Design of New Biologically Active Molecules by Recursive Neural Networks

Alessio Micheli, Alessandro Sperduti, Antonina Starita
University of Pisa
Dipartimento di Informatica
Corso Italia 40, 56125, Pisa, Italy
E-mail: { *micheli,perso,starita* } *@di.unipi.it*

Anna Maria Bianucci
University of Pisa
Dipartimento di Scienze Farmaceutiche
Via Bonanno 6, 56126, Pisa, Italy
E-mail: bianucci@farm.unipi.it

Abstract

In this paper, we face the design of novel molecules belonging to the class of adenine analogues (8-azaadenine derivates), that present a widespread potential theraupetic interest, in the new perspective offered by recursive neural networks for Quantitative Structure-Activity Relationships analysis. The generality and flexibility of the method used to process structured domains allows us to propose new solutions to the representation problem of this set of compounds and to obtain good prediction results, as it has been proved by the comparison with the values obtained 'a posteriori' after synthesis and biological essays of designed molecules.

1 Introduction

The purpose of this paper is aimed to explore a real-world problem: the design of new biologically active molecules by recursive neural networks (RNN) which appear to supply a well suited tool for Quantitative Structure-Activity Relationships (QSAR) analysis. RNN allows to combine the flexibility of a connectionist model and the representational power of a structured domain in a learning system. In particular, RNN allows to deal with the prediction of molecular activity on the basis of the adaptive processing of chemical compounds represented as labeled trees.

The use of methods to deal with structured domain in chemical application gives an opportunity to obtain a direct and adaptive relationship between molecular structures and their properties avoiding the use of any fixed coding scheme of the structure in preestablished features, offering a new perspective to the QSAR studies.

In fact, the generality and flexibility of a structured representation, allows one to deal with heterogeneous compounds and heterogeneous problems using the same approache. This advantage is not at the expense of predictive accuracy: our previous results [1] [2] compare favorably versus the traditional QSAR treatment, for the analysis of benzodiazepines [1]. 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 [2].

Successive studies on the internal representation developed by the RNN (realized by a CCS algorithm) applied to QSAR studies of benzodiazepines were conducted using principal component analysis [3]. 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.

After the completion of these preliminary steps of assessment of the RNN as new approach to the QSAR analysis, we propose an application of our model as prediction tool for new molecules. Here we present the design of novel molecules belonging to the class of adenine analogues (8-azaadenine derivatives), many of which act as antagonists of the human A1 adenosine receptors. They present a widespread potential therapeutic interest. Just as few examples, we will mention that they may be exploited as potassium sparing diuretics, with kidney-protective properties, or in therapy of degenerative diseases such as the Alzheimer's one [4]. We expect that the design of new antagonists take a big advantage from the use of predictive tools such as the one supplied by the QSAR models. In fact, a reliable quantitative prediction of the activity before the compound is made is of great interest to reduce the cost of drugs development.

The generality and flexibility of the structured representation allows us to deal with a class of compound characterized by a quite high morphological complexity, e.g. tautomeric forms.

The main aspect of this work is the completion of the QSAR analysis followed by the prediction step for a small theoretical library of compounds. The molecules, which a significant activity had been predicted for, have been subsequently synthesized and biologically essayed. Some qualitative analysis are also conducted for chiral substituent and the tautomeric forms.

In Sec. 2 we summarize the description of the RNN model in the framework of the QSAR methodologies and of the novelty of our approach to QSAR. It is followed by the description of the characteristics of the application task under analysis and by the presentation of the adopted representation of the molecular structure. Simulation results, both for quantitative and qualitative aspects, are reported in Sec. 3 and followed by the conclusion.

2 Method

2.1 The Recursive Neural Model for QSAR
In this section we briefly present the approaches to the QSAR analysis using an uniform formalism, viewed as functional transduction, both for the traditional and the new approach based on recursive neural networks. We then describe how to realize this functional transduction by the cascade correlation for structures (CCS) neural networks.

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.:

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

The function $\mathcal{T} : \mathcal{I} \to \mathcal{O}$ 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 equation (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 of 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

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

where $\tau()$ is the *encoding* function from the domain of the chemical compounds to the descriptors space, while g is the

mapping function from the descriptors space to the biological activity space. This corresponds to the traditional QSAR approach [5, 6], 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. Usually, the number and type of descriptors used to represent the chemical compound depend on the specific QSAR problem at hand and the selection is performed by an expert.

More in detail, the encoding process requires the solution of two subtasks. The aim of the first one is to represent explicitly 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 [5], a molecule is first of all 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

$$\tau() = \tau_E \circ \tau_R, \qquad (3)$$

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.

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 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.

The RNN [7] allows a new approach to the problem. In fact this model is able to take directly as input the graph generated by τ_R and to implement adaptively both τ_E and g.

In our approach we define a function τ_E that allows to progressively encode an input structure, e.g. a tree, using at each step a neural computational model τ_{NN}. Before defining these functions we provide a proper instantiation of their input and output domains. 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 and that each DOAG possess a super-source, i.e. a vertex s such that every vertex in the graph can be reached by a directed path starting from s. In the case of trees the super-source is defined by his root node. 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, the class of functions which can be realized by a recursive neural network can be characterized as the class of functional graph transductions described in the form $g \circ \tau_E$, where $\tau_E : \mathcal{G} \to I\!\!R^m$ is the encoding function, and $g : I\!\!R^m \to I\!\!R$ the output function.

The function τ_{NN} is used to process a single node of a given structure. Given a node in a DOAG, τ_{NN} uses the information available at the current node: *i)* the numerical label attached to the node (in $I\!\!R^n$), *ii)* the numerical code for each subgraph of the node (in $I\!\!R^m$), and produces a code in $I\!\!R^m$. As a result, if k is the maximum out-degree of DOAGs in $\mathcal{G}$, τ_{NN} is defined as

$$\tau_{NN} : I\!\!R^n \times \underbrace{I\!\!R^m \times \cdots \times I\!\!R^m}_{k \text{ times}} \to I\!\!R^m \qquad (4)$$

The composition of τ_{NN} used to encode a structured set of nodes, e.g. a tree X, is defined by the following recursive definition of τ_E:

$$\tau_E(X) = \begin{cases} 0 & \text{if } X \text{ is empty} \\ \tau_{NN}(X_{root}, \tau_E(X^{(1)}), \ldots, \tau_E(X^{(k)})) \end{cases}$$

$$(5)$$

where 0 is the null vector in $I\!\!R^m$, *root* is the root node (or super-source of the tree X), X_{root} is the label attached to the root, and $X^{(1)}, \ldots, X^{(k)}$ are the subgraphs pointed by *root*. Note that the same definition may be applied to DOAG once the super-source s corresponds to the root of the tree.

The equation 5 comprehensively defines the functionality of the recursive neural network. The recursive definition of τ_E determines a systematic visit of the input DOAG. It guides the application of τ_{NN} to each node of the structures, from the frontier to the super-source of the input DOAG, allowing the neural model to incrementally compute a numerical code for the whole structure. Since the computation of each code depends only on the current node and nodes descending by it, the encoding is *causal*. This property, with the *stationarity* assumption, i.e. the computation that produces the code is the same for all the nodes, allows to choose a uniform and quite simple neural realization for each step of τ_E through the neural definition of τ_{NN}.

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

$$x = \tau_{NN}(l, x^{(1)}, \ldots, x^{(k)}) = \Phi(Wl + \sum_{j=1}^{o} \widehat{W}_j x^{(j)} + \theta),$$

$$(6)$$

where $\Phi_i(v) = \varphi(v_i)$ (sigmoidal function), $l \in I\!\!R^n$ is a label, $\theta \in I\!\!R^m$ is the bias vector, $W \in I\!\!R^{m \times n}$ is the weight matrix associated with the label space, $x^{(j)} \in I\!\!R^m$ are the vectorial codes obtained by the application of the encoding function τ_E to the subgraphs $X^{(j)}$ (i.e., $x^{(j)} = \tau_E(X^{(j)})$), and $\widehat{W}_j \in I\!\!R^{m \times m}$ is the weight matrix associated with the jth subgraph space.

There are different ways to realize the recursive neural network [7]. 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 [8, 9]. The built neural network has a hidden layer composed of recursive (hidden) units. The recursive hidden units compute the values of τ_E (in $I\!\!R^m$) for each input DOAG. 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 order to realize the function g is possible to choose any known mathematical model. In the class of neurocomputing models the function g may be realized using a multilayer network to perform regression or classification tasks. In the CCS model we use a single linear output neuron to realize a regression model: $g(x) = m^T x + \beta$, where $m \in I\!\!R^m$ and $\beta \in I\!\!R$ is the output threshold. A graphical representation of the architecture developed by the CCS algorithm, when 3 hidden units are added to the networks, is shown in Figure 1. 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 1.

A complete description of the CCS algorithm and a formulation of the learning method and equations can be found in [7, 2].

Summarizing, using the RNN we realize the τ_E function through an adaptive mapping. The process can consider both the graph topology (connectivity) and the atom types (or the chemical functionalities). Since the encoding func-
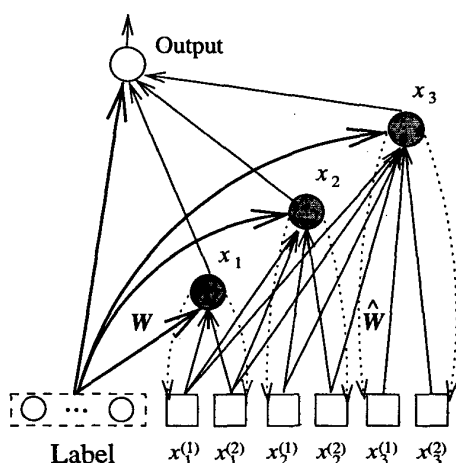
Figure 1: Architecture of a CCS with 3 hidden units ($m = 3$) and $k = 2$. Recursive hidden units (shadowed) generate the code of the input graph (function τ_E). The hidden units are added to the network during the training. The box elements are used to store the output of the hidden units, i.e. the code $x_i^{(j)}$ that represent the context according to the graph topology. The output unit realize the function g and produce the final prediction value.

tion (τ_E) is learnt by the neural network together with the mapping function (g), we allow the automatic generation of numerical descriptors which are specific for the regression (QSAR) task to be solved. Differently from the previous approaches, 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.

2.2 The Task: QSAR for 8-azaadenine

As mentioned in Section 1, the class of molecules studied for our analysis is made up of 8-azaadenine derivatives. The selected target property was the affinity toward the receptor expressed as binding constant (Ki) values. The binding constant is usually measured in nanomolar (nM) units, which is a concentration unit. It is a measure of the strength of the binding between the ligands (in our case antagonists) and the receptor that constitutes the target biological macromolecule. Lower Ki values correspond to more active ligands. The strength of the binding is one of the requirements needed to be satisfied in order to obtain a drug exploitable in therapy. It is measured 'in vitro', while the whole molecular properties needed are subsequently measured 'in vivo'. Usually the QSAR studies are focused in 'in vitro' data, since the 'in vivo' data bases are not large enough for QSAR analysis.

We face the QSAR task according to the following steps: (a) Composition of data set: the 117 molecules, of known Kis, were selected among the ones synthesized and biologically essayed in the Medicinal Chemistry group which we co-operate to (already published compounds). These data set present some novel characteristics: *i*) chiral compounds are included and the target values refer to the racemic mixture; pure enantiomers are described by the same structural formula, but present specular arrangements of the atoms; the biological activity may differ significantly from each other; *ii*) eight compounds undergo a tautomeric equilibrium. This means that the morphology of these molecules can be associated to two different structures. In general, we cannot indicate *a priori* if only one tautomer or both of them, and, if so, in which ratio, are effective on the receptor. In fact the environment surrounding the ligand in its bound state contributes to determine the tautomeric ratio. For each compound undergoing tautomeric equilibrium two different representations were used which the same value of the target property was associated to (see Sec. 2.3). The total number of structure becames 125.

(b) Training of the CCS model using 91 molecules (plus 5 tautomeric forms) and , in parallel,

(c) the analysis and validation steps, performed with a test set composed by around 20% of the total number of data (26 molecules, plus 3 tautomeric forms).

(d) Design of a small virtual library of new compounds ("prediction set"). In our preliminary attempt this set was built with the following guidelines: *i*) taking into account new suggestions from qualitative SAR; *ii*) assaying the predictive power of the neural model at the sampling limits of the training data; *iii*) maximize the molecular diversity; *iv*) minimize the development cost for the synthesis laboratories and maximize the rapidity of such synthesis.

(e) We have used the CCS model obtained from the step (b) and (c) to compute the target property (Ki) for data of the prediction set. The compound designed were then synthesized and biologically essayed. We report the synthesis of the quantitative results in Sec. 3 .

(f) We have drawn some qualitative results on specific compounds belonging to the prediction set, as reported in Sec. 3.

2.3 Molecular Structure Representation

The definition of an appropriate function τ_R for the specific set of molecules studied is discussed in [1] and [2] using some simple rules according to the standard conventions used in chemistry (IUPAC nomenclature). The goal of such rules is to obtain an unique structured representation of each substituent fragment as labeled ordered tree. Basically each node represent an atom or a cycle or a functional group and each arc a bound of the molecules fragment. The shared nu-

cleous of the molecular group (template) is the root of the whole molecular tree. An example of representation for a 8-azaadenine compound is shown in Figure 2 for the first tautomeric form; the template is labeled "AZA". Substituents at positions 2, 6 and 9 are responsible for the diversity of the library. The number used to indicate the substituent position of the nucleus are inherited from the IUPAC nomenclature of the endogenous ligand adenosine. For the second tautomeric form the substructure in position 7 and 9 are exchanged. The two tautomeric forms correspond to the same target value.

Concerning the label attached to each node, a bipolar localist representation encoding the types of the chemical objects has been used. For example, the label for the N atom would be something like [-1,-1,....,-1,1,-1,....,-1,-1].
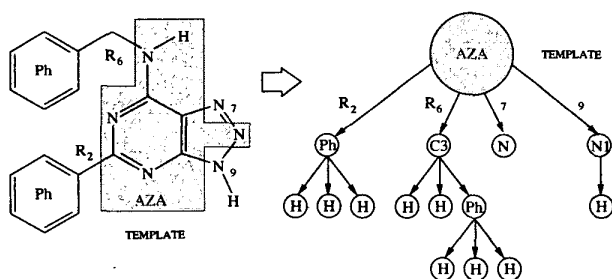

Figure 2: Example of a 8-azaadenine (left side) with the template-nucleus (shadowed) and his representation as labeled tree (right side).

3 Results

Due to the low number of training data and to avoid overfitting, several expedients were used. First of all, no connection between hidden units were allowed. Then the gain of the sigmoids of the hidden units were set to 0.4. Finally, an incremental strategy (*i-strategy*) on the number of training epochs was adopted for each new inserted hidden node. Allowing few epochs to the first nodes we avoid the increase of the weight values. The advanteges of this strategy are already shown in [3]. The work of Bartlett [10] gives theoretical support for techniques, like the *i-strategy*, that allows to produce networks with small weights. As a result we could continue learning, adding new hidden unit in CCS, without overtraining the model. No significant increase in the test error was encontered in our experiments. However, the large tollerance in the precison of the target values in the data set (experimental values) and the limitations in the sampling quality of the sub-components of the structures present in the small training set do not allow to expect an optimal generalization performance.

As target output for the networks we used $log(Ki)$ normalized into the range $[0, 1]$. An initial set of preliminary trials were performed in order to determine an admissible range for the learning parameters. However, no effort was done to optimize these parameters. Five trials were carried out for the simulation using different random initialization for the connection weights.

The main statistics computed over all the simulations for the training set are reported in Table 1. Statistics on the number of inserted hidden units are reported, as well as on the expected mean absolute error and the standard deviation of error (S). Note that the mean absolute error and S are obtained by averaging over the single trials; also the minimum and maximum values of the mean absolute error over these five trials are reported. The maximum error is 1%, as fixed by the criteria to stop the learning.

Table 1: Results Obtained on the Training Set by CCS.

#Units Mean(Min-Max)	Mean Error(Min-Max)	S
22.6 (20-29)	0.0020 (0.0009-0.0028)	0.0026

The results for the corresponding test set are reported in Table 2, adding a column for the maximum absolute error. For the test data set we also report on the second row the statistics for the committee model (based on a simply ensemble averaging method) obtained using the mean expected output as computed over the five trials.

Table 2: Results Obtained on the Test Set by CCS and by the Ensemble Averaging Method.

Model	Mean Error (Min-Max)	Max Error (Min-Max)	S
CCS	0.1124 (0.0932-0.1230)	0.3006 (0.2512-0.3561)	0.1402
Committee	0.0650	0.2227	0.1201

Our main interest is on prediction set: the results for four significant compounds are reported in Table 3. For each compound we report the target (experimental measured activity), the mean output over the 5 trials (predicted value), the error value, and the correspondent values of the target and the output re-expressed in the original Ki scale.

Note that the differences between the predicted and the experimentally measured activity were below the usually encountered experimental error. Three compounds (number 2,3, and 4) in Table 3 are predicted to be quite active ligand and the experimental results have confirmed this prevision.

Table 3: Results Obtained for Each Compound of the Prediction Set by CCS.

#	Target	Output	%Error	Target Ki	Output Ki
1	0.7312	0.7486	-1.74	354	439.4
2	0.5516	0.5584	-0.67	38	41.3
3	0.5443	0.5266	1.77	34.7	27.8
4	0.5054	0.5745	-6.90	21.4	50.4

We have tested other new compounds to extract qualitative information. Since we enclose in the virtual library some pure enantiomers, although in the training data mostly racemic mixtures were considered, we tested both enantiomer forms (R and S). In particular the compounds number 1 and 3 in Table 3 are pure enantiomers of type R. The correspondent S enantiomers are less active ligands. We found in both cases that the S enantiomers were badly predicted, and this observation suggests that the model has learnt to predict the values of the R enantiomers also from the racemic mixtures sampled in the training data set. This result is compatible with what already known from biological data.

Concerning the evaluation of results, we note a gap between the test and training set accuracy, even though no overtraining is encountered. As mentioned above, high tolerance in the target values and sampling incompleteness can explain these results. However, a simply committee technique, like ensemble averaging method, allows to obtain a mean error of 6.5% for test set data, that is a result at the state of the art in QSAR analysis. Moreover we could obtain very reliable results for prediction set. This suggests to focus future works on the data set and the theoretical library enlargement in order to obtain a more rational and complete sampling of structure instances.

We have to point out, respect to the accuracy, that the precision, which the binding constant can be experimentally obtained with, is quite low. It affects obviously the quality of the training data. Nevertheless also the precision in the prediction doesn't need to be high: predictions that can be successfully exploited in drug design may be affected for example by an error up to 10 times the binding constant value, when its value is below 50 nM, o up to 3 times when its value is about 150-200 nM. It fact the need is usually to design ligands with binding constant below few hundred nM. Our results (errors) are largely below the above tolerance needed for a correct classification of active drugs.

With regard to the ability to discriminate between different tautomers shown by the model we observed that both in the training and in the test set the mean error calculated for each group of tautomers has the same order of the mean error calculated for the molecules not undergoing tautomeric equilibrium. Even though the number of enantiomers is limited, we can infer that both tautomeric forms may be active on the receptor.

Summarizing, we can conclude that the good fit with the experimental data and the qualitative results confirm the ability shown by the model of capturing relevant features of the whole molecular morphology.

4 Conclusions

The present experiments allow to test the potentiality of the connectionist approach for the challenge task of a rational design of new drugs by predictive models. Our results are a further step of assessment of the approach and allow the connectionist methods to enter deeply in the drug discovery process.

References

[1] A.M. Bianucci, A. Micheli, A. Sperduti, and A. Starita. Quantitative structure-activity relationships of benzodiazepines by recursive cascade correlation. In *Proceedings of IJCNN*, pp 117-122, 1998.

[2] 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.

[3] 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.

[4] C. E. Muller. A1 adenosine receptor antagonists. *Exp. Opin. Ther. Patents*, 7(5): 419-440, 1997.

[5] 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.

[6] 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.

[7] A. Sperduti and A. Starita. Supervised neural networks for the classification of structures. *IEEE Trans on Neural Networks*, 8(3):714-735, 1997.

[8] S. E. Fahlman and C. Lebiere. The cascade-correlation learning architecture. *Advances in Neural Information Processing Systems 2*, pp 524-532. San Mateo, CA: Morgan Kaufmann, 1990.

[9] S. E. Fahlman. The recurrent cascade-correlation architecture. *Advances in Neural Information Processing Systems 3*, pp 190-196, San Mateo, CA: Morgan Kaufmann, 1991.

[10] P. L. Bartelett. The sample complexity of pattern classification with neural networks: the size of weights is more important than the size of the network. *IEEE Trans on Inform. Theory*, 44(2):525-536, 1998.