Prediction of Chemical-Physical Properties by Neural Networks for Structures

Celia Duce,[1] *Alessio Micheli,*[2] *Roberto Solaro,**[1] *Antonina Starita,*[2] *Maria Rosaria Tiné*[1]

Summary: Here we present an overview of a new approach to cheminformatics based on recursive neural networks. This approach allows for combining the flexibility and advantages of neural networks with the representational power of structured domains. Current advances, which include applications to the prediction of the solvation free energy of small molecules in water and of the glass transition temperature of (meth)acrylic polymers are reported.

Keywords: biomaterials; cheminformatics; QSPR/QSAR; recursive neural networks

Introduction

In the biomedical field there is an increasing demand for new methods for drug delivery and for substituting or integrating damaged tissues and organs. Conventional dosage forms, which still dominate pharmaceutical products, are not able to control either the rate of drug delivery or the target area of drug administration. Accumulation at non-target sites may lead to adverse reactions and undesirable side effects. Therefore, suitable delivery systems need to be designed in order to enhance the drug bioavailability with substantial abatement of undesired adverse side effects.[1] For example, the biodistribution of active principles can be modified by incorporation within polymeric nanoparticles.[2] Indeed, drug-loaded nanoparticles exposing targeting moieties could effectively and selectively drive the active principle at the desired site of action.[3,4]

On the other hand, the advent of synthetic polymers and biocompatible metals in the latter part of the twentieth century has changed the entire character of health care. Materials are needed for the removal of congenital defects and for the replacement of tissues that have been either damaged or destroyed by pathological processes.[5] Medical devices may replace a damaged part of anatomy; simulate a missing part; aid in tissue healing; aid in diagnosis. Since the performances of the implant biomaterial must fulfill the functions of the body part to be replaced, the requirements that they must satisfy are very stringent. Their characteristics must be different and quite complex. For example, together with proper mechanical and physical-chemical properties, they must be biocompatible and/or hemocompatible.[6] Moreover, the applications of most of the familiar polymers with relatively simple repeat unit structures, have reached their limits, so that the chemical structures of polymers, suitable for advanced applications, have increased in complexity. Consequently, the development of predictive methods to evaluate the most promising candidates for specific applications, has gained urgency.

In time, significant efforts have been spent on the development of Quantitative Structure-Activity/Property Relationship (QSAR/QSPR) techniques in order to predict physical, chemical, biological, biomedical, and technological properties of

[1] University of Pisa, Department of Chemistry and Industrial Chemistry, Via Risorgimento 35, 56126 Pisa, Italy E-mail: rosola@dcci.unipi.it
Fax: (+39) 050 2219260
E-mail: rosola@dcci.unipi.it

[2] University of Pisa, Department of Informatics, Largo B. Pontecorvo 3, 56127 Pisa, Italy

molecules. The aim of QSAR/QSPR is to find an appropriate function, which given a proper representation of a molecule can predict a selected property.

Traditional QSAR/QSPR approaches, employing standard regression methods (from linear regression to standard neural network), take fixed-size numerical vectors as input. Consequently, all molecules must be reduced to vectors of the same dimension by using a suitable group of molecular descriptors. The molecule can be represented by using different encoding approaches, such as, the selection of physical-chemical, geometrical, and electronic properties, the calculation of topological or connectivity indices, as well as the occurrence of each group in the molecular structure.

The need for molecular descriptors limits the type of modeled molecules and determines the applicability of the method. In particular, both experimental and computed descriptors are numerical values that encode only partial aspects of the molecular structure. Additionally, the number and types of numerical descriptors used to represent chemical compounds are strictly dependent on the (target) property under study; for this reason, the models are not target-invariant. Hence, an expert has to start again from scratch the process of choosing suitable descriptors whenever a different property is investigated.

It is possible to overcome most of these issues by predicting properties directly from the molecular structure. To this aim we use Recursive Neural Network (RecNN) methods that take a labeled structured representation of molecules as input.

The first successful applications of the RecNN model were achieved predicting the boiling points of linear and branched alkanes and the pharmacological activity of a series of substituted benzodiazepines.[7–10] More recently, further advancements have been done to deal with a widest set of molecular structures and to address different chemical tasks.

In particular, we tested the RecNN-based method for small molecules by applying it to the prediction of standard free energy, $\Delta_{\text{solv}}G^{\circ}$, of solvation in water of a set of organic compounds.[11–13] The standard free energy of solvation in water was selected as the target property because of the availability of a large dataset of reliable literature data. Indeed, a homogeneous and critically reviewed data base is needed in order to reliably assess which performance can be obtained by the application of the proposed model to a given problem. Furthermore, $\Delta_{\text{solv}}G^{\circ}$ of small organic molecules is of significant interest in drug design,[14] in the analysis of protein folding and binding,[15] and in the development of force fields by computer simulation.[16]

Afterwards, we extended the method to macromolecules by investigating the glass transition temperature, Tg, of a set of acyclic hydrocarbon-chain polymers. The wide class of different chemical data and QSPR/QSAR problems faced by these applications provide the support to show the generality of the approach.

Method

Before entering in the description of the results, let us introduce the basics of the Recursive Neural Networks (RecNN) model we used to tackle the learning in structured domain for chemical data.[7–9]

Recursive neural networks extend standard neural networks (NN) to deal with structured domains, i.e. input patterns can be a class of graphs, trees, sequences, etc. We choose to describe the molecules by a 2-D graph directly inferred from the structural formula. The molecular structures are represented in terms of labeled Directed Positional Acyclic Graphs (DPAGs).[17] In such structures, for each vertex (or *node*) a total order is defined on the edges leaving from it and a position is assigned to each edge. We assume a bounded out-degree and that each DPAG has a super-source, i.e. a vertex *s* such that every vertex of the graph can be reached by a directed path starting from *s*. *Labels* are tuples of variables attached to vertexes.

Let $\mathcal{R}^n$ denote the label space. Here, the classes of used DPAGs are in form of *k-ary trees*. *k-ary trees* (*trees* in the following) are rooted positional trees with finite out-degree *k*, i.e. *k* is the maximum number of children for each node. The super-source is the root of the tree. Vertexes with zero out-degree are *leaves* of the tree.

In the framework of the QSPR/QSAR analysis, and according to the RecNN architecture, the processing of a RecNN can be presented as the sequential application of two functions, an *encoding function* and a *mapping function*. Let us consider a realization of the two functions by a recursive neural network with *m* hidden neurons. The *encoding* of an input structure, e.g. a tree T, is made by the hidden units computing a numerical code ($\boldsymbol{x}$ in $\mathcal{R}^m$) for each vertex of T by using information of both the vertex label ($\boldsymbol{l}$ in $\mathcal{R}^n$) and, recursively, the code, denoted as $x^{(j)}$ in $\mathcal{R}^m$, of the sub-trees descending from the current vertex. The *encoding function*, i.e. the output $\boldsymbol{x}$ of the hidden units for a vertex *v* (the code of *v*), is computed as:

$$x = \Phi\left(Wl + \sum_{j=1}^{k} \hat{\mathrm{W}}^{j} x^{(j)}\right) \tag{1}$$

where Φ is a set of *m* sigmoid functions, $\boldsymbol{W}$ in $\mathcal{R}^{\mathrm{m}\times\mathrm{n}}$ is the weight (free-parameters) matrix associated with the label space, and $\hat{\mathrm{W}}^{j}$ in $\mathcal{R}^{\mathrm{m}\times\mathrm{m}}$ is the weight (free-parameters) matrix associated with the *j*-th sub-tree space. The *bias* is included in the label $\boldsymbol{l}$. Through Equation 1 the encoding function is recursively computed for all the vertexes of the input structure and a code for the whole structure is returned at the root.

The encoding process of the RecNN is graphically represented in Figure 1 for two input structures representing acrylic acid and poly(methacrylic acid), respectively. Note that the encoding process mimics the morphology of each compound. As shown in Figure 1, the encoding is a bottom-up process starting from leaves (black arrows). This corresponds to a visit (*traversal*) of the input tree according to an inverse topological order.

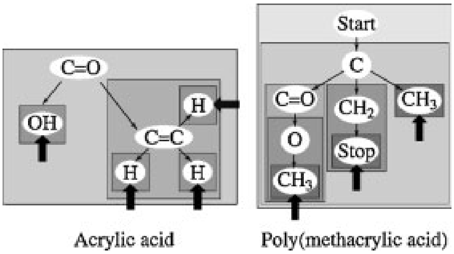

Figure 1.
Examples of the encoding process unfolded through the molecular structures. Each box includes the sub-trees progressively encoded by the RecNN.

The model described in Equation 1 is applied for each step of the traversal. At the root, this process computes a code of the whole molecular structure. The code is then mapped to the output property value by the output function $y = g(\boldsymbol{x})$.

In order to realize the output *mapping function* for the regression model, we use a single linear output neuron:

$$y = g(x) = \boldsymbol{A}x + \theta \tag{2}$$

where $\boldsymbol{A}$ is a weight (free-parameters) matrix in $\mathcal{R}^m$ and θ is the output threshold.

Different architectures of the neural network that realize the encoding function can be considered. In particular, in the following we used a constructive approach, a Recursive Cascade Correlation method,[9,17] which adds the hidden recursive neurons during the training of the model. Since this method automatically determines the number *m* of hidden units, it has been found particularly useful in applications when no information is given on the complexity of the problem.

We can summarize the characteristics of the RecNN approach in the context of QSPR/QSAR application by the following main points:

- RecNN take directly a structured representation of molecules as input;
- Recursive models can learn (by tuning the free-parameters) how to encode the structured representation of molecules

according to the given QSPR/QSAR task;

- Through the *encoding* and *mapping* functions, the RecNN models a direct and adaptive relationship between molecular structures and target properties;

Hence, RecNN discover by learning the specific structural descriptors (numerical code) for the QSPR/QSAR task at hand. As a result, no *a priori* definition and/or selection of properties by an expert are needed.

Results and Discussion

Concerning the task of predicting the free energies of solvation in water, the molecular structures of almost 300 acyclic organic compounds were represented in terms of labelled rooted trees. To this aim we studied an appropriate set of rules in order to build a unique chemical tree for each molecule. Since labelled structures are high abstract graphical tools we could describe a molecule at different levels of detail, such as atom bonds and/or chemical groups. In particular, each compound was divided into defined atomic groups. Each group corresponds to a vertex of the tree and each bond between them corresponds to an edge. We chose the smallest number of atomic groups able to build the greatest number of molecules in a reasonably compact form. The labels of vertex are categorical attribute distinguishing the symbols of the atomic groups. We decided to place the root in the functional group characteristic of the class to which the molecule pertains; all the other groups descending from it are considered branches. For a more detailed description of the representation rules see references 11–13.

The whole data set, containing alkanes, alkenes, alkynes, alcohols, ethers, thiols, thioethers, aldehydes, ketones, carboxylic acids, esters, amines, amides, haloalkanes, nitriles, and nitroalkanes was divided into disjoint training and test sets for the learning and the validation process respectively. The target values in the training set were expressed in kJ mol^{-1} and ranged from -40.63 to 14.94 kJ mol^{-1}.

The molecules were selected so that the test set was representative of the different molecular sizes, topologies and functional groups, provided they were present in the training set. In order to have a significant evaluation of the results, sixteen trials were carried out for RecNN simulations involving each one of the different experiments, and the results were averaged over the different trials. In order to assess the whole system (including the data cleaning aspects) several experiments were carried out encompassing data sets of increasing size and various splitting of the data. The current comprehensive result presents 0.09 Kj mol^{-1} mean absolute error, 0.54 kJ mol^{-1} maximum absolute error, 0.999 correlation coefficient (R), and 0.14 kJ mol^{-1} standard deviation (S) for a training set of 236 compounds; 1.00 kJ mol^{-1} mean absolute error, 6.46 kJ mol^{-1} maximum absolute error, 0.988 correlation coefficient, and 1.62 kJ mol^{-1} standard deviation for a test set of 60 compounds. To assess the occurrence of over-fitting, we used a higher error tolerance for training; the obtained results indicate that no improvements on the test set are due to a lower fitting.

These results are satisfactory considering that the data used for training the system were reproduced within the experimental error; the data of test sets were predicted with a mean absolute error and standard deviation in good qualitative agreement with methods based on state-of-the-art neural networks that in turn definitely show better performances than others QSAR-QSPR models, both for regression and prediction purposes.[18–25] Furthermore, the generality of the proposed approach stems from the fact that the RecNN model takes the molecular structure directly as input, whereas the results of the standard neural network model mainly depend on the availability of the best molecular descriptors related to the property under study.

The principal component analysis (PCA) of the internal representation of molecules (i.e. the output of the encoding function) built by the neural network showed that the RecNN is able to cluster the molecules not only by considering the chemical similarity of the molecular trees, but also by abstracting chemical information from the relationships between structures and targets learned by the model. For instance, the solvent accessibility of polar groups and their ability to act as hydrogen bond donor and/or acceptor are highlighted by the distribution of polar molecules in the representation space developed by the RecNN.[11–13]

The success obtained with small molecules encouraged us to extend the approach to the prediction of polymer properties. In the present research, the glass transition temperature (Tg) of a set of acyclic polymers including polyacrylates, polymethacrylates, polyacrylamides, polymethacrylamides, and some α- and β-substituted polyacrylics and polymethacrylics was investigated. Acrylic and methacrylic polymers were chosen because of the availability of a large number of experimental data, which allows for testing the potential of our RecNN model with macromolecules. On the other hand, it is well known that the glass-rubber transition is of considerable technological significance. In fact, the Tg determines the utilization limits of rubbers and thermoplastic materials. For instance, the Tg of materials designed to replace soft and hard tissues must be lower than and well above body temperature, respectively.

Several standard regression methods are reported in the literature for the prediction of polymer Tg.[26–34] These methods use molecular descriptors for the representation of the molecular structure. The limitations associated with standard QSAR/QSPR methods already evidenced in the treatment of small molecules are exacerbated in the case of polymers. Molecular descriptors are indeed inadequate tools for the complete description of the whole macromolecular structure in that they can be only evaluated for one repeating unit or for a short repeating unit sequence at the best. Moreover, material properties are not only intrinsic to the polymer chemical structure, but they also depend on average characteristics of the polymer, such as, molecular weight, polydispersity index, stereo-regularity, repeating unit distribution. As a consequence, these methods are mostly used for amorphous polymeric materials and are not applied to copolymers, which convey repeating units with different molecular structures. On the other hand, direct treatment of structured data, as it is possible with our RecNN model, enables to by-pass the limitations associated with the use of molecular descriptors.

The representation of each polymer was based on the 2D graph of its repeating unit treated as a small molecule. In particular, each repeating unit was decomposed by using the same atomic groups, labels, and priority rules defined for low molecular weight compounds. With respect to the small molecule representation, the most relevant innovation was the positioning of the tree root. Indeed, the root was not placed on the highest-priority chemical group, but on an additional super-source vertex (the group ‘‘Start’’), not related to the molecular graph (see Figure 1). The super-source label conveys information on the average macromolecule characteristics. This allows the model to account for both the repeating unit detailed 2D structure and macromolecule average characteristics. In the first application of this approach, we encoded the information of polymer stereoregularity (tacticity) in the super-source label as the fraction of *rr* dyads. This extension of the representation of small molecules to macromolecules further points to the flexibility of a structured representation approach.

The whole data set was divided in disjoint training and test sets for the learning and the validation process respectively. The training set consisted of 127 compounds and the test set of 27 compounds. About 20 and 11% of compounds in the training set and in the test set, respectively are present in more than one tacticity form. The macromolecules were selected in order to make the test set representative of the different molecular sizes, topologies, and functional groups present in the training set. The target values in the training set and in the test set ranged from 162 to 501 K and from 198 to 413 K, respectively. Two experiments were carried out with different maximum tolerance for training and sixteen trials were performed for each experiment.

Glass transition temperatures reported in the literature can be affected by the presence of additives, fillers, unreacted monomers and/or impurities, the specimen size and thermal history, and the determination method. The experimental uncertainty can be estimated in the 10–20 K range. Moreover, the Tg values reported by different authors for the same polymer differ in some cases by as much as 80 K. Consequently, in the first and in the second experiment, learning was stopped when the maximum error for each compound of the training set was below 20 and 110 K, respectively. The same training and test sets were used in both experiments. The results averaged over the different trials are collected in Table 1.

The results obtained in experiment 1 were satisfactory with a mean error of about 2 K and 14 K for the training and the test set, respectively. The standard deviation for the test set was 19 K, corresponding to 6.5% of the average experimental Tg (292 K). The second experiment was performed to empirically assess if a lower fitting could be useful to tackle and to evaluate the noise and outliers of experimental data. In this case the model is underexploited since a very rough fitting is imposed. The mean errors of both the training and the test sets of Experiment 2 were obviously higher than those of the previous experiment.

A punctual analysis of the mean error for each compound of the training set showed that the absolute errors increased with the number of carbon atoms in the side chain and the greatest absolute errors (all above the mean training error) corresponded to polymers with long alkyl chains. It is worth noting that recent literature researches on polyacrylics and polymethacrylics polymers question that the reported Tg values of long side chain vinyl polymers really correspond to a glass transition process.

The results obtained in predicting the Tg are promising. Indeed, the potential of the RecNN model to take effectively into account the extent and type of stereoregularity of the polymer chains in the encoding of molecular structures is of paramount importance because of the impact of these features on several properties of the materials. For instance, very often stereoregular polymers are highly crystalline, whereas atactic polymers are amorphous. On the other hand, methods able to correlate the Tg of polymers with their tacticity are lacking in literature.

Concluding Remarks

The reported results highlight the greater generality and flexibility of our method and

Table 1.
Mean and maximum absolute error, correlation coefficient (R), standard deviation of error (S), in the training and test sets of the different experiments.[a)]

Exp.	Max tolerance	Training set				Test set			
	(K)	Mean abs. error[b)]	Max abs. error[c)]	R[d)]	S	Mean abs. error[b)]	Max abs. error[c)]	R[d)]	S
1	20	1.88	19.10	0.9985	3.56	13.57	55.86	0.948	18.90
2	110	14.83	89.58	0.9490	20.87	20.84	78.66	0.897	27.48

[a)] All the statistical parameters, except R, are expressed in K.
[b)] Mean absolute error defined as the mean of the absolute residuals.
[c)] Maximum absolute error.
[d)] Linear correlation coefficient between the experimental and calculated values.

of the adopted representation with respect to standard literature methods. In fact, the RecNN method overcomes the limits of a vectorial representation of the data by treating molecules of variable structure and size. In particular, we have shown the possibility to treat small molecules and polymers with the same fundamental approach. In the latter case, the method allows for taking into account both the molecular structure of the repeating unit and the mean macromolecular characteristics, and for the simultaneous handling of polymers for which one or more values of the considered average property exist. Moreover the molecular representation can be naturally extended to the treatment of all types (random, alternating, block) of copolymers.

Finally, an overview of the presented results shows the potential of machine learning methods capable of handling structured data in contributing to the building of flexible tools for the prediction of chemical-physical properties.

Acknowledgements: The financial support by MIUR Cofin2003 Project "Integrated multidisciplinary design and realization of biologically active systems for biomedical applications" is gratefully acknowledged.

[1] V.H.L. Lee, "*Peptide and Protein Drug Delivery*", Marcel Dekker, New York, 1990.

[2] R. Solaro, F. Chiellini, F. Signori, C. Fiumi, R. Bizzarri, E. Chiellini, *J. Mat. Sci. Mat. Med.*, **2003**, *14*, 705.

[3] D. F. Ranney, *Biochem. Pharmacol.* **2000**, *59*, 105–114.

[4] G. E. Francis, C. Delgado, *Drug Targeting. Strategies, Principles, and Applications*, Humana Press, Totwa, NJ, 2000.

[5] H. Ringsdorf, *J. Polym. Sci.*, **1975**, *51*, 135.

[6] R.M. Nerem, *Med. Biol. Eng. Comp.* **1992**, *30*, CE-8-CE12.

[7] A. Micheli, A. Sperduti, A. Starita, A. M. Bianucci, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 202.

[8] A. M Bianucci, A. Micheli, A. Sperduti, A. Starita, in "*Soft Computing Approaches in Chemistry*", Sztandera L. M., Cartwright H. M., Eds., Springer-Verlag: Heidelberg 2003, p. 265.

[9] A. Micheli, "*Recursive Processing of Structured Domains in Machine Learning*", PhD thesis. TD-13/03, Dept. of Computer Science, University of Pisa, 2003.

[10] A. M. Bianucci, A. Micheli, A. Sperduti, A. Starita, *Appl. Int. J.* **2000**, *12*, 117.

[11] L. Bernazzani, C. Duce, A. Micheli, V. Mollica, A. Sperduti, A. Starita, M. R. Tiné, TR-04-16. Dept. of Computer Science, University of Pisa, 2004.

[12] C. Duce, "*Physical chemical methods in the rational design of new materials: QSAR and calorimetric approaches*", PhD Thesis, Dept. of Chemistry and Industrial Chemistry, University of Pisa, 2005.

[13] L. Bernazzani, C. Duce, A. Micheli, V. Mollica, A. Sperduti, A. Starita, M. R. Tiné, in preparation.

[14] Kollman P. A., 1996, *Acc. Chem. Res.*, *29*, 461–469.

[15] Eisenberg D., and McLachlan A. D., 1986, Nature (London), 319, 199–203.

[16] Jorgenson W. L., and Tirado-Rives J., 1988, J. Am. Chem. Soc., 110, 1657.

[17] A. Sperduti, A. Starita, *IEEE Transactions on Neural Networks* **1997**, *8*, 714.

[18] M. H. Abraham, J. A. Platts, *J. Org. Chem.* **2001**, *66*, 3484.

[19] J. Hine, P. K. Mookerjee, *J. Org.Chem.* **1975**, *40*, 292.

[20] S. Cabani, P. Gianni, V. Mollica, L. Lepori, *J. Solution Chem.* **1981**, *10*, 563.

[21] G. Klopman, H. Zhu, *J. Chem. Inf. Comput.Sci.* **2001**, *41*, 439.

[22] N. Nirmalakhandan, R. A. Brennan, R. E. Speece, *Water Res.* **1997**, *31*(6), 1471.

[23] D. Yaffe, Y. Cohen, G. Espinosa, A. Arenas, F. Giralt, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 85.

[24] N. J. English, D. G. Carroll, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1150.

[25] A. R. Katritzky, A. A. Oliferenko, P. V. Oliferenko, R. Petrukhin, D. B. Tatham, U. Uko Maran, A. L. Lomaka, W. E. Jr., Acree, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794.

[26] J. Bicerano, "*Prediction of polymer properties*", 3rd ed., Marcel Dekker, New York 2002.

[27] D. R. Wiff, M. S. Altieri, I. J. Goldfarb, *J. Polym. Sci: Polym. Phys. Ed.*, **1985**, *23*, 1165.

[28] D.W. Van Krevelen, "*Properties of Polymers-Their Estimation and Correlation with Chemical Structure*", 2nd ed., Elsevier, New York 1976.

[29] M. G. Koehler, A. J. Hopfinger, *Polymer* **1989**, *30*, 116.

[30] A.R. Katritzky, S. Sild, V.S. Lobanov, M. Karelson, *J. Chem. Inf. Comput. Sci* **1998**, *38*, 300.

[31] P. Camelio, C. C. Cypcar, V. Lazzeri, B. Waegel, *J. Polym. Sci.: Part A: Polym. Chem.* **1997**, *35*, 2579.

[32] S.J. Joyce, D.J. Osguthorpe, J.A. Padgett, G. J. Price, *J. Chem. Soc., Faraday Trans.*, **1995**, *91*, 2491.

[33] B.E. Mattioni, P.C. Jurs, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 232.

[34] B. G. Sumpter, D. W. Noid, *J. Thermal Anal.* **1996**, *46*, 833.