A Novel Approach to QSPR/QSAR Based on Neural Networks for Structures [*]

Anna Maria Bianucci[1], Alessio Micheli[2], Alessandro Sperduti[2], and Antonina Starita[2]

[1] Dipartimento di Scienze Farmaceutiche, Università di Pisa, Via Bonanno 6, 56126, Pisa, Italy
[2] Dipartimento di Informatica, Università di Pisa, Corso Italia, 40, 56125 Pisa, Italy

Abstract. We present a novel approach based on neural networks for structures to QSPR (quantitative structure-property relationships) and QSAR (quantitative structure-activity relationships) analysis. We face two quite different chemical applications using the same model, i.e. predicting the boiling point of a class of alkanes and QSAR of a class of benzodiazepines. The model, Cascade Correlation for structures, is a class of recursive neural networks recently proposed for the processing of structured domains. Through the use of this model we can represent and process the structure of chemical compounds in the form of labeled trees. We report our results on preliminary applications to show that the model, adopting this representation of molecular structure, can adaptively capture significant topological aspects and chemical functionalities for each specific class of the molecules, just on the basis of the association between the molecular morphology and the target property.

1 Introduction

The possibility of relating some significant aspects of molecular structures to any particular behaviour 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. For instance, 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. For this reason the achievement of good predictive models constitutes a big task for both the basic and the applied research.

Many mathematical models were developed in the past years 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 non-direct correlation of the molecular structure to the target property. In these models some physico-chemical properties were used as molecular descriptors. They should be better classified as property-property or property-

[*] Partially supported by MURST grant 9903244848 and MM09308497.

activity relationships 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 physico-chemical properties [1–3], subsequently has been 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 remove some structural information which may be relevant. Moreover, the *a priori* definition of the encoding process has other several drawbacks. For example, when the encoding is performed by using *topological indexes*, they need to be properly designed by an *expert* through a very expensive *trial and error approach.* Thus this approach needs an expert, which may not be available, or may be very expensive, or even may be misleading if the expert knowledge is not correct. Finally, changing the class of chemical compounds under study, or the computational task, will of course mean that all the above steps must be performed from scratch. More general vectorial representation of graphs, with unicity properties, may be very difficult to map on the target values.

A completely different approach is possible facing directly the processing of structured domain in the machine learning systems. While algorithms that manipulate symbolic information are capable of dealing with highly structured data, they very often are not able to deal with noisy and incomplete data. Moreover, they are usually not suited to deal with domains where both categorical (symbols) and numerical entities coexist and have the same relevance for the solution of the problem.

Neural networks are universally recognized as tools suited for dealing with noisy and incomplete data, especially in contexts where numerical variables play a relevant role in the solution of the problem. In addition to this capability, when used for classification and/or prediction tasks, they do not need a formal specification of the problem, just requiring a set of examples showing samples of the function to be learned. Unfortunately, neural networks are mostly regarded as learning models for domains in which instances are organized into *static* data structures, like records or fixed-size arrays, and thus they do not seem suited to deal with structured domains. Recurrent neural networks, that generalize feedforward networks to sequences (a particular case of dynamically structured data) are perhaps the best known exception.

In recent years, however, there has been some effort in trying to extend the computational capabilities of neural networks to structured domains. While the earlier approaches were able to deal with some aspects of processing of structured information, none of them established a practical and efficient way of dealing with structured information. A more powerful approach, at least for classification and prediction tasks, was proposed in [8] and further extended in [9]. In these works a generalization of recurrent neural networks for processing sequences to the case of directed graphs is presented. The basic idea behind this generalization is the extension of the concept of *unfolding* from the domain of sequences to the domain of directed ordered graphs (DOGs). We will give more details on these types of neural networks for the class of directed ordered acyclic graphs (DOAGs) in Section 2.2.

The possibility of processing structured information using neural networks is appealing for several reasons. First of all, neural networks are universal approximators; in addition, they are able to learn from a set of examples and very often, by using the correct methodology for training, they are able to reach a quite high generalization performance. Finally, as already mentioned above, they are able to deal with noisy and incomplete, or even ambiguous, data.

All these capabilities are particularly useful when dealing with prediction tasks in Chemistry, where data are usually gathered experimentally and the compounds can naturally be represented as labeled graphs. Each node of the graph is an atom or a group of atoms, while edges represent bonds between atoms. So neural networks for processing of structures seem to have the computational capabilities to deal with chemical domains. The prediction model can face one fundamental problem in Chemistry: the prediction of the physical-chemical properties, chemical reactivity or biological activity of molecules, leading to *Quantitative Structure-Property Relationship* (QSPR), or *Quantitative Structure-Activity Relationship* (QSAR) studies. Recursive neural networks [8] face this problem by simultaneously learning both how to represent and how to classify structured patterns. The specificity of the proposed approach stems from the ability of these networks to automatically encode the structural information depending on the computational problem at hand, i.e., the representation of the molecular structures is not defined a priori, but learned on the basis of the training set. This ability is proved in this paper by the application of Cascade Correlation for structures (CCS) [8] to two radically different QSAR/QSPR problems: the prediction of the non-specific activity (affinity) towards the benzodiazepine/GABA$_A$ receptor by a group of benzodiazepines (Bz) [10], and the prediction of the boiling point for a group of acyclic hydrocarbons (alkanes)[11].

It must be stressed that the generality and flexibility of a structured representation, allows one to deal with heterogeneous compounds and heterogeneous problems using the same approaches. This advantage is not at the expense of predictive accuracy, in fact our results [12] [13] compare favorably versus the traditional QSAR treatment, for the analysis of benzodiazepines, based on equations [10]. It is also competitive with results on QSPR problems (such as, the prediction of the boiling point of alkanes) where the *a priori* analytical knowledge allows the use of suitable 'ad hoc' representations as input to standard neural networks [11].

Successive studies on the internal representation developed by the recursive neural networks (realized by a Cascade Correlation algorithm) applied to QSAR studies of benzodiazepines were conducted using principal component analysis [14]. This study allows us to deal with the issue of the relationship between the developed neural numerical codes and the qualitative aspects of the QSAR problem. The results show that the recursive neural network is able to discover relevant structural features just on the basis of the associations between the molecular morphology and the target property (affinity). In particular the characteristics of many substituents affecting the activity of benzodiazepines, already highlighted by previous QSAR studies, were correctly recognized by the model. This is a further step towards the assessment of the model as a new tool for the rational design of new molecules.

The chapter is organized as follows. Section 2 begins with an outline of the traditional QSPR/QSAR approach and it is followed by the introduction of the new QSPR/QSAR approach based on recursive neural networks. General representational issues for chemical compounds are discussed in Section 3. The first computational tasks faced in this paper, i.e., the prediction of the boiling point for alkanes, including representation rules and experimental results, is explained in Section 4. Similarly, the application to the QSAR problem of the prediction of the affinity towards the benzodiazepine/GABA$_A$ receptor is explained in Section 5, where we present also the study of the internal representations developed by the neural model through Principal Component Analysis. Discussion of the results and conclusions are contained in Section 6 and Section 7, respectively.

2 Recursive Neural Networks in QSPR/QSAR

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

Without loss of generality, for the sake of a simpler exposition and due to their relevance, we mainly focus the following explanations and examples on QSAR studies. However, thought QSPR deals with general properties instead of activity, the following considerations are valid both for QSPR and QSAR analysis.

2.1 Toward a New QSPR/QSAR Approach

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

$$Activity = \mathcal{T}(Structure). \tag{1}$$

More generally QSPR assumes that any molecular property, such as physical-chemical properties, can be related to the structure of the compounds, i.e.:

$$Property = \mathcal{T}(Structure). \tag{2}$$

The function $\mathcal{T} : \mathcal{I} \rightarrow \mathcal{O}$ is therefore a functional transduction from an input structured domain $\mathcal{I}$, where molecules are represented, to an output domain $\mathcal{O}$, such as the real number set. In equations 1 and 2 the term "structure" stresses the importance of the use of global information about molecular shape, atom connectivities and chemical functionalities as understood in the QSPR/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 non-linear regression tools (e.g., equational modeling, and feed-forward neural networks).

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

$$\mathcal{T}() = g(\tau()) \tag{3}$$

where $\tau()$ is the *encoding* function from the domain of the chemical structures to the descriptor space, while g is the *mapping* function from the descriptors space to the biological activity space. This corresponds to the traditional QSPR/QSAR approaches, as summarized in Fig. 1 for the QSAR, 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 physico-chemical or geometrical and electronic properties, the calculation of topological indices, or an explicit vector based representation of molecular connectivity (see the examples in Section 4.2). The question mark in the picture shown in Fig. 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

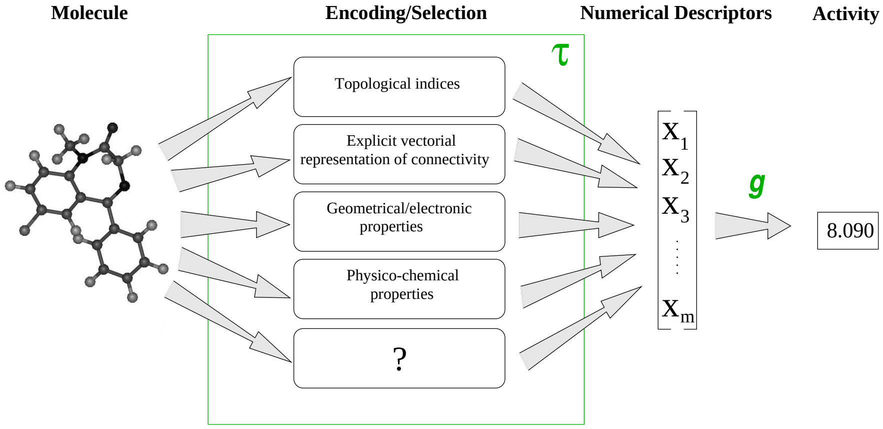
Molecule
Encoding/Selection
Numerical Descriptors
Activity
τ
Topological indices
Explicit vectorial representation of connectivity
Geometrical/electronic properties
Physico-chemical properties
?
X1
X2
X3
Xm
g
8.090

Fig. 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.

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

$$\tau() = \tau_E(\tau_R()), \tag{4}$$

where τ_R extracts a specific structural aspect 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.

We could sort the traditional approaches on the basis of the evolution toward the use of more direct representations of the molecular structures. Summarizing, we can mention models based on physico-chemical properties [15–18], that may be obtained as combinations of fragment contributions, on topological indices [19,11], or matricial [20] graph representations, and finally a template-based approach [21]. This last model uses a neural network which partially mimics the chemical structures of the analyzed compounds by means of a common molecular template, statically defined for all the compounds.

The mathematical and computational tools used in QSPR/QSAR approches are quite different from each other and include equation based models [1,2] and neural network based models [22–24].

However, in traditional QSPR/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 in order to adapt them to the regression problem required by the QSAR study. So, even if the chemical graph is clearly 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 chapter 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 [8], which are able to take directly as input the graph generated by τ_R and to implement adaptively both τ_E and g.

In order to exemplify the above concepts, in Fig. 2, we show the outline of the proposed approach assuming that a given molecule is represented by τ_R as a labeled tree[1]. 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 it 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 Fig. 1 and the proposed new scheme reported in Fig. 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 Fig. 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 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 it is equal to 1 if and only if the representation refers to that entity; otherwise it is set to -1; e.g., assuming that the fluorine atom (F) is associated with the i-th component and the chlorine atom (Cl) is

[1] The definition of an appropriate function τ_R for the specific set of molecules studied in this paper is discussed in Section 3.

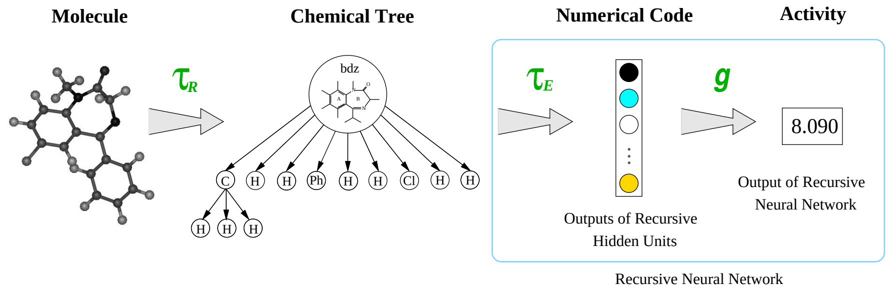

Fig. 2. The new QSAR scheme using recursive neural networks is shown: the molecule, after a structural coding phase driven by ad hoc rules (τ_R), is directly processed by the recursive neural network through the adaptive encoding function τ_E. The internal representation developed by the recursive neural 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).

associated with the j-th component, the fluorine atom is represented by the following vector $[\underbrace{-1,-1,...,-1,}_{i-1}1,-1,...,-1,-1]$, while the chlorine 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 Fig. 3 we exemplify the above visit on an input tree where the labels are not explicitly represented. First the leaves (nodes 4, 5, 6 and 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, 6, and 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 grey 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.

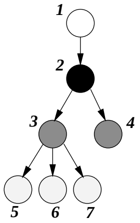

Fig. 3. The coding process: a code is progressively generated for each node by using the code already produced for its descendants. Nodes colored with different grey levels are used to denote the time when the code of each node is used as state information for the current node: e.g., the code for node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

Note that the way the encoding function acts on a specific tree, such as the tree in Fig. 3, is specified in terms of how the encoding function acts on the sub-trees 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.

2.2 The Recursive Neural Network Model

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 vertices. Let $I\!R^n$ denote the label space.

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

Finally, we define the encoding function as

$$\tau_E : \mathcal{G} \rightarrow I\!R^m \tag{5}$$

and the output function g as

$$g : I\!R^m \to I\!R. \tag{6}$$

The use of a stationary and causal model for τ_E allows us to choose a uniform and quite simple neural realization for each step of τ_E through the definition of a recursive neural network model. In order 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

$$I\!R^n \times \underbrace{I\!R^m \times \cdots \times I\!R^m}_{\text{k times}}$$

and produces a code in $I\!R^m$.

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

$$\boldsymbol{x} = \boldsymbol{F}\left(\boldsymbol{W}\boldsymbol{l} + \sum_{j=1}^{k} \widehat{\boldsymbol{W}}_j \boldsymbol{x}^{(j)} + \boldsymbol{\theta}\right), \tag{7}$$

where $\boldsymbol{l} \in I\!R^n$ is the label (external input) associated with the current node, $\boldsymbol{W} \in I\!R^{m\times n}$ is the weight matrix associated with the label space, $\widehat{\boldsymbol{W}}_j \in I\!R^{m\times m}$ is the recursive weight matrix associated with the j-th subgraph code, $\boldsymbol{x}^{(j)} \in I\!R^m$ is the code computed for the j-th subgraph of the current node, $\boldsymbol{\theta} \in I\!R^m$ is the bias vector, and $\boldsymbol{F}(\boldsymbol{y})_i = f(\boldsymbol{y}_i)$ where $f(\cdot)$ is a sigmoidal nonlinear function.

Specifically, let us study what happens for a single recursive neuron with $m = 1$. The simplest non-linear neural realization of the recursive model is given by

$$x = f(\sum_{i=1}^{n} w_i l_i + \sum_{j=1}^{k} \hat{w}_j x^{(j)} + \theta), \tag{8}$$

where f is a sigmoidal function, w_i are the weights associated to the label space, $\hat{w}_j$ are the weights associated to the subgraphs spaces, θ is the bias, $\boldsymbol{l}$ is the current input label, $x^{(1)}, \ldots, x^{(k)}$ are the encoded representations of subgraphs , and x is the encoding of the current structure. A graphical representation of the single recursive neuron is given in Fig. 4.

Using equation (7) 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.

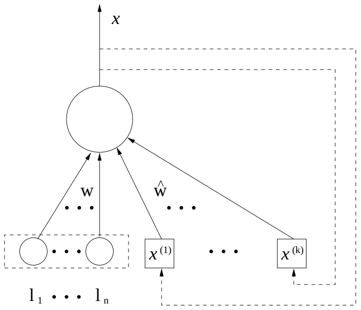

Fig. 4. A graphical representation of a recursive neuron.

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 $m = 1$ are shown in Fig. 5. The examples involve two substituents (-CH_3 and -CH_2-CF_3) for the benzodiazepine class of molecules studied in this work. More complete examples are in Fig. 6, based on the same substituents, where two neurons are involved ($m = 2$) and a representation of the numerical vectors with $n = 3$ for the encoding of the symbol is reported. 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 [21]. Notice that the weight matrices are shared by different encoding networks (see Fig. 5), 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 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.

There are different ways to realize the recursive neural network ([8]). 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 Corre-

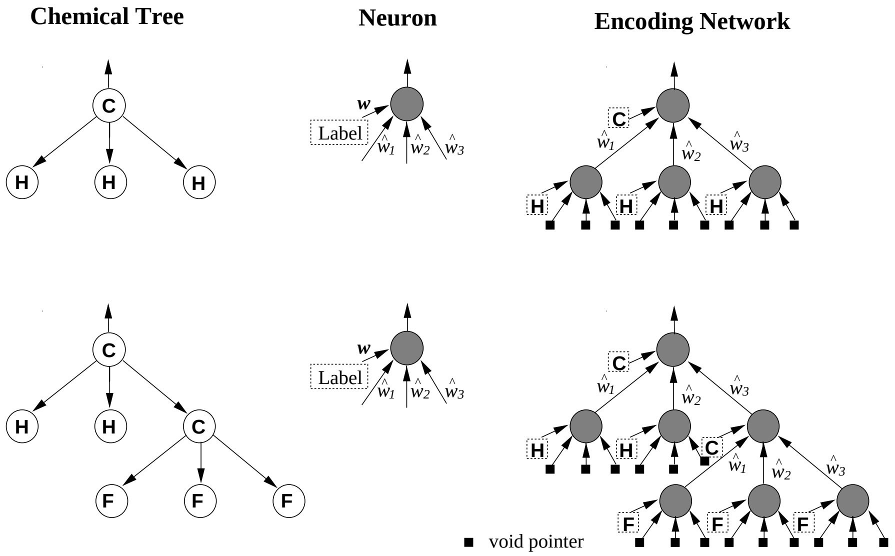

Fig. 5. Examples of encoding networks (left side) for the chemical fragments -CH_3 and -CH_2-CF_3 with $m = 1$. The fragments are assumed to be represented by the chemical trees shown on the left side of the figure. The encoding network are obtained by replicating (unfolding) the recursive neuron for each node in the chemical trees (as shown by the multiple occurrences of the weights). The black squares represents void pointers which are encoded as null vectors (in this case, the void pointer is equal to 0). The labels, here represented as symbols, are supposed to be encoded through suitable numerical vectors. The output of each encoding network is the code computed for the corresponding chemical fragments.

lation 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 $\mathbb{R}^m$) for each input DOAG, as shown in Fig. 5 or in Fig. 6. 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 QSPR/QSAR task. In the Cascade Correlation for structures (CCS) model, in order to realize the function g, we use a single standard linear output neuron. A complete description of the Cascade Correlation for structures algorithm and a formulation of the learning method and equations can be found in [27,8]

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 QSPR/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

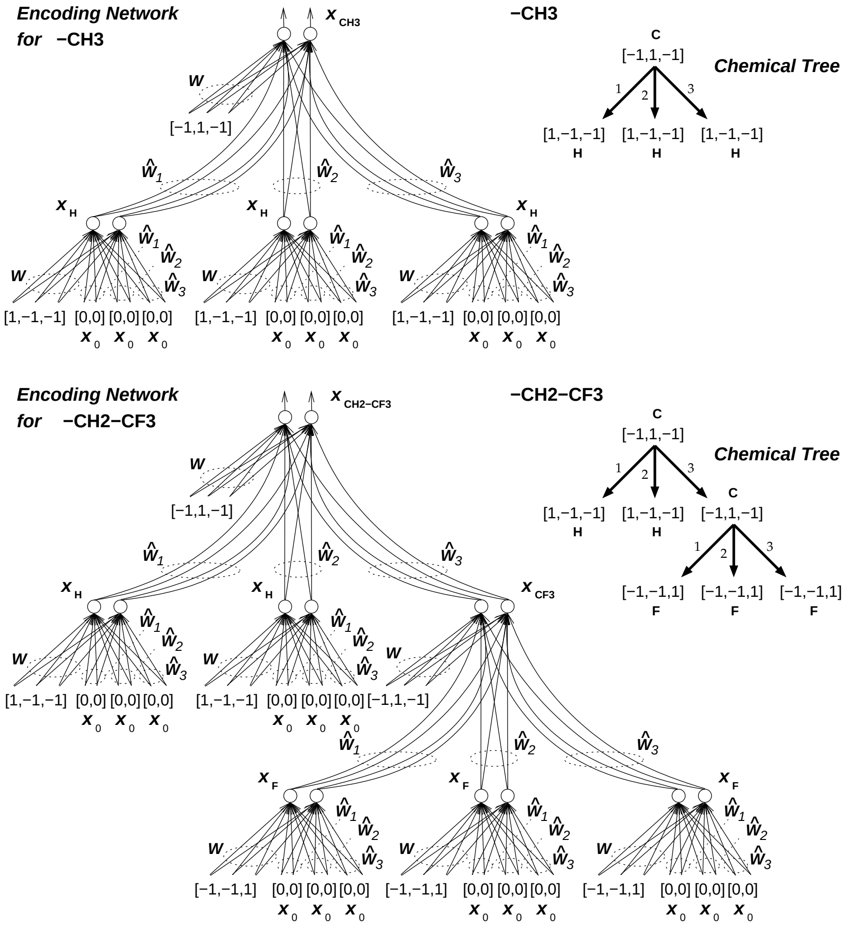

Fig. 6. Examples of encoding networks with $n = 3$ and $m = 2$ (left side) for the chemical fragments -CH_3 and -CH_2-CF_3. The labels of the chemical trees represent the atom types: H is represented by $[1, -1, -1]$, C by $[-1, 1, -1]$ and F by $[-1, -1, 1]$. Void subgraphs are encoded by the null vector $\mathbf{x}_0$. The output of each encoding network is the code computed for the corresponding chemical fragments (i.e., $\mathbf{x}_{CH3}$ and $\mathbf{x}_{CH2-CF3}$, respectively).

used as input for a linear regression model realized by the output unit (the $g()$ function), as shown in Fig. 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, expecially for QSAR, the following issues:

- chemical meaningfulness of the numerical descriptors produced by the recursive neural network;
- relationship between the developed numerical codes and the qualitative aspects of the QSAR problem.

Those issues were partially addressed in [14] by studying the internal representations developed by the recursive neural network trained on a specific family of benzodiazepines. Examples of such results are reported in Section 5.4.

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.

3 Representational Issues

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 Fig. 2. Since the functions τ_E and g are automatically developed by the model, in the new QSPR/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 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 the 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.

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.

For alkanes, where each carbon-hydrogens can correspond to a node of the tree, the root of the tree can be determined by the first carbon-hydrogens group according to the IUPAC nomenclature system, cycles are absent and the total order over the edges can be based on the size of the sub-compounds.

In the case of benzodiazepines, the major atom group that repeats unchanged throughout the class of analyzed compounds (common template) constitutes the root of the tree [2]. 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 physico-chemical 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 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 benzodiazepines data set, concerning the first problem, since cycles mainly constitute some common shared template of the benzodiazepines compounds, it is reasonable to represent them as a single node where the attached label codifies information about their chemical nature [3]. 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.

Rules that allows to define the function τ_R according to the above ideas will be specified in each section of the two different task (alkanes, Section 4.2, and benzodizepines, Section 5.2).

[2] 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.

[3] We distinguish different principal heterocycles or cycles that appear as substituents using different labels.

4 QSPR Analysis of Alkanes

4.1 QSPR Task: Alkanes

To assess the true performance of standard neural networks in QSPR, they are usually tested on well known physical properties. A typical example is the prediction of the boiling point of alkanes. The prediction task is well characterized for this class of compounds, since the boiling points of hydrocarbons depend upon molecular size and molecular shape, and vary regularly within a series of compounds, which means that there is a clear correlation between molecular shape and boiling point. Moreover, the relatively simple structure of these compounds is amenable to very compact representations such as topological indexes and/or vectorial codes, which are capable of retaining the relevant information for prediction. For these reasons, multilayer feed-forward networks using 'ad hoc' representations yield very good performances.

In order to perform a comparison with our method, we decided to use as reference point the work described in [11] which uses multilayer feed-forward networks. The data set used in [11] comprised all the 150 alkanes with 10 carbon atoms. Cherqaoui *et al.* use a vectorial code representation of alkanes obtained by encoding the chemical graph (tree) with suppressed hydrogens through an "*N-tuple*" code (see Fig. 7). Each component of the vectorial code, which in this case is of dimension 10, represents the number of carbon bonds for each carbon atom. The last components are filled by zeros when the number of atoms of the compound is less than 10. The uniqueness of the code is guaranteed by keeping a lexicographic order.

This representation for alkanes is particularly efficient for the prediction of the boiling point since it is well known that the boiling point is strongly correlated with the number of carbon atoms and the branching of the molecular structure. However, the same representation could be useless for a different class of compounds and different tasks.

4.2 Representation of Alkanes

We observe that the hydrogens suppressed graphs of alcane molecules are trees and they can be represented as ordered rooted trees by the following minimal set of rules:

1. the carbon-hydrogens groups (H, C, CH, CH_2, CH_3) are associated with graph vertexes while bonds between carbon atoms are represented by edges;
2. the root of the tree is defined as the first vertex of the main chain (i.e., the longest chain present in the compound) numbered from one end to the other according to IUPAC rules (the direction is chosen so to assign the lowest numbers possible to side chains, resorting, when needed, to

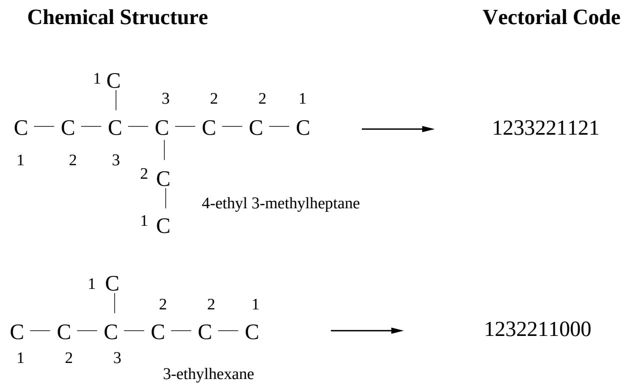

Fig. 7. Example of derivation of the vectorial code (*N-tuple*) for two alkanes. The vectorial code is obtained starting from a chemical graph where hydrogen atoms are "suppressed". The numbers represent the degree of each node in the graph.

a lexicographic order); moreover, if there are two or more side chains in equivalent positions, instead of using the IUPAC alphabetical order of the radicals names, we adopt an order based on the size of the side chains (i.e., depth of substructure);

3. the orientation of the edges follows the increasing levels of the trees;
4. the total order on the subtrees of each node is defined according to the depth of the substructure; we impose a total order on the three possible side chains occurring in our data set: methyl < ethyl < isopropyl.

Examples of representations for alkanes are shown in Fig. 8.

The complete lists of the compounds, according with our represenation, along with the target and the ouput results are reported in [13].

4.3 Experimental Results (Alkanes)

As the target output for the networks we used the boiling point in Celsius degrees normalized into the range $[-1.64, 1.74]$. A bipolar localist representation encoding the atom types was used.

For the sake of comparison, we tested the prediction ability using exactly the same 10-fold cross validation (15 compounds for each fold) used in [11]. Moreover, we repeated the procedure for four times. Learning was stopped when the maximum absolute error for a single compound was below 0.08.

The obtained results for the training data are reported in Table 1 and compared with the results obtained by different approaches, i.e., the results obtained by Cherqaoui *et. al.* using 'ad hoc' Neural Networks, two different equations based on connectivity (χ) topological indexes, and multilinear regression over the vectorial code for alkanes. The results obtained on the test set are shown in Table 2 and compared with the MLP results obtained by

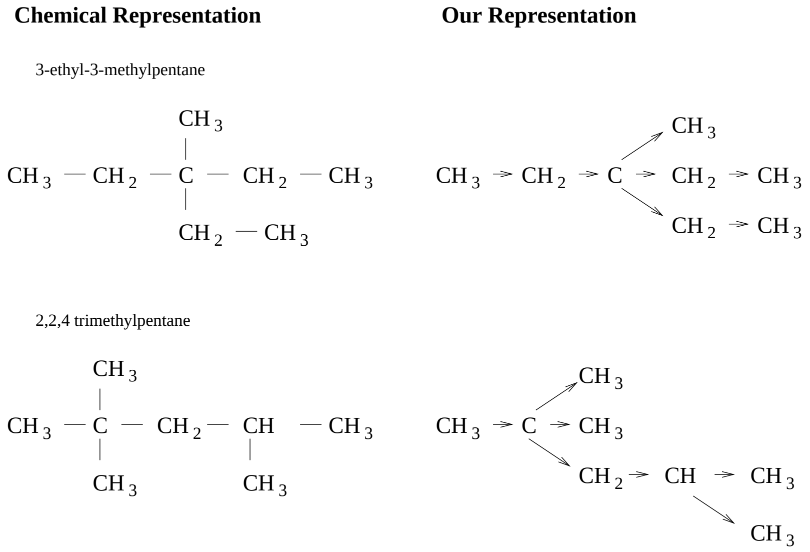

Fig. 8. Example of representations for alkanes.

Cherqaoui *et. al.* For completeness we have reported the cumulative results from a set of several trials of our model in row 3 of Table 2. It must be pointed out that the results are computed by removing the methane compound from the test set (for the MLP and CCS in Table 2), since it turns out to be an outlier. Particularly, from the point of view of our new approach that considers the structure of compounds, methane (CH_4) is so structurally small that it does not represent a typical element in the class of alkanes.

Model	#Units	Mean Abs. Error	R	S
CCS (Mean)	110.7	1.98	0.99987	2.51
Best MLP	7	2.22	0.99852	2.64
Top. Index 1			0.9916	6.36
Top. Index 2			0.9945	5.15
MLR			0.9917	6.51

Table 1. Results obtained for alkanes on training data set by Cascade Correlation for structure (CCS), by Cherqaoui *et. al.* using 'ad hoc' neural networks (MLP), by using topological indexes and by using multi linear regression. The data are obtained by a 10-fold cross-validation with 15 compounds for each fold. The correlation coefficient (R) and the standard deviation of error (S) are reported.

The results are presented in full, with t residual errors for each compound, in [13]. Examples of training and test curves for two different instances of Cascade Correlation networks trained over the same fold, are shown in Fig. 9.

Model	Mean Abs. Error	Max Abs. Error	R	S
Best MLP	3.01	10.42	0.9966	3.49
Best CCS	2.74	13.27	0.9966	3.5
Mean CCS	3.71	30.33	0.9917	5.43

Table 2. Results obtained for alkanes on test data set by Cascade Correlation for structure (CCS) and by 'ad hoc' neural networks (MLP). The data are obtained by a 10 fold cross-validation with 15 compounds for each fold. The last row of the Table is computed over four different cross-validation evaluations.

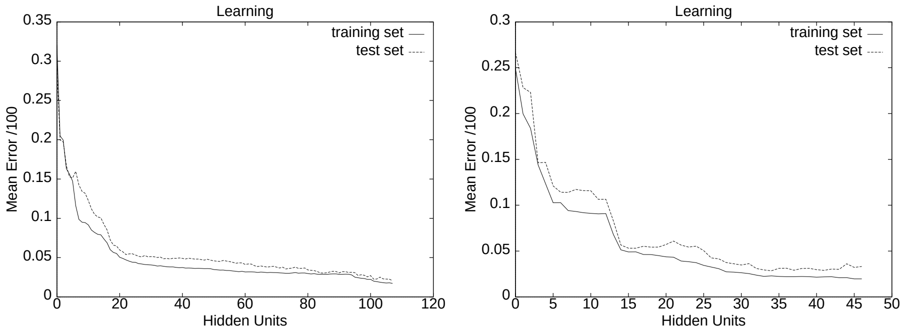

Fig. 9. Mean training and test error for two different instances of Recursive Cascade Correlation networks trained over the same fold. The mean error is plotted versus the number of inserted hidden units.

5 QSAR Analysis of Benzodiazepines

5.1 QSAR Task: 1,4-benzodiazepin-2-ones

Due to the strong therapeutic interest [10] and to the multiplicity of SAR studies of this class of compounds, benzodiazepines (Bz) were chosen as the starting application domain for QSAR analysis. At this stage, a group of 1,4-benzodiazepin-2- ones, previously studied by Hadjipavlou-Litina and Hansch [10] through traditional QSAR equations, was selected for testing our model, the evaluation of the method being the initial step of its application. The task is the prediction of the non-specific activity (affinity) towards the $Bz/GABA_A$ receptor. The affinity can be expressed as logarithm of the reciprocal of the drug concentration C (mol./liter) able to give a fixed biological response[4]. The data set analyzed by Hadjipavlou-Litina and Hansch (see Table 2 of [10]) 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. The complete list of the compounds, the training and test set used, and the ouput results are reported in [14].

All the molecules present a common template consisting of the Bz nucleus (in three compounds the A ring of the Bz nucleus consists of a thienyl instead

[4] In order to characterize the fixed response, the drug concentration able to give half of the maximum response (IC_{50}) is commonly used.

of a phenyl group) and they differ in a variety of substituents at the positions shown at the left side of Fig. 10.

5.2 Representation of Benzodiazepines

The labeled tree representation of a Bz is obtained by the following minimal set of rules:

1. the root of the tree represents the Bz nucleus;
2. the root has 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[5] 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;
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, *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, 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, naf, respectively, for fragments of Phenyl, 2-pyridyl, Cyclohexenyl, Cyclohexyl and Naphthyl. For labeling we use a bipolar localist representation, as shown in Section 2.

Examples of representations for benzodiazepines (or substituents) which comply with the above rules are shown in Fig. 10 (compound #60 in Table 5 in the Appendix) and in Fig. 5.

5.3 Experimental Results (Benzodiazepines)

In this section we briefly summarize experimental results obtained for the QSAR task [13,14].

For the analysis of the data set described in Section 5, four different splittings in disjoint training and test sets of the data were used (Data set I, II, II, and IV, respectively). Specifically, the first test set (5 compounds) has been chosen as it contains the same compounds used by Hadjipavlou-Litina and Hansch. The second data set is obtained from Data set I by removing 4 racemic compounds from the training set and one racemic compound from

[5] The multiplicity of the bound is implicitly encoded in the structure of the subtree.

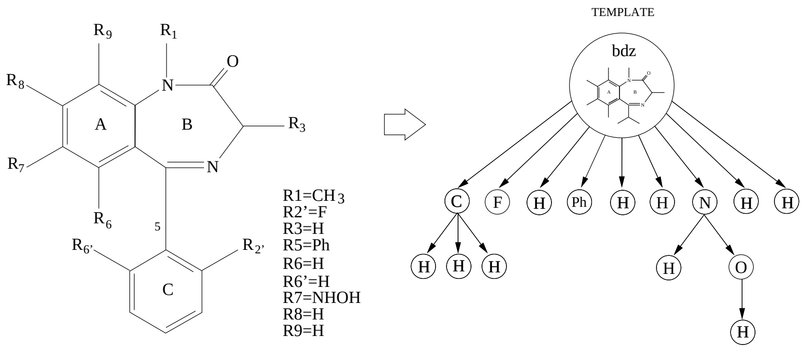

Fig. 10. Example of representation for a benzodiazepine.

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 (5 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 (4 compounds) has been randomly chosen so to test the sensitivity of the network to different learning conditions. The training set III, with the used numbering of the molecules, is reported in Table 5 in the Appendix.

As 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 3. 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 row, respectively. For each data set, statistics on the number of inserted hidden units are reported for the Cascade Correlation 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 Mean Abs. Error, R and S for Cascade Correlation 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 4. In 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 11 and 12 we have plotted the error of the network versus the desired target for data set I and III. Moreover, for the sake of comparison, in Fig. 11 the error obtained using an equational approach [10] on data set I is reported as well.

Each point referring to the neural networks 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).

Training Set	Mean #Units	Mean Abs. Error	R	S
	(Min-Max)	(Min-Max)		
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

Table 3. Results obtained for benzodiazepines on training data set I by Hadjipavlou-Litina and Hansch (HLH, first row), by a "null model" (second row) and on all the training data sets by Cascade Correlation for structures. The mean absolute error, the correlation coefficient (R) and the standard deviation of error (S) are reported.

Test Set	Data #	Mean Abs. Error	Mean Max Abs. Error	S
		(Min-Max)	(Min-Max)	
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

Table 4. Results obtained for benzodiazepines on test data set I by Hadjipavlou-Litina and Hansch (HLH, first row), by a "null model" (second row) and on all the test data sets by Cascade Correlation for structures. The mean absolute error, the mean of the maximum of the absolute error, and the standard deviation of error (S) are reported.

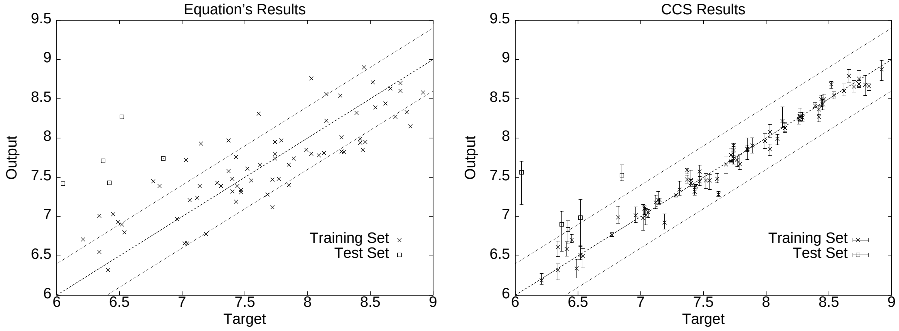

Fig. 11. Output of the models proposed by Hadjipavlou-Litina and Hansch (left) and for the Cascade Correlation for structures network (CCS) (right) versus the desired target; both models use the same training and test sets (data set I). Each point in the right plot represents the mean expected output for Cascade Correlation network, together with the deviation range (minimum and maximum values), as computed over six trials. The tolerance region is shown on the plots.

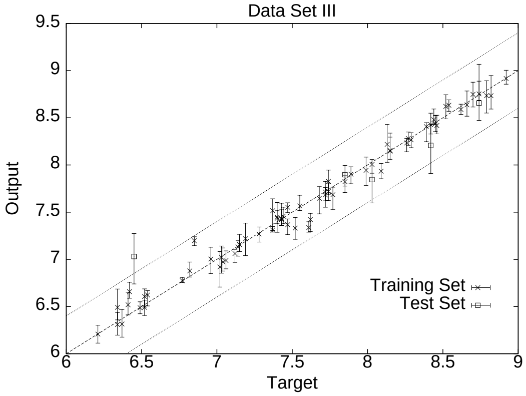

Fig. 12. Output of the models for Cascade Correlation network (CCS) versus the desired target using the data set III . Each point in the plot represents the mean expected output for Cascade Correlation network, together with the deviation range (minimum and maximum values), as computed over six trials. Note that the test data are spread across the input range.

Due to the small number of training examples we considered various learning strategies in order to avoid or mitigate the overfitting problem. We fully described the adopted strategies in [13] and [14]. Basically we control the gains of the sigmoids, and the increase of the weight values through an incremental strategy on the number of training epochs for each new inserted hidden node. The improvement in the learning behavior using our strategies is analyzed in [14].

5.4 Internal Representation Analysis

In order to understand the degree to which the proposed model is able to capture relevant domain knowledge from the training data, we investigated the internal representations, i.e. the output of hidden units, developed by the neural network trained with the selected set of benzodizepines.

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 to each compound and its subcomponents.

For this investigation we performed a Principal Component Analysis (PCA) of the internal representations. Due to the relatively large dimensionality of the representational space (typically around 20-30 hidden units are inserted by the training algorithm), we studied 2-D plots of the first two principal components. The aim was to show, as 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.

From previous SAR studies some positions of the Bz nucleus are recognized to be the ones where substituents play significant roles in determining the biological activity also in relation to their specific chemical characteristics: positions 1, 7 and $2'$ ([10] and references therein). Within the class of compounds analyzed the above mentioned positions appear to be widely sampled.

In brief, the most important characteristics required for substituents at position 1 concern lipophylicity and steric hindrance, while the ones required for substituents at position 7 and $2'$ (or $2'$ and $6'$), mostly concern the electronic effect. Lipophylicity ($\pi = logP$) and electronic effect of the substituents (Hammett σ constant) constitute the most popular physico-chemical descriptors employed in the traditional equation based Hansch approach [1,2]. Substitutions at positions 6, 8, and 9 are known, instead, to decrease the affinity.

What we were interested in finding, through the analysis of the first two principal components was the presence of clusters possibly containing molecules grouped according to a classification amenable to the two above mentioned descriptors. As a first approach we reduced the relevant molecular descriptor to very simple entities, in order to make the analysis as clear as possible. From this perspective we collected into a unique class the lipophylicity (π) and steric hindrance descriptors, and only considered an on-off classification (molecules with a hydrogen atom or molecules with substituents, mostly lipophylic, at position 1). In the more detailed analysis reported in [14] we reduced the scale of the substituent effect values (Hammett σ scale) to a few sub-classes corresponding to the effect that the substituents produce on well known chemical reactions (electrophilic substitution in aromatic compounds). But for the results reported here we again considered an on-off classification, i.e. presence or absence of halogens atoms (F, Cl, Br, I). In fact halogens atoms strongly affect the σ values.

We then focused our interest on the analysis of the molecules on the basis of the substituents at position $2'$ or $2'$ and $6'$. We considered three cases: *(i)* molecules with only one halogen substituent (at position $2'$), *(ii)* molecules with two halogen substituents (at the symmetrical $2'$ and $6'$ positions) and *(iii)* molecules bearing no-halogen substituents at these positions.

The principal components of the internal representations developed by the Cascade Correlation for structures (outputs of recursive hidden neurons) were analyzed for all the six experiments on data set III mentioned in Section (5.3).

A representative plot of the first two principal components is shown in Fig. 13. It shows the biologically active molecules analyzed (compounds associated to a target) and the relevant molecular fragments. Examples involving more experimental trials are described in [14].

Analysis of the plot shows that molecules and fragments are clustered on the basis of both morphological differences and specifc chemical features, that can not be inferred directly by the observation of the molecular graph, rather only by the association of molecular structures and targets.

The plot (see Fig. 13) appears to be split in 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 group containing compounds associated to a target is divided, in turn, in two sub-groups, highlighted in the plot shown in Fig. 13 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 Bz 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 little sub-group on the right side of the plot contains compounds characterized by thienyl, instead of the phenyl, for the group A ring of the Bz nucleus.

Both the biggest clusters contain molecules divided in turn into smaller homogeneous sub-clusters on the basis of the presence of substituents at the other significant positions of the Bz nucleos previously mentioned.

In Fig. 14 we observe that each of the two big clusters identified in the previous plots is sub-clustered 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 sub-groups so identified

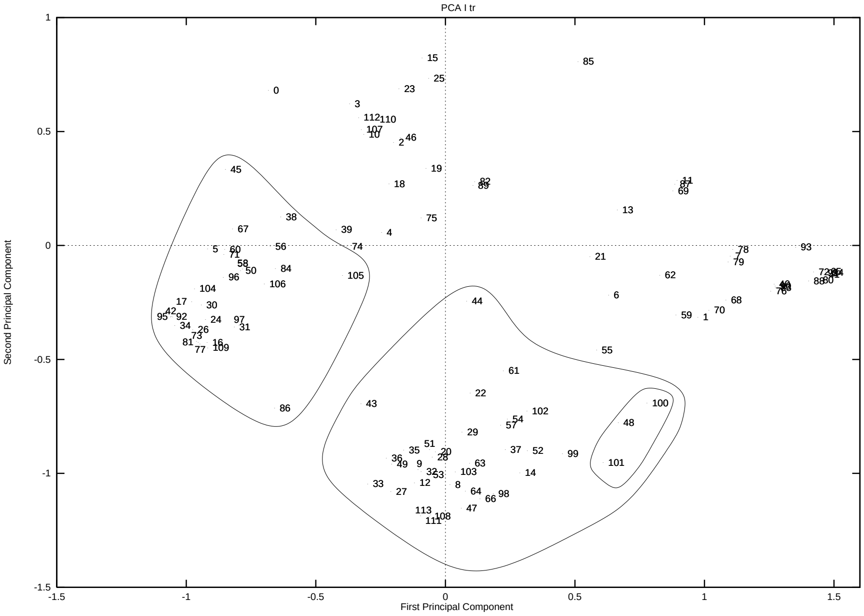

Fig. 13. Principal component analysis of training compounds used in the experiment I derived from 28 output values of hidden neurons. Compounds characterized by R_1=H (left side of the plot) and compounds bearing a substituent at position 1 (lower side of the plot) are grouped by contour lines. The circled sub-cluster on the right side includes compounds where the A ring of the Bz nucleus is a thienyl group instead of a phenyl one. See Table 5 in the Appendix for compound numbering.

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 plot shown in Fig. 15 allows us to focus the analysis on the presence and the type of substituent at position $2'$ and $2' - 6'$: once again quite homogeneous sub-groups were found. The sub-groups appear only slightly overlapping. Compounds characterized by the presence of only one halogen at position $2'$ are marked by little boxes, and compounds characterized by the simultaneous presence of halogens at position $2'$ and $6'$ are marked by a cross within little boxes.

The analysis of positions 6, 8, and 9, shows sub-groups still characterized by a certain degree of homogeneity, as reported in [14].

It is noticeable that the differences in analogous plots showing the results obtained from distinct experiments (corresponding to different realizations of the model) only consist of rotations and/or translations of the clusters with respect to each other, i.e. the molecules are still homogeneously clustered on the basis of the substituent effects. For details see [14].

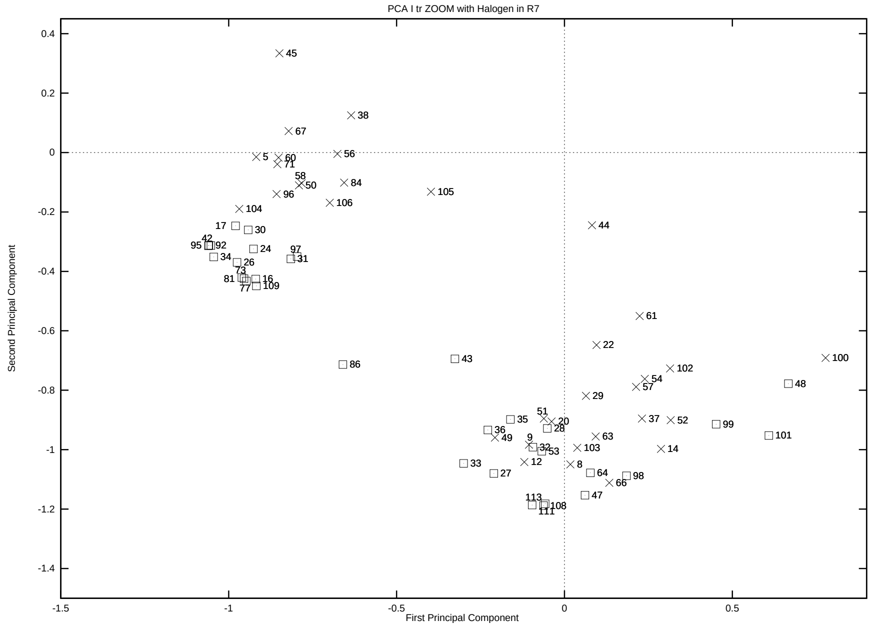

Fig. 14. An expanded view of the circled areas in Fig. 13. Compounds characterized by R_7 = halogen are marked by little boxes; compounds where R_7 is not a halogen are marked by times signs. Compounds bearing a halogen atom at position 7 appear to be located at the (left) lower side of each group.

6 Discussion

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 most poorly predicted compound is the one bearing hydrogen atoms in place of substituents which play an important role in determining affinity. 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 contains a Naphthyl group as C ring which never occurs in the training set. This explains the high variance observed in the prediction.

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 Principal Component Analysis, as expected, we found that the global distribution of molecules and fragments in the plots of the two

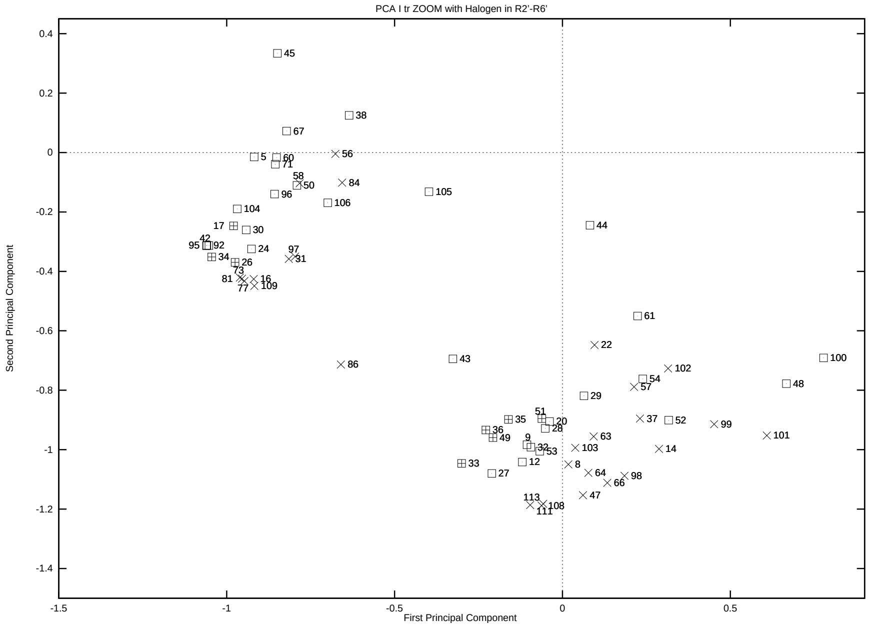

Fig. 15. An expanded view of the circled areas of the plot in Fig. 13. Compounds characterized by $R_{2'}$ = halogen are marked by boxes; compounds bearing halogen atoms both at position $2'$ and $6'$ are marked by plus signs in boxes, and compounds where $R_{2'}$ and $R_{6'}$ are not an halogen are marked by times signs. Compounds bearing halogen atoms at positions $2'$ or $2'$ and $6'$ appear to be located at the (left) upper side of each group.

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.

The behavior of the model for the prediction of the boiling point of alkanes demonstrates the ability of the model to be competitive with respect to ‘ad hoc’ techniques. In fact, the obtained results compare favorably with the approach proposed by Cherqaoui *et. al.* bearing in mind that the vectorial representation of alkanes retains the structural information which is known to be relevant to the prediction of the boiling point.

We would like to stress that the experimental results seem to confirm that our approach allows the prediction, without substantial modifications, both for QSAR and QSPR tasks, obtaining competitive or even better results than traditional approaches.

7 Conclusions

We have demonstrated that the application of neural networks for structures to QSAR/QSPR tasks allows the treatment of different computational tasks by using the same basic representations for chemical compounds, obtaining improved prediction results with respect to traditional equational approaches for QSAR and competitive results with respect to ‘ad hoc’ designed representations and MLP networks in QSPR. It must be stressed that for QSAR, no physico-chemical descriptor was used by our model, however, it is still possible to use them by the insertion into the representation of the compounds.

The main advantage of the proposed approach with respect to topological indexes is that in our case no *a priori* definition of structural features is required. Specifically, since the learning phase involves both the encoding and the regression process, the numerical encoding for the chemical structures devised by the encoding network are optimized with respect to the prediction task. Of course, this is not the case for topological indexes which need to be devised and optimized through a trial and error procedure by experts in the fields of application. Moreover, in our approach it is possible to store into the label attached to each node information at different levels of abstraction, such as the atom types or functional groups, allowing a flexible treatment of different aspects of the chemical functionality.

The capability of the model in extracting structural features which are significant for the target correlation is shown by the PCA of internal representation. 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 physico-chemical 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.

Concerning a comparison with respect to approaches based on feedforward networks, the main advantage resides in the fact that the encoding of chemical structures does not depend on a fixed vectorial or template based representation. In fact, due to the dynamical nature of the computational model, our approach is able to adapt the encoding process to the specific morphology of each single compound.

Moreover, the generality of the compound representations used by our approach allows the simultaneous treatment of chemically heterogeneous compounds. Finally, our approach must be regarded as a major step towards a fully structural representation and treatment of the chemical compounds using neural networks.

References

1. C. Hansch, P.P. Maloney, T. Fujita, and R.M. Muir. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature*, 194:178–180, 1962.
2. C. Hansch and T. Fujita. Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.*, 86:1616–1626, 1964.
3. S.M. Free Jr. and J.W. Wilson. A mathematical contribution to structure-activity studies. *J. Med. Chem.*, 7:395–399, 1964.
4. L. H. Hall and L. B. Kier. *Reviews in Computational Chemistry*, chapter 9, The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling, pp 367–422. VCH Publishers, Inc.: New York, 1991.
5. D. H. Rouvray. Should we have designs on topological indices ? In R. B. King, editor, *Chemical Applications of Topology and Graph Theory*, pp 159–177. Elsevier Science Publishing Company, 1983.
6. V. R. Magnuson, D. K. Harris, and S. C. Basak. Topological indices based on neighborhood symmetry: Chemical and biological application. In R. B. King, editor, *Chemical Applications of Topology and Graph Theory*, pp 178–191. Elsevier Science Publishing Company, 1983.
7. M. Barysz, G. Jashari, R. S. Lall, V. K. Srivastava, and N. Trinajstic. On the distance matrix of molecules containing heteroatoms. In R. B. King, editor, *Chemical Applications of Topology and Graph Theory*, pp 222–230. Elsevier Science Publishing Company, 1983.
8. A. Sperduti and A. Starita. Supervised neural networks for the classification of structures. *IEEE Trans on Neural Networks*, 8(3):714–735, 1997.
9. P. Frasconi, M. Gori, and A. Sperduti. A general framework for adaptive processing of data structures. In *IEEE Trans on Neural Networks*, 9: 768–785, 1998.
10. D. Hadjipavlou-Litina and C. Hansch. Quantitative Structure-Activity Relationships of the benzodiazepines. A review and reevaluation. *Chemical Reviews*, 94(6):1483–1505, 1994.
11. D. Cherqaoui and D. Villemin. Use of neural network to determine the boiling point of alkanes. *J. Chem. Soc. Faraday Trans.*, 90(1):97–102, 1994.
12. A.M. Bianucci, A. Micheli, A. Sperduti, and A. Starita. Quantitative structure-activity relationships of benzodiazepines by recursive cascade correlation. In *IEEE International Joint Conference on Neural Networks*, pp 117–122, 1998.
13. A.M. Bianucci, A. Micheli, A. Sperduti, and A. Starita. Application of cascade correlation networks for structures to chemistry. *Journal of Applied Intelligence*, 12:117–147, 2000.
14. A. Micheli, A. Sperduti, A. Starita, and A.M. Bianucci. Analysis of the internal representations developed by neural networks for structures applied to quantitative structure-activity relationship studies of benzodiazepines. *Journal of Chemical Information and Computer Sciences*, 41(1):202–218, January 2001.

15. Y. Suzuki T. Aoyama and H. Ichikawa. Neural networks applied to quantitative structure-activity relationships. *J. Med. Chem.*, 33:2583–2590, 1990.
16. Ajay. A unified framework for using neural networks to build QSARs. *J. Med. Chem.*, 36:3565–3571, 1993.
17. K. L. Peterson. Quantitative structure-activity relationships in carboquinones and benzodiazepines using counter-propagation neural networks. *J. Chem. Inf. Comput. Sci.*, 35(5):896–904, 1995.
18. A. F. Duprat, T. Huynh, and G. Dreyfus. Towards a Principled Methodology for Neural Network Design and Performance Evaluation in QSAR; Application to the Prediction of LogP. *J. Chem. Inf. Comput. Sci.*, pp 854–866, 1998.
19. Shuhui Liu, Ruisheng Zhang, Mancang Liu, and Zhide Hu. Neural networks-topological indices approach to the prediction of properties of alkene. *J. Chem. Inf. Comput. Sci.*, 37:1146–1151, 1997.
20. D. W. Elrod, G. M. Maggiora, and R. G. Trenary. Application of neural networks in chemistry. 1. prediction of electrophilic aromatic substitution reactions. *J. Chem. Inf. Comput. Sci.*, 30:447–484, 1990.
21. V. Kvasnička and J. Pospichal. Application of neural networks in chemistry.prediction of product distribution of nitration in a series of monosubstituted benzenes. *J. Mol. Struct. (Theochem)*, 235:227–242, 1991.
22. James Devillers, editor. *Neural Networks in QSAR and Drug Design.* Academic Press, London, 1996.
23. J. Zupan and J. Gasteiger. *Neural Networks for Chemists: an introduction.* VCH Publishers, NY(USA), 1993.
24. J. A. Burns and G. M. Whitesides. Feed-forward neural networks in chemistry: Mathematical system for classification and pattern recognition. *Chemical Reviews*, 93(8):2583–2601, 1993.
25. S. E. Fahlman and C. Lebiere. The cascade-correlation learning architecture. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pp 524–532. San Mateo, CA: Morgan Kaufmann, 1990.
26. S. E. Fahlman. The recurrent cascade-correlation architecture. In R.P. Lippmann, J.E. Moody, and D.S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pp 190–196, San Mateo, CA, 1991. Morgan Kaufmann Publishers.
27. A. Sperduti, D. Majidi, and A. Starita. Extended cascade-correlation for syntactic and structural pattern recognition. In Petra Perner, Patrick Wang, and Azriel Rosenfeld, editors, *Advances in Structural and Syntactical Pattern Recognition*, volume 1121 of *Lecture notes in Computer Science*, pp 90–99. Springer-Verlag, Berlin, 1996.

A Appendix

In the following the training set for benzodiazepines data used in data set III are reported. We report in the tables the numbers associated to compounds (not their fragments) as used in Fig. 13, Fig. 14, and Fig. 15.

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 * in Table 5).

Table 5. Training Data Set III

#	R1	R3/R6	R7	R8/R9	R2'	R6'	Log 1/C
5	$-CH_3$		$-CN$		$-F$		7.52
8			$-CH{=}CH_2$				7.62
9					$-F$		7.68
12			$-COCH_3$		$-F$		7.74
14			$-CF_3$				7.89
16	$-CH_3$		$-Cl$				8.09
17	$-CH_3$		$-Cl$		$-Cl$	$-Cl$	8.26
20			$-N_3$		$-F$		8.27
22			$-NO_2$		$-CF_3$		8.45
24	$-CH_3$		$-I$		$-F$		8.54
26	$-CH_3$		$-Br$		$-F$	$-F$	8.62
27			$-Cl$		$-F$		8.70
28			$-Cl$		$-Cl$		8.74
29			$-NO_2$		$-F$		8.82
30	$-CH_3$		$-F$		$-F$		8.29
31	$-CH_3$		$-F$				7.77
32			$-F$		$-F$		8.13
33			$-Cl$		$-F$	$-F$	8.79
34	$-CH_3$		$-Cl$		$-F$	$-F$	8.39
35			$-Cl$		$-Cl$	$-F$	8.52
36			$-Cl$		$-Cl$	$-Cl$	8.15
37			$-NO_2$				7.99
38	$-CH_3$		$-NO_2$		$-Cl$		8.66
42	$-CH_2CH_2OH$		$-Cl$		$-F$		7.61
43		R3= $-(s)CH_3$	$-Cl$		$-F$		8.46
44		R3= $-(s)CH_3$	$-NO_2$		$-Cl$		8.92
45	$-CH_3$	R3= $-(s)CH_3$	$-NO_2$		$-F$		8.15
47*			$-Br$				7.74
48			$-Cl$		$-Cl$		8.03
49					$-F$	$-F$	7.72
50	$-CH_3$				$-Cl$		8.42
51				R8= $-Cl$	$-F$	$-F$	7.55
52				R8= $-CH_3$	$-F$		7.72
53			$-Cl$	R8= $-Cl$	$-F$		8.44
54			$-CH_3$	R8= $-Cl$	$-F$		7.85
56	$-CH_3$		$-NH_2$				6.34
57			$-NH_2$				6.41
58	$-CH_3$		$-CN$				6.42
60	$-CH_3$		$-NHOH$		$-F$		7.02
61			$-NH_2$		$-Cl$		7.12
63			$-CHO$				7.37
64			$-F$				7.40
66			$-C_2H_5$				7.44
67	$-CH_3$		$-NH_2$		$-F$		7.19
71	$-CH_3$		$-NHCONHCH_3$		$-F$		6.34
73	$-CH_2-CF_3$		$-Cl$				7.04
77	$-CH_2-C\equiv CH$		$-Cl$				7.03
81	$-CH_2C_3H_5$		$-Cl$				6.96
84	$-CH_2OCH_3$		$-NO_2$				6.37
86	$-C(CH_3)_3$		$-Cl$				6.21
92	$-(CH_2)_2OCH_2CONH_2$		$-Cl$		$-F$		7.37
95	$-CH_2CHOHCH_2OH$		$-Cl$		$-F$		6.85
96	$-CH_3$	R6= $-Cl$		R8= $-Cl$	$-F$		6.52
97	$-CH_3$		$-Cl$	R8= $-Cl$			7.40
98			$-Cl$	R9= $-Cl$			7.43
99			$-Cl$	R9= $-CH_3$			7.28
100					$-Cl$		7.43
101			$-Cl$				7.15
102		R6= $-CH_3$	$-CH_3$				6.77
103		R6= $-Cl$					6.49
104	$-CH_3$	R6= $-Cl$			$-F$		6.82
105	$-C(CH_3)_3$		$-NO_2$		$-Cl$		6.52
106	$-CH_3$			R9= $-Cl$	$-F$		7.14
108*			$-Cl$				7.47
109*	$-CH_3$		$-Cl$				7.47
111*			$-Cl$				7.06
113*			$-Cl$				6.54