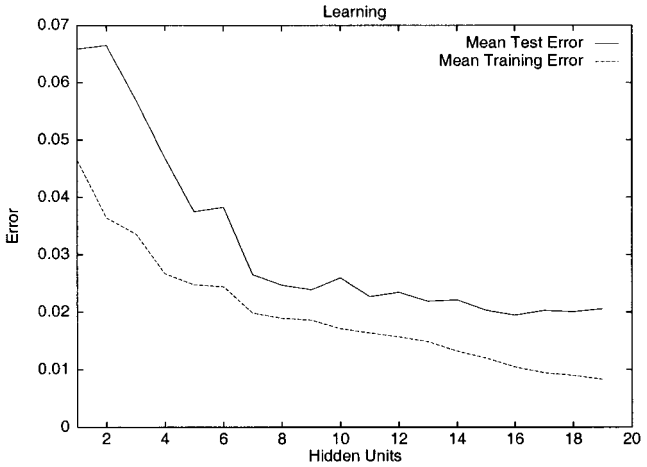

Figure 8. Mean training and test errors for a CC for structures network trained on data set III. The mean error is plotted versus the number of inserted hidden units.

used. When different types of cycles are present at corresponding positions of the molecular structure throughout the class of analyzed compounds, different labels are used to describe them.

The representational scheme described above basically solves all the representational problems i−iii. In fact, with reference to the benzodiazepine data set, concerning the first problem, since cycles mainly constitute some common shared template of the benzodiazepine compounds, it is reasonable to represent them as a single node where the attached label codifies information about their chemical nature. [We distinguish different principal heterocycles or cycles that appear as substituents using different labels.] The second problem was solved using the major common template as the root of a tree representing a benzodiazepine molecule. Finally, the total order over the edges follows a set of rules mainly based on the size of the molecular fragments.

More precisely, the labeled tree representation is obtained by the following minimal set of rules: (1) The root of the tree represents the Bz nucleus. (2) The root does have as many subtrees as substituents on the Bz nucleus, sorted according to the order conventionally followed in chemistry (standard IUPAC numbering of substituent positions). (3) Each explicitly represented atom (or any other common atomic group) of a substituent corresponds to a node, and each explicitly represented bond (the multiplicity of the bound is implicitly encoded in the structure of the subtree) to an edge. The root of each subtree that represents the substituent is the atom directly connected to the common template, and the orientation of the edges follows the increasing levels of the trees. (4) Different atoms (or any other common atomic group) are represented by different labels, and each node in the trees has a label associated with it. (5) The total order on the subtrees of each node is hierarchically defined according to (i) the subtree's depth, (ii) the number of nodes of the subtree, and (iii) the atomic weight of the subtree's root.

In the analyzed data set different labels are used for the following atoms: C, N, O, F, Cl, Br, I, and H. Moreover, we use a different label for each of the following atomic groups: bdz (Bz nucleus), bdztg (Bz nucleus where the A ring is a thienyl group instead of a phenyl one), and ph, py, cya, and naf, respectively, for fragments of phenyl, 2-pyridyl, cyclohexenyl, cyclohexyl, and naphthyl. For labeling we use a bipolar localist representation, as shown in section II.

Examples of representations for benzodiazepines (or substituents) which comply with the above rules are shown in Figure 5 (compound no. 60 in Table 3 in the Appendix) and in Figure 4.

IV. EXPERIMENTAL RESULTS: INTERNAL REPRESENTATION ANALYSIS

In this section, after recalling experimental results obtained by using the sum of square errors as global error function, we demonstrate the ability of the proposed model to learn from the training data relevant knowledge about the application domain. This is shown by studying the internal representations developed by the model through PCA.

A. Model Evaluation. In this section we briefly summarize experimental results obtained for the QSAR task.[23]

For the analysis of the data set described in section III, four different splittings in disjoint training and test sets of the data were used (data sets I, II, II, and IV, respectively). Specifically, the first test set (five compounds) has been chosen as it contains the same compounds used by Hadjipavlou-Litina and Hansch. The second data set is obtained

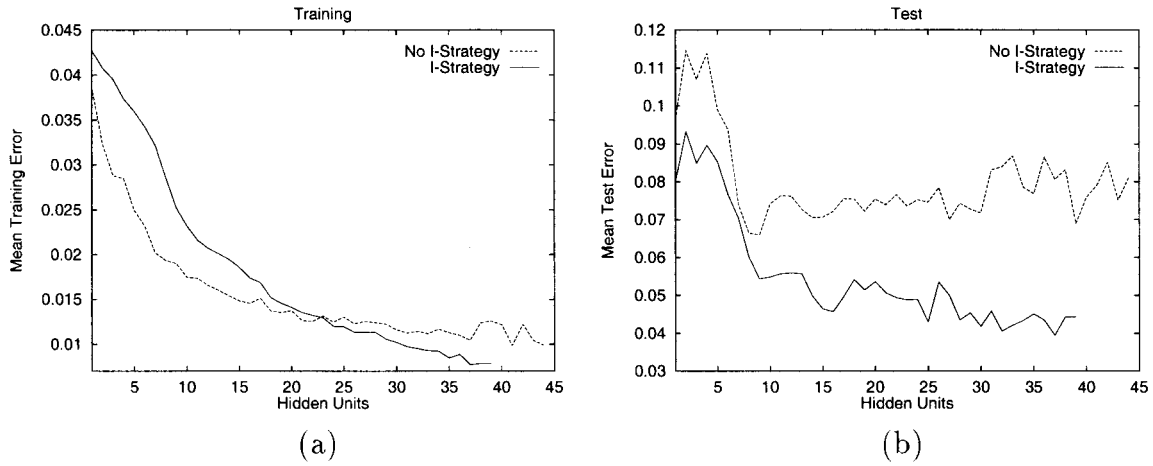

Figure 9. Mean training and test errors for two different instances, using or not using the i-strategy, of CC for structures networks trained over data set I. The mean error is plotted versus the number of inserted hidden units. The i-strategy allows one to reach a lower training error using less units (a). The test error decreases as a function of the number of hidden units only when using the i-strategy (b).

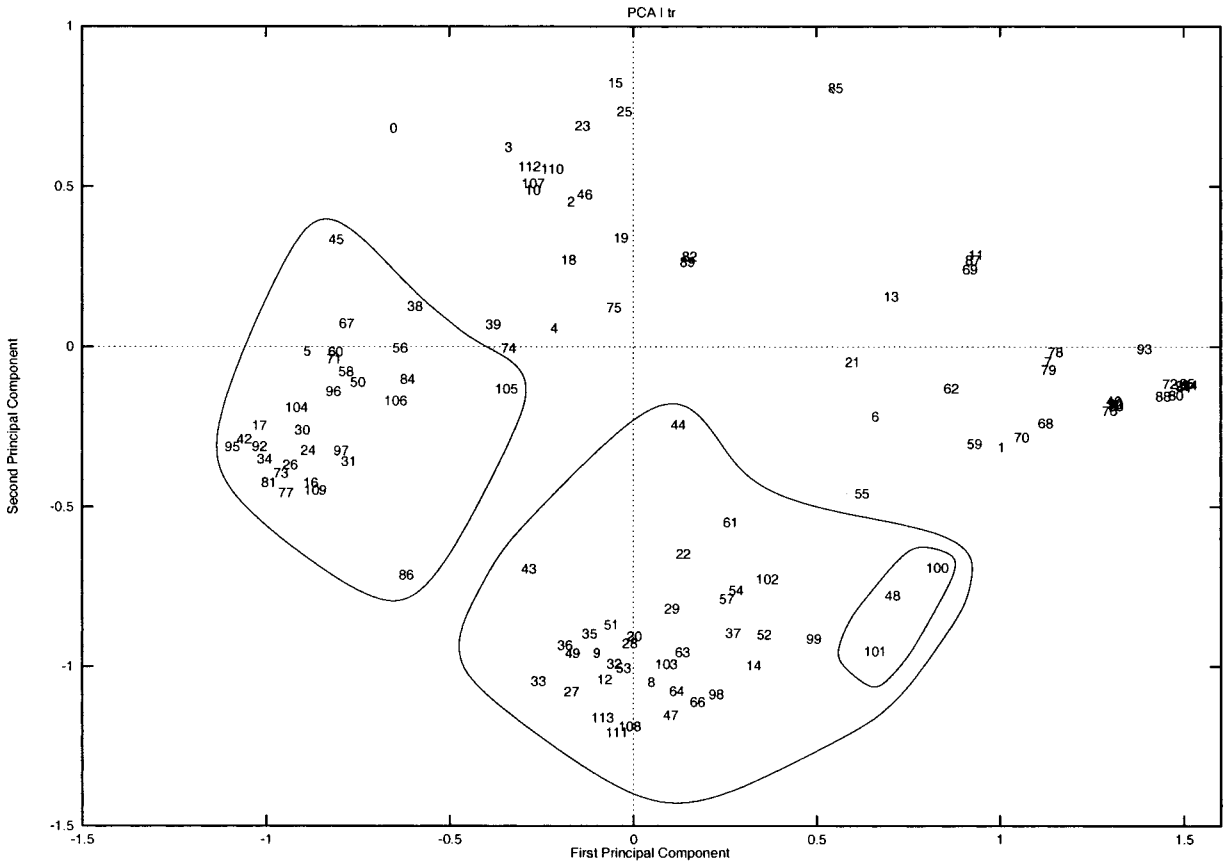

Figure 10. Principal component plot of training compounds used in experiment I derived from 28 output values of hidden neurons. Compounds characterized by $R_1 = H$ (left side of the plot) and compounds bearing a substituent at position 1 (lower side of the plot) are grouped by contour lines. The circled subcluster on the right side includes compounds where the A ring of the benzodiazepine nucleus is a thienyl group instead of a phenyl group. See Table 3 in the Appendix for compound numbering.

from data set I by removing four racemic compounds from the training set and one racemic compound from the test set. This allows the experimentation of our approach without the racemic compounds, which are commonly recognized to introduce ambiguous information. The test set of data set III (five compounds) has been selected as it simultaneously shows a significant molecular diversity and a wide range of affinity values. Furthermore, the included compounds were selected so that substituents, already known to increase the affinity on given positions, appear in turn in place of H atoms, which allows the decoupling of the effect of each substituent. So, a good generalization on this test set means that the network is able to capture the relevant aspects for the prediction. The test set of data set IV (four compounds) has been randomly chosen to test the sensitivity of the network to different learning conditions.

As the target output for the networks we used $\log(1/C)$. Six trials were carried out for the simulation involving each one of the different training sets. The initial connection weights used in each simulation were randomly set. Learning was stopped when the maximum error for a single compound was below 0.4. This tolerance is largely below the minimal tolerance needed for a correct classification of active drugs.

The main statistics computed over all the simulations for the training sets are reported in Table 1. Specifically, the results obtained by Hadjipavlou-Litina and Hansch, as well as the results obtained by the null model, i.e., the model in which the expected mean value of the target is used to perform the prediction, are reported in the first and second rows, respectively. For each data set, statistics on the number of inserted hidden units are reported for the CC for structures network. The mean absolute error (mean abs error), the correlation coefficient (R), and the standard deviation of error (S), as defined in regression analysis, are reported in the last three columns, respectively. Note that the mean abs error, R, and S for the CC for structures are obtained by averaging over the performed trials (six trials); the minimum and maximum values of the mean absolute error over these six trials are reported as well.

The results for the corresponding test sets are reported in Table 2. In the case of small test data sets the correlation coefficient is not meaningful so we prefer to report the maximum absolute error for the test data (max abs error), calculated as the average over the six trials, and the corresponding minimum and maximum values of the maximum absolute error obtained for each trial.

In Figures 6 and 7 we have plotted the error of the network versus the desired target for data sets I and III. Moreover, for the sake of comparison, in Figure 6 the error obtained using an equational approach[21] on data set I is reported as well.

Each point referring to the neural network models in the plots represents the average error, together with the deviation range, as computed over the six trials (i.e., the extremes of the deviation range correspond to the minimum and maximum output values computed over the six trials for each compound).

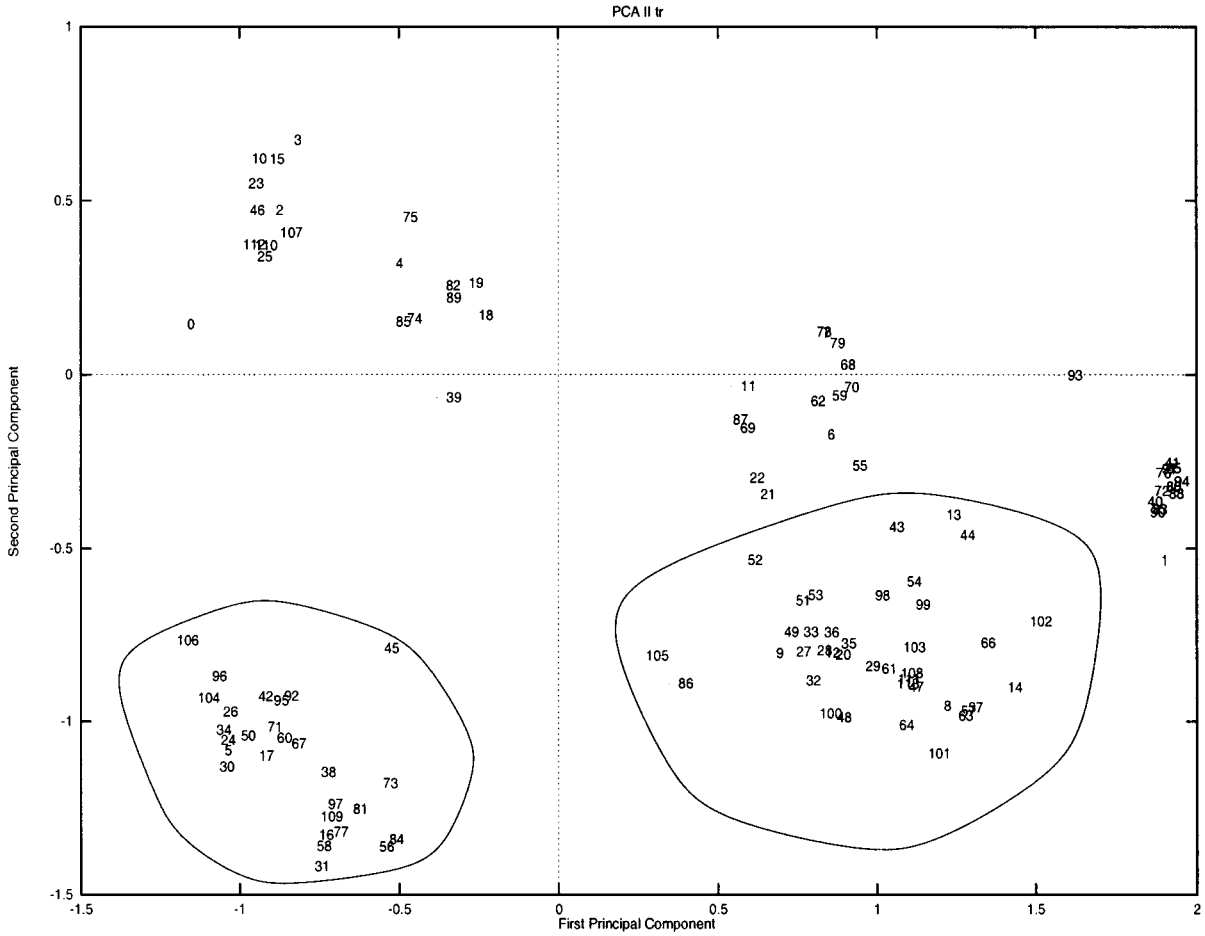

Figure 11. Principal component plot of training compounds used in experiment II derived from 30 output values of hidden neurons. Compounds characterized by $R_1 = H$ (left lower side of the plot) and compounds bearing a substituent at position 1 (right lower side of the plot) are grouped by contour lines. See Table 3 in the Appendix for compound numbering.

Regarding the evaluation of the performance of the proposed model for the treatment of benzodiazepines, from the comparison with the results obtained by the traditional equational treatment, we can observe a strong improvement in the fitting of the molecules included both in the training set and in the test set. The experimental results suggest a significant improvement over traditional QSAR techniques. Good results were obtained also for data set III, where the worse predicted compound is the one bearing hydrogen atoms in place of substituents which play an important role in determining affinity (compound no. 1ts in Table 5 in the Appendix). Finally, the soundness of the proposed model was confirmed by the experimental results obtained for data set IV, where the only compound which showed the maximum variance through the trials (compound no. 113 in Table 3 in the Appendix) contains a naphthyl group as the C ring, which never occurs in the training set. This explains the high variance observed in the prediction.

To complete the model evaluation, we present in the following a brief analysis of the learning behavior, especially considering the problem of overfitting, by discussing some learning curves for the recursive neural networks.

Typically in the CC for structures algorithm, as well as in other neural networks, the training error decreases as a function of the number of hidden units. This is related to the number of free parameters, which increases with the number of hidden units. If the number of parameters is too high with respect to the complexity of the training set, the network tends to overfit the data, basically learning the peculiar regularities of the specific training set instead of general domain knowledge. In this case, overfitting can be easily recognized by screening the performance of the network on a test set during learning: initially the test error decreases along with the training error, till it reaches a minimum, and then it starts to increase, while the training error keeps decreasing. The increase in the test error indicates that the network is starting to learn regularities in the training set which are not of general validity, thus showing overfitting.

A very common approach to avoid overfitting is to stop training as soon as the minimum on the test error is reached. This approach, however, implies the availability of enough data for both the training and test sets. Since in our case we do not have enough data, we adopted a specific strategy, called the *i-strategy*,[23] to avoid overfitting. The plot in Figure 8 shows a typical training session for the CC for structures. From the plot it is clear that adding new hidden units does not bring overfitting.

Basically, the i-strategy can be understood as an incremental strategy on the number of training epochs for each new inserted hidden unit. This is done because allowing few epochs to the first units avoids the increase of the weight values and the subsequent saturation of the units. On the other hand, units introduced late, which work with small

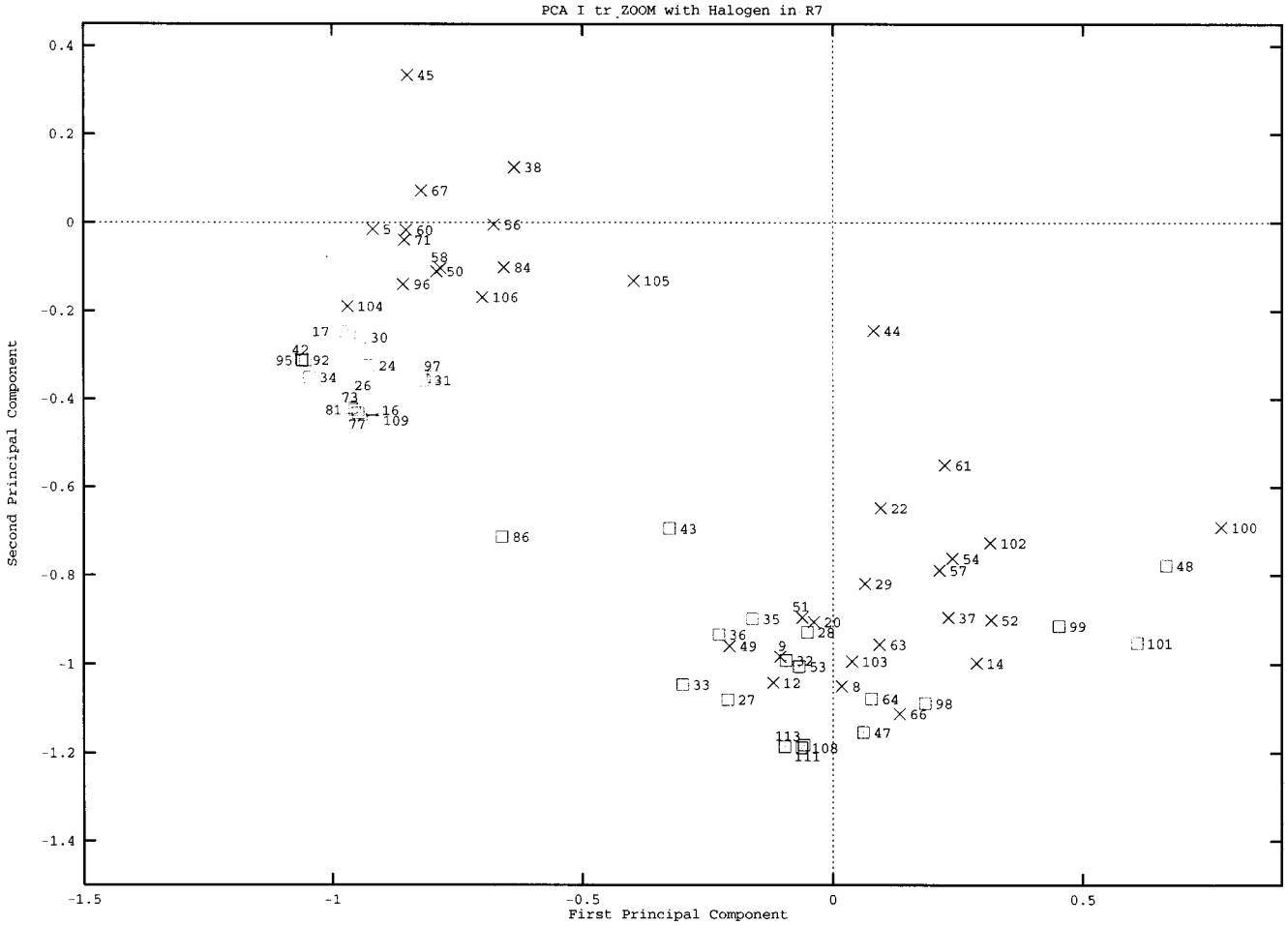

Figure 12. Zoom of circled areas of the plot reported in Figure 10. Compounds characterized by R_7 = halogen are marked by boxes; compounds where R_7 is not a halogen are marked by times signs. Compounds bearing a halogen atom at position 7 appear to be located at the (left) lower side of each group.

gradients due to the reduction of the residual error, take an advantage from the increased number of epochs.

In Figure 9, plot a shows typical learning curves for the training set with and without adopting the i-strategy. The training error with the i-strategy is higher than the training error without the i-strategy for the first inserted hidden units; however, with the increase in the number of hidden units, this relationship is inverted. Moreover, plot b in Figure 9, which reports learning curves for the test set, clearly shows that overfitting does occur when training does not use the i-strategy, while it does not occur when training uses the i-strategy. The global result is that, using the i-strategy, the better training error shown in Figure 9a is combined with a better generalization performance.

Concerning the other learning parameters, an initial set of preliminary trials were performed to determine an admissible range. However, no effort was done to optimize these parameters.

B. Internal Representations and Domain Knowledge. In the following, to understand whether the proposed model is able to capture relevant domain knowledge from the training data, we investigate the internal representations, i.e., the output of hidden units, developed by the neural network trained with the selected set of benzodiazepines.

The outputs of hidden units correspond to the encoding values generated for each compound or molecular fragments in the data set. Some of these fragments exactly correspond to the substituents attached to the main common template; other fragments are part of the substituents and do not have any chemical meaning.

Since the information about the morphological characteristics of the chemical compounds is directly given in input to the model as labeled trees, it is possible to perform a direct analysis of the computed values for these numerical codes associated with each compound and its subcomponents. For this investigation, due to the relatively large dimensionality of the representational space (typically around 30 hidden units are inserted by the training algorithm), we performed a PCA of the internal representations and studied 2-D plots of the first two principal components. The aim was to show, in a first approximation, the relative distance and position of internal representations and how they cluster within the representational space of the model. We expect the configurations of the points in the plots to approximately describe the knowledge learned by the neural network from the training data.

In the following we briefly recall some knowledge about the QSAR task. This exposition will be useful to evaluate how much of this knowledge the proposed model and training algorithm were able to capture.

From previous SAR studies, the presence or absence of substituents at some particular positions on the benzodiaz-

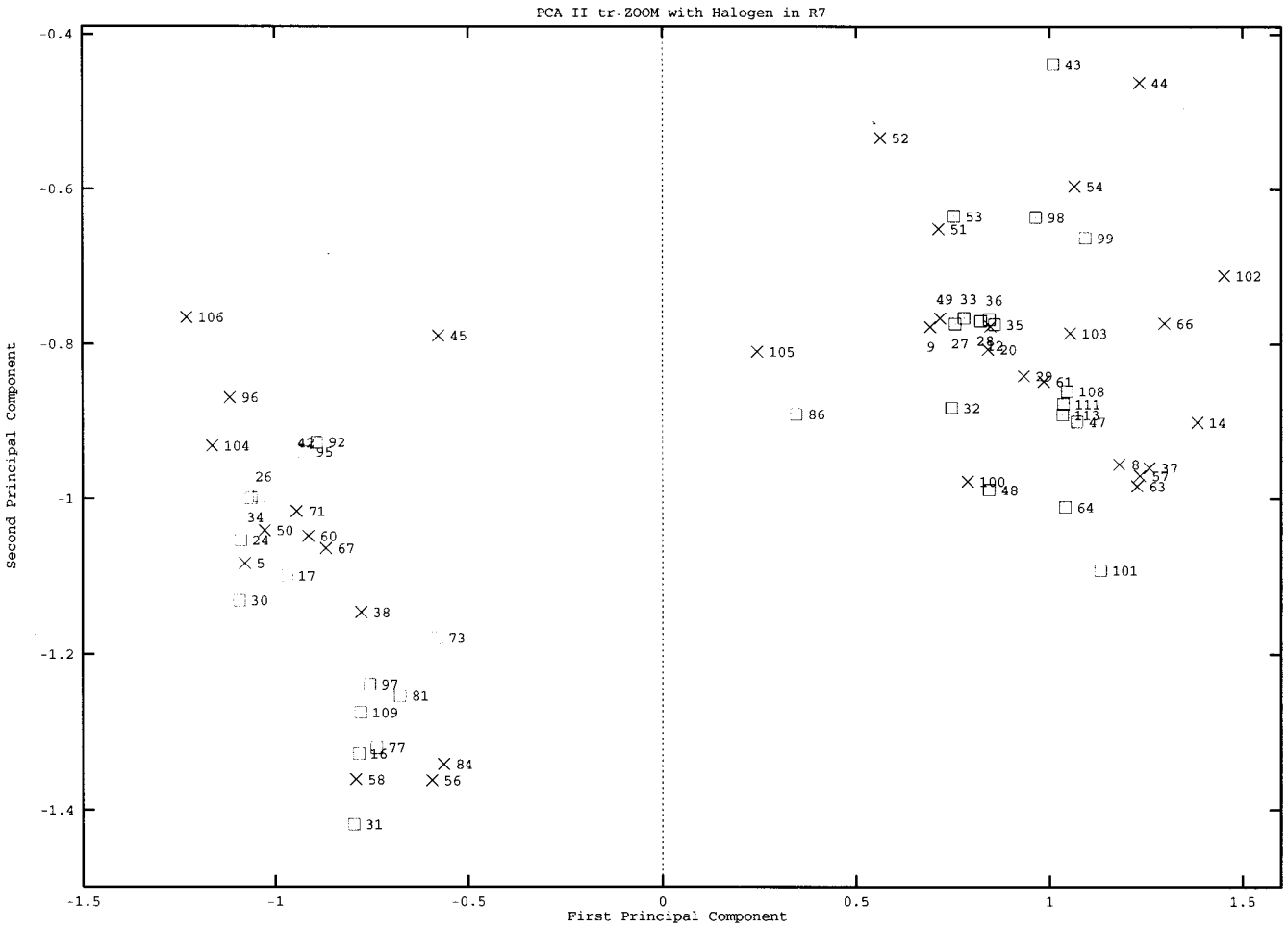

Figure 13. Zoom of circled areas of the plot reported in Figure 11. Compounds characterized by R_7 = halogen are marked by boxes; compounds where R_7 is not a halogen are marked by times signs.

epine nucleus are known to increase the affinity toward the receptor and/or affect the overall biological activity of benzodiazepine derivatives. Some of the SAR aspects that will be useful in the discussion of our results are summarized in ref 21. Specifically, a small lipophilic substituent at position 1 (B ring) is known to cause a moderate increase in efficacy (the methyl group seems to be optimal). The importance of electron-withdrawing substituents (such as Cl, Br, NO_2, and CF_3) at position 7 (A ring) was also pointed out since the earliest "in vivo" studies. Finally a single substitution at position 2′, or a double substitution at positions 2′ and 6′ on the C ring, by small-sized halogen atoms (F, Cl), strongly increases the activity. Instead, substitutions at positions 6, 8, and 9 of the A ring induce strong loss of activity. Within the class of compounds analyzed, the above-mentioned positions appear to be widely sampled. To rationalize the results, we need to establish a simple classification of the substituents. All the substituents were so classified on the basis of the effect produced on the electronic arrangement of the benzodiazepine nucleus. The effect that the substituents produce on the reactions of electrophilic substitution in aromatic compounds was chosen as the descriptor of the relevant molecular property. The above reactions constitute an important class in the organic chemistry area that can be taken as a probe for an evaluation of the electronic arrangement of the aromatic molecular regions involved in the reactions themselves.

The molecular descriptor so identified reflects substituent properties that are quite significant even from the point of view of the drug−receptor interaction. In fact this interaction is affected either by possible intermolecular hydrogen bonds or by possible hydrophobic interactions. The electronic effect of the substituent plays a significant role in all these interactions, mainly in molecules containing wide aromatic regions like benzodiazepines do.

The atoms or atomic groups at position 7 were so classified according to the following: (i) hydrogen atom (H) (considered the neutral element in our scale); (ii) halogen atoms (F, Cl, Br, I) (deactivating and ortho−para directing substituents for the electrophilic substitution); (iii) atomic groups such as $-NO_2$, $-CN$, $-CHO$, $-COCH_3$ (acethyl group), and $-CF_3$ (deactivating and meta directing groups for the electrophilic substitution); (iv) atomic groups such as $-NH_2$, $-NHOH$, $-NHCONHCH_3$, $-CH_3$, and $-C_2H_5$ (activating and ortho−para directing groups for the electrophilic substitution).

Here the substituent effect is roughly classified by grouping all substituents into the above four classes. It may be considered the qualitative analogue of the molecular descriptor associated with the substituent electronic effect quantified by Hammett's (σ) constants and currently used in classical QSAR.

Our PCA plots were analyzed taking into account the above aspects.

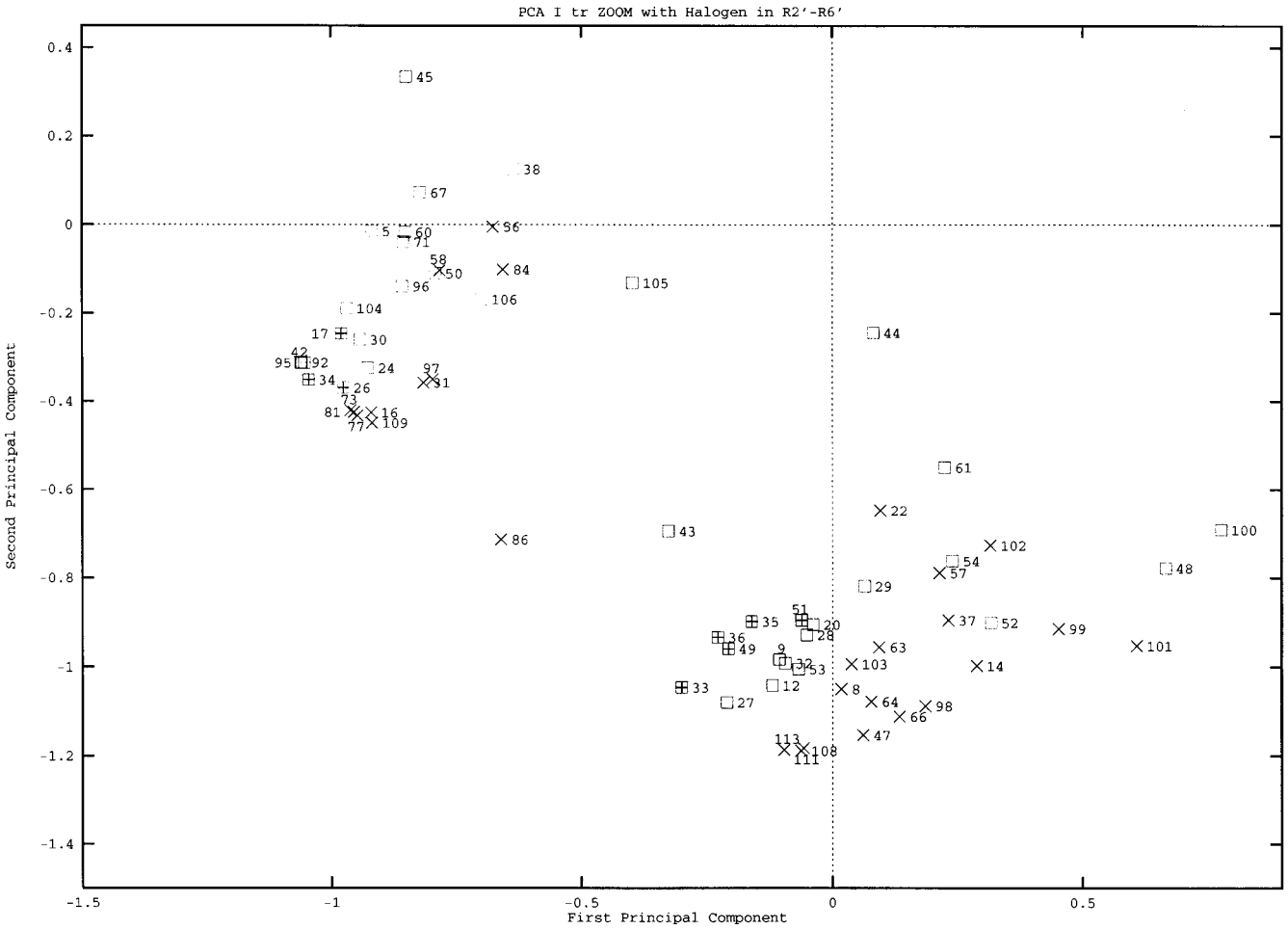

Figure 14. Zoom of circled areas of the plot reported in Figure 10. Compounds characterized by $R_{2'}$ = halogen are marked by boxes, compounds bearing halogen atoms at both positions 2′ and 6′ are marked by plus signs in boxes, and compounds where $R_{2'}$ and $R_{6'}$ are not halogens are marked by times signs. Compounds bearing halogen atoms at position 2′ or positions 2′ and 6′ appear to be located at the (left) upper side of each group.

C. Study of Internal Representations by PCA. The principal components of the internal representations developed by the CC for structures (outputs of recursive hidden neurons) were analyzed for all six experiments on data set III mentioned in the Model Evaluation subsection (IV.A.). Plots involving the first two principal components from two experiments are reported in Figures 10 and 11. They show the biologically active molecules analyzed (compounds associated with a target) and the relevant molecular fragments.

From the plots it can be seen that molecules and fragments are clustered on the basis of structural differences directly appearing at a simple observation of the molecular morphology. Furthermore, the molecules are grouped on the basis of some different chemical features that cannot be inferred directly by the observation of the molecular graph, rather only by the association of molecular structures and targets.

The plot obtained on the basis of experiment I (see Figure 10) appears to be split into two big clusters: all the substituents or molecular fragments approximately fall into its triangular upper right side, while all compounds to which a target is associated (molecules) approximately fall into its triangular lower left side; the plot obtained on the basis of experiment II (see Figure 11) appears to be split into two big groups as well, although with slight differences.

The group containing compounds associated with a target is divided, in turn, into two subgroups, highlighted in the plot shown in Figure 10 by contour lines. On the left side we find all the molecules bearing a methyl substituent or other alkyl groups at position 1 of the benzodiazepine nucleus (the alkyl groups may be substituted in turn and may show bigger steric hindrance and/or different chemical features). In a central region of the plot we find all the molecules that bear no substituents at position 1. The above distribution may be considered quite significant, as position 1 plays an important role in the structure−activity relationships of benzodiazepines, as mentioned above.

On the right side of the plot shown in Figure 10 we find a little subgroup containing the three molecules characterized by a different root of the tree. In these compounds the A ring of the benzodiazepine nucleus is a thienyl instead of a phenyl group, which is present in all the remaining molecules.

A similar plot for PCA of the internal representations developed in experiment II is reported in Figure 11, where the cluster of compounds bearing a substituent at position 1 is located at the lower left corner.

Both the biggest clusters containing molecules are divided in turn into smaller quite homogeneous subclusters on the basis of the possible presence of substituents at the other

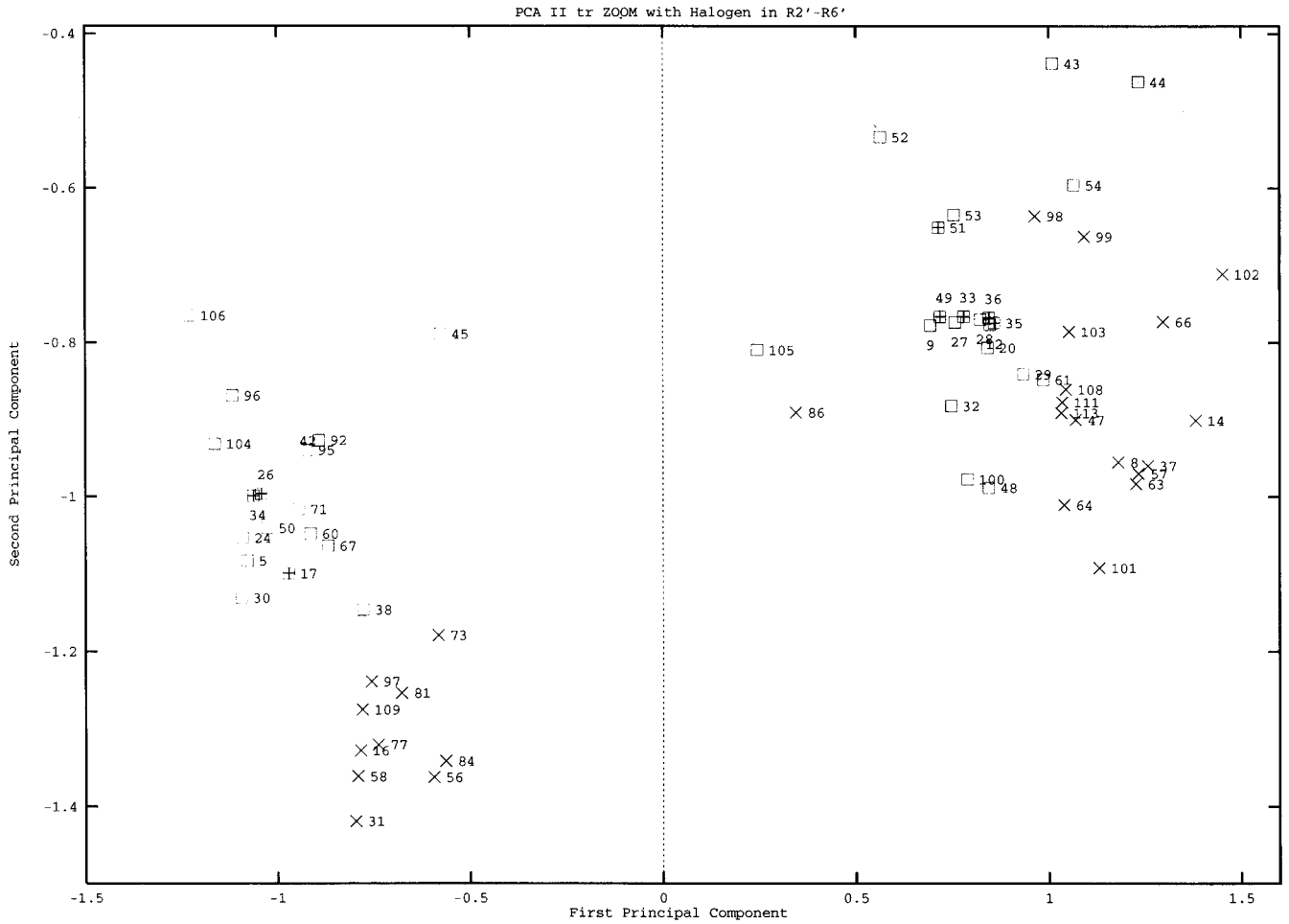

Figure 15. Zoom of circled areas of the plot reported in Figure 11. Compounds where $R_{2'}$ = halogen are marked by boxes, compounds bearing halogen atoms at both positions 2′ and 6′ are marked by plus signs in boxes, and compounds where $R_{2'}$ and $R_{6'}$ are not halogens are marked by times signs. Compounds bearing halogen atoms at position 2′ or positions 2′ and 6′ appear to be located at the left upper side of each group.

significant positions of the benzodiazepine nucleus previously mentioned.

In the plot shown in Figure 12 we observe that each one of the two big clusters identified in the previous plots is subclustered on the basis of which kind of atom or atomic group is present at position 7. Compounds characterized by the presence of a halogen atom at position 7 are marked by little boxes, while little crosses are used to mark the remaining compounds. The subgroups so identified only partially overlap; mostly it is possible to find regions of the plot where molecules characterized by one or another kind of substituent prevail. The corresponding plot for results obtained from experiment II is reported in Figure 13.

Position 2′ (substitution on the C ring of the benzodiazepine nucleus) and, in the case of double substitution, position 6′, symmetrical to position 2′ with respect to the 2-fold axis in the C ring, represent further key positions for the affinity of benzodiazepines.

The plots shown in Figures 14 and 15 allow us to focus the analysis on the presence and the type of substituent at position 2′ and positions 2′−6′: once again quite homogeneous subgroups were found. The subgroups appear only slightly overlapping in the case of experiment I, while they appear quite well defined in the case of experiment II. Compounds characterized by the presence of only one halogen at position 2′ are marked by boxes, and compounds characterized by the simultaneous presence of halogens at positions 2′ and 6′ are marked by plus signs within boxes.

Finally substitutions at positions 6, 8, and 9 were analyzed in data from experiment I. Molecules characterized by substituents at these positions (including a few cases of simultaneous/multiple substitutions), even poorly sampled, are divided by PCA into subgroups still showing a certain degree of homogeneity. Molecules bearing a substituent at position 9 always fall on the right side of each subgroup for data obtained from experiment I. Molecules bearing an activating and ortho−para directing substituent, at any one of the above positions, fall on the right side of the sub-group with respect to molecules bearing, at the same positions, halogen atoms which are deactivating and ortho−para directing substituents.

It may be noteworthy to observe that, in all six experiments, analogous types of clustering are found: all the molecules are homogeneously clustered on the basis of the substituent effects. The differences in analogous plots showing the results obtained from distinct experiments only consist of rotations and/or translations of the clusters with respect to each other, as we can observe from the comparison of plots reported in Figures 10 and 11. This is partially due

Table 3. Training Data Set III

no.	R_1	R_3/R_6	R_7	R_8/R_9	$R_{2'}$	$R_{6'}$	log(1/*C*)	mean output	min ouput	max output
5	$-CH_3$		$-CN$		$-F$		7.52	7.331	7.212	7.443
8			$-CH=CH_2$				7.62	7.423	7.294	7.487
9					$-F$		7.68	7.646	7.482	7.770
12			$-COCH_3$		$-F$		7.74	7.827	7.690	7.946
14			$-CF_3$				7.89	7.902	7.801	7.980
16	$-CH_3$		$-Cl$				8.09	7.933	7.854	8.015
17	$-CH_3$		$-Cl$		$-Cl$	$-Cl$	8.26	8.234	8.140	8.282
20			$-N_3$		$-F$		8.27	8.282	8.209	8.349
22			$-NO_2$		$-CF_3$		8.45	8.437	8.353	8.523
24	$-CH_3$		$-I$		$-F$		8.54	8.633	8.569	8.691
26	$-CH_3$		$-Br$		$-F$	$-F$	8.62	8.588	8.535	8.644
27			$-Cl$		$-F$		8.70	8.748	8.614	8.877
28			$-Cl$		$-Cl$		8.74	8.754	8.684	8.888
29			$-NO_2$		$-F$		8.82	8.734	8.593	8.948
30	$-CH_3$		$-F$		$-F$		8.29	8.266	8.183	8.342
31	$-CH_3$		$-F$				7.77	7.685	7.529	7.765
32			$-F$		$-F$		8.13	8.218	8.028	8.429
33			$-Cl$		$-F$	$-F$	8.79	8.734	8.550	8.893
34	$-CH_3$		$-Cl$		$-F$	$-F$	8.39	8.406	8.247	8.448
35			$-Cl$		$-Cl$	$-F$	8.52	8.621	8.489	8.744
36			$-Cl$		$-Cl$	$-Cl$	8.15	8.146	8.060	8.338
37			$-NO_2$				7.99	7.943	7.783	8.084
38	$-CH_3$		$-NO_2$		$-Cl$		8.66	8.637	8.517	8.784
42	$-CH_2CH_2OH$		$-Cl$		$-F$		7.61	7.337	7.288	7.398
43		$R_3 = -(s)CH_3$	$-Cl$		$-F$		8.46	8.419	8.329	8.527
44		$R_3 = -(s)CH_3$	$-NO_2$		$-Cl$		8.92	8.916	8.856	9.004
45	$-CH_3$	$R_3 = -(s)CH_3$	$-NO_2$		$-F$		8.15	8.154	8.053	8.295
47*			$-Br$				7.74	7.720	7.623	7.807
48			$-Cl$		$-Cl$		8.03	8.006	7.913	8.057
49					$-F$	$-F$	7.72	7.714	7.620	7.825
50	$-CH_3$				$-Cl$		8.42	8.426	8.328	8.549
51				$R_8 = -Cl$	$-F$	$-F$	7.55	7.566	7.518	7.680
52				$R_8 = -CH_3$	$-F$		7.72	7.683	7.552	7.761

to the different projections of the principal components in a two-dimensional space.

In this regard it has to be pointed out that the substituent effects on the target molecular properties (e.g., affinity) combine with each other in very complex ways. Nevertheless, the well-defined clustering observed in most of the PCA plots suggests that each single effect may be easily extracted by the model, in its different realizations corresponding to the six experiments, offering a quite direct analysis of the structure−property relationships. The analogies found in PCA of different experiments appear to be particularly significant. It shows that the capability of the model in extracting structural features which are significant for the target correlation is quite independent from the different realizations of the model itself.

V. CONCLUSIONS

With regard to the performance of the proposed model, we observed a noticeable improvement for the QSAR task in comparison to the results obtained by the traditional Hansch equation-based model. Although the improvement in performance can be explained by the use of a nonlinear model, and doubts can be raised on the necessity of using the proposed model instead of a standard neural network, we observe that our model allowed us to study directly the correlation between the morphological characteristics of the chemical compounds and the biological activity of interest. This correlation could be studied only indirectly by using a standard neural network, since standard neural networks require physicochemical descriptors or a priori defined topological indices.

Moreover, the structural information supplied to the model by the DOAG representation is more direct and richer than that contained in physicochemical descriptors or in topological indices, usually exploited in most of the QSAR models, based either on equations or on neural networks. The use of these last two types of molecular descriptors does not ensure that all the significant structural features are included in the analysis, while the representation of the molecular structure proposed here seems to answer in an optimal way most of the already mentioned typical QSAR problems. Due to the possibility of directly processing chemical structures, the model appears to be a powerful tool as it is able to consider all the meaningful elements for the identification of the structure−activity relationship. Thus, it allows one to avoid one of the main problems encountered in QSAR, i.e., the identification of a proper set of molecular descriptors that is simultaneously complete and nonredundant.

Flexibility in representing the chemical compounds is another interesting feature of the proposed approach. In fact, the possibility of explicity representing only selected atoms allows the user to adjust the amount of structural information supplied to the model in accordance with the problem at hand, and so to optimize the use of computational resources. It also allows the user to keep information about atom types and about atom connectivities at the desired level of detail (through the DOAG representation). Any other kind of already known 2-D QSAR model does not allow that in a comparable amount. On the other hand, the use of DOAGs as molecular structure descriptors does not supply any 3-D structural information (considered, instead, in the 3-D QSAR models), which may be deceptive in all these cases in which

Table 4. Training Data Set III Continued

no.	R_1	R_3/R_6	R_7	R_8/R_9	$R_{2'}$	$R_{6'}$	log(1/*C*)	mean output	min output	max output
53			−Cl	$R_8 = -Cl$	−F		8.44	8.479	8.431	8.593
54			$-CH_3$	$R_8 = -Cl$	−F		7.85	7.824	7.777	7.886
56	$-CH_3$		$-NH_2$				6.34	6.309	6.197	6.437
57			$-NH_2$				6.41	6.519	6.412	6.651
58	$-CH_3$		−CN				6.42	6.657	6.588	6.757
60	$-CH_3$		−NHOH		−F		7.02	6.920	6.706	7.082
61			$-NH_2$		−Cl		7.12	7.062	6.968	7.106
63			−CHO				7.37	7.514	7.306	7.641
64			−F				7.40	7.434	7.286	7.602
66			$-C_2H_5$				7.44	7.457	7.378	7.553
67	$-CH_3$		$-NH_2$		−F		7.19	7.218	7.040	7.384
71	$-CH_3$		$-NHCONHCH_3$		−F		6.34	6.492	6.358	6.685
73	$-CH_2-CF_3$		−Cl				7.04	6.973	6.876	7.121
77	$-CH_2-C\equiv CH$		−Cl				7.03	7.023	6.854	7.141
81	$-CH_2C_3H_5$		−Cl				6.96	7.004	6.850	7.129
84	$-CH_2OCH_3$		$-NO_2$				6.37	6.315	6.176	6.469
86	$-C(CH_3)_3$		−Cl				6.21	6.208	6.112	6.303
92	$-(CH_2)_2OCH_2CONH_2$		−Cl		−F		7.37	7.310	7.289	7.348
95	$-CH_2CHOHCH_2OH$		−Cl		−F		6.85	7.197	7.147	7.240
96	$-CH_3$	$R_6 = -Cl$		$R_8 = -Cl$	−F		6.52	6.491	6.406	6.568
97	$-CH_3$		−Cl	$R_8 = -Cl$			7.40	7.447	7.378	7.516
98			−Cl	$R_9 = -Cl$			7.43	7.431	7.359	7.595
99			−Cl	$R_9 = -CH_3$			7.28	7.269	7.184	7.343
100					−Cl		7.43	7.423	7.362	7.531
101			−Cl				7.15	7.157	7.082	7.265
102		$R_6 = -CH_3$	$-CH_3$				6.77	6.773	6.754	6.808
103		$R_6 = -Cl$					6.49	6.488	6.423	6.550
104	$-CH_3$	$R_6 = -Cl$			−F		6.82	6.878	6.812	6.971
105	$-C(CH_3)_3$		$-NO_2$		−Cl		6.52	6.607	6.493	6.687
106	$-CH_3$			$R_9 = -Cl$	−F		7.14	7.133	7.034	7.198
108*			−Cl				7.47	7.367	7.264	7.428
109*	$-CH_3$		−Cl				7.47	7.552	7.498	7.599
111*			−Cl				7.06	6.988	6.900	7.089
113*			−Cl				6.54	6.622	6.506	6.669

Table 5. Test Data Set III

no.	R_1	R_3/R_6	R_7	R_8/R_9	$R_{2'}$	$R_{6'}$	log(1/*C*)	mean output	min output	max output
1ts							6.45	7.030	6.739	7.273
6ts	$-CH_3$		$-NO_2$		−F		8.42	8.207	7.910	8.428
8ts			$-NO_2$		−Cl		8.74	8.655	8.472	9.067
9ts			−Cl				8.03	7.845	7.597	8.054
10ts	$-CH_3$				−F		7.85	7.897	7.708	7.996

Table 6. Racemic Compounds for Data Set I

R_1	R_3/R_6	R_7	R_8/R_9	$R_{2'}$	$R_{6'}$	log(1/*C*)
$-CH_3$	$R_3 = -(rac)CH_{3-}$	−Cl				7.31
	$R_3 = -(rac)OH$	−Cl				7.74
$-CH_3$	$R_3 = -(rac)OH$	−Cl				7.79
$-CH_3$	$R_3 = -(rac)Cl$	−Cl		−F		8.27
$-CH_3$	$R_3 = -(rac)OCON(CH_3)_2$	−Cl				6.05

the biologically active conformation of the molecules cannot be correctly guessed.

The ability of recursive neural networks to automatically discover useful numerical representations of the input structures at the hidden layer is the key feature of the adaptive solution to the QSAR task. By analyzing these representations through PCA, as expected, we found that the global distribution of molecules and fragments in the plots of the two first principal components reflects the expected capability of the model in detecting homogeneous structural features that can be directly observed on the basis of the molecular morphology. However, the most remarkable aspect is that the distribution reflects its ability in detecting the similar characteristics of the substituents not directly related to the molecular morphology, such as electronic effects produced by halogen atoms. It has to be recalled here that halogen atoms are represented and distinguished, with respect to each other, only by four different labels, which do not contain any evident information regarding their very homogeneous electronic properties.

In this regard the analysis of the principal components shows that the neural network used here for QSAR studies is capable of capturing in most cases the physicochemical meaning of the above-mentioned substituents even when the use of different labels does not allow a direct grouping of substituents into chemically homogeneous classes.

Globally, we can observe that the characteristics of many substituents affecting the activity of benzodiazepines, already highlighted by previous QSAR studies, were correctly recognized by the model; i.e., the numerical code developed by the recursive neural network is effectively related to the qualitative aspect of the QSAR problem.

We can conclude that, although the method presented here is at an early stage of its development, the proposed neural network is able to supply a well-suited tool for QSAR

analysis. This evaluation is supported both by its performance in comparison to other models previously used for the analysis of the same class of molecules and by the results of PCA.

VI. APPENDIX

In Tables 3−5, the training and test sets for benzodiazepine data used in data set III are reported.

Since the numbering refers both to the analyzed compounds and to the fragments generated by the preprocessing, the set of numbers associated with the molecules reported in the tables is not complete (only the compounds are reported).

Note that the C ring, located at position 5, is a phenyl group in all the analyzed compounds except in compounds 47, 108, 109, 111, and 113, where it is replaced by 2-pyridyl, cyclohexenyl, cyclohexenyl, cyclohexyl, and naphthyl, respectively (marked by an asterisk in Tables 3 and 4).

Table 6 reports the racemic compounds used in data set I.

REFERENCES AND NOTES

(1) Hansch, C.; Maloney, P.; Fujita, T.; Muir, R. M. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature* **1962**, *194*, 178−180.

(2) Hansch, C.; Fujita, T. Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(3) Free, S. M., Jr.; Wilson, J. W. A mathematical contribution to structure−activity studies. *J. Med. Chem.* **1964**, *7*, 395−399.

(4) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Modeling. *Reviews in Computational Chemistry*; VCH Publishers: New York, 1991; Chapter 9, pp 367−422.

(5) Rouvray, D. H. Should we have designs on topological indices? In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier Science Publishing Co.: New York, 1983; pp 159−177.

(6) Magnuson, V. R.; Harris, D. K.; Basak, S. C. Topological indices based on neighborhood symmetry: Chemical and biological application. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier Science Publishing Co.: New York, 1983; pp 178−191.

(7) Barysz, M.; Jashari, G.; Lall, R. S.; Srivastava, V. K.; Trinajstic, N. On the distance matrix of molecules containing heteroatoms. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier Science Publishing Co.: New York, 1983; pp 222−230.

(8) Devillers, J., Ed. *Neural Networks in QSAR and Drug Design*; Academic Press: London, 1996.

(9) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: an introduction*; VCH Publishers: New York, 1993.

(10) Burns, J. A.; Whitesides, G. M. Feed-forward neural networks in chemistry: Mathematical system for classification and pattern recognition. *Chem. Rev.* **1993**, *93* (8), 2583−2601.

(11) Suzuki, Y.; Aoyama, T.; Ichikawa, H. Neural networks applied to quantitative structure−activity relationships. *J. Med. Chem.* **1990**, *33*, 2583−2590.

(12) Ajay. A unified framework for using neural networks to build QSARs. *J. Med. Chem.* **1993**, *36*, 3565−3571.

(13) Peterson, Keith L. Quantitative structure−activity relationships in carboquinones and benzodiazepines using counter-propagation neural networks. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (5), 896−904.

(14) Duprat, A. F.; Huynh, T.; Dreyfus, G. Towards a Principled Methodology for Neural Network Design and Performance Evaluation in QSAR; Application to the Prediction of LogP. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 854−866.

(15) Liu, S.; Zhang, R.; Liu, M.; Hu, Z. Neural networks-topological indices approach to the prediction of properties of alkene. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1146−1151.

(16) Cherqaoui, D.; Villemin, D. Use of neural network to determine the boiling point of alkanes. *J. Chem. Soc., Faraday Trans.* **1994**, *90* (1), 97−102.

(17) Elrod, D. W. Maggiora, G. M.; Trenary, R. G. Application of neural networks in chemistry. 1. prediction of electrophilic aromatic substitution reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 447−484.

(18) Kvasnička, V.; Pospichal, J. Application of neural networks in chemistry. Prediction of product distribution of nitration in a series of monosubstituted benzenes. *J. Mol. Struct.: THEOCHEM* **1991**, *235*, 227−242.

(19) Sperduti, A.; A. Starita. Supervised neural networks for the classification of structures. *IEEE Trans. Neural Networks* **1997**, *8* (3), 714−735.

(20) Sperduti, A. Majidi, D.; Starita, A. Extended cascade-correlation for syntactic and structural pattern recognition. In *Advances in Structural and Syntactical Pattern Recognition*; Perner, P., Wang, P., Rosenfeld, A., Eds.; Lecture notes in Computer Science Vol. 1121; Springer-Verlag: Berlin, 1996; pp 90−99.

(21) Hadjipavlou-Litina, Dimitra; Hansch, Corwin Quantitative Structure−Activity Relationships of the benzodiazepines. A review and reevaluation. *Chem. Rev.* **1994**, *94* (6), 1483−1505.

(22) Bianucci, A. M.; Micheli, A.; Sperduti, A.; Starita, A. Quantitative structure−activity relationships of benzodiazepines by recursive cascade correlation. *IEEE International Joint Conference on Neural Networks*; 1998; pp 117−122.

(23) Bianucci, A. M.; Micheli, A.; Sperduti, A.; Starita, A. Application of cascade correlation networks for structures to chemistry. *J. Appl. Intell.* **2000**, *12*, 117−147.

(24) Jollife, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.

(25) Fahlman, S. E.; Lebiere, C. The cascade-correlation learning architecture. In *Advances in Neural Information Processing Systems 2*; Touretzky, D. S., Ed.; Morgan Kaufmann Publishers: San Mateo, CA, 1990; pp 524−532.

(26) Fahlman, S. E. The recurrent cascade-correlation architecture. In *Advances in Neural Information Processing Systems 3*; Lippmann, R. P., Moody, J. E., Touretzky, D. S., Eds.; Morgan Kaufmann Publishers: San Mateo, CA, pp 190−196.

(27) Sternbach, L. H. The benzodiazepine story. *J. Med. Chem.* **1979**, *22*, 1−7.

(28) Evans, B. E.; Rittle, K. E.; Whitter, W. L.; Veber, D. E.; Anderson, P. S.; Freidinger, R. M.; Bock, M. G.; Dipardo, R. M. Benzodiazepine gastrin and brain cholecystochinin receptor ligands. *J. Med. Chem.* **1989**, *32*, 13−16.

(29) Lenox, R. H.; Kornecki, E.; Ehrlich, Y. H. Platelet-activating factor-induced aggregation of human platelet specifically inhibited by triazolo benzodiazepines. *Science* **1984**, *226*, 1454−1456.

(30) Hooly, M.; Slice, L.W.; Sherman, M. I.; Richman, D. D.; Potash, M. J.; Volsky, D. J.; Hsu, M. C.; Schutt, A. D. Inhibition of HIV replication in acute and chronic infections in vitro a tat antagonist. *Science* **1991**, *254*, 1799−1802.

(31) Desmyter, J.; Schols, D.; Kukla, M. J.; Breslin, H. J.; Raeymaekers, A.; Van Gelder, J.; Woestenborghs, R.; Heykants, J.; Schellekens, K.; Janssen, M. A. C.; Clercq, E. D.; Janssen, P. A. J.; Pauwels, R.; Andries, K. Potent and selective inhibition of hiv-1 replication in vitro by a novel series of tetrahydro imidazo[4,5,1-jk][1,4]-benzodiazepin-2(1h)-one and thione (tibo) derivatives. *Nature* **1990**, *343*, 470−474.

(32) Fryer, R. I. In *Comprehensive Medicinal Chemistry: Ligand Interactions at the Benzodiazepines Receptor*; Ramsden, C. A., Ed.; Pergamon Press: New York, 1990; Chapter 12.8, p 539.

(33) Fryer, R. I. *The Benzodiazepines: From Molecular Biology to Clinical practice*; Raven Press: New York, 1983; p 7.

CI9903399