Journal of Molecular Graphics and Modelling 27 (2009) 797–802

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Evaluation of hierarchical structured representations for QSPR studies of small molecules and polymers by recursive neural networks

Carlo Bertinetto [a], Celia Duce [a], Alessio Micheli [b], Roberto Solaro [a], Antonina Starita [b], Maria Rosaria Tiné [a,*]

[a] Department of Chemistry and Industrial Chemistry, University of Pisa, via Risorgimento 35, 56126 Pisa, Italy
[b] Department of Computer Science, University of Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy

ARTICLE INFO

Article history:
Received 14 May 2008
Received in revised form 28 November 2008
Accepted 1 December 2008
Available online 9 December 2008

Keywords:
QSPR
Recursive neural network
Molecular representation
Cyclic structure
Cheminformatics
Poly(meth)acrylates
Ionic liquids

ABSTRACT

This paper reports some recent results from the empirical evaluation of different types of structured molecular representations used in QSPR analysis through a recursive neural network (RNN) model, which allows for their direct use without the need for measuring or computing molecular descriptors. This RNN methodology has been applied to the prediction of the properties of small molecules and polymers. In particular, three different descriptions of cyclic moieties, namely *group*, *template* and *cycle break* have been proposed. The effectiveness of the proposed method in dealing with different representations of chemical structures, either specifically designed or of more general use, has been demonstrated by its application to data sets encompassing various types of cyclic structures. For each class of experiments a test set with data that were not used for the development of the model was used for validation, and the comparisons have been based on the test results. The reported results highlight the flexibility of the RNN in directly treating different classes of structured input data without using input descriptors.

1. Introduction

In previous studies, we exploited a machine learning method based on recursive neural networks (RNNs) to perform quantitative structure–property relationship (QSPR) analysis [1–10]. This methodology was successfully assessed and compared to literature approaches by applying it to very different data sets of low molecular weight compounds [1–7] and polymers [7–10]. Unlike traditional QSPR approaches, the RNN-based method can directly process a hierarchical structured representation of molecules without the need of pre-defined molecular descriptors, thus providing a more direct, general, automatic and effective prediction tool.

A structured molecular representation is constituted by chemical fragments (groups) connected by a topological structure and shared by all or some of the data set compounds. Current well-known standard notation systems (such as SMILES [11–14] and InChI [15–17]) show that a general representation of chemical compounds can be based on hierarchical (acyclic) structures. The use of hierarchical labeled structures as class of data introduces both constraints and flexibility to the molecular representation. In particular, the choice of fragments, i.e. the level of detail by which chemical groups are represented in the structures, simultaneously determines the level of chemical information, the fragment sampling in the data set, the structure size and complexity. A successful representation seeks a good balance among these often conflicting issues.

In this respect, an important test bed for our method is given by the occurrence of cyclic structures. Cyclic moieties can be represented into hierarchical structures according to different solutions to the above-indicated balancing issue. The aim of this paper is to probe some of these options by empirical evaluation over different types of molecules using our well-established RNN method for hierarchical structures. These representations are applied to the prediction of the melting point of substituted ionic liquids and the glass transition temperature (Tg) of (meth)acrylic polymers. All investigated data sets contain various types of cyclic moieties. The influence of the different representations on the prediction results will be examined.

2. Method

The main characteristics of the RNN model for QSPR/QSAR analysis are fully described in Refs. [1–4] and outlined in Refs. [5–10]. RNNs have the intrinsic capability of dealing directly with labeled hierarchical structured representations of molecules (in

* Corresponding author. Tel.: +39 0502219268; fax: +39 0502219260.
E-mail address: mrt@dcci.unipi.it (M.R. Tiné).

1093-3263/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.12.001

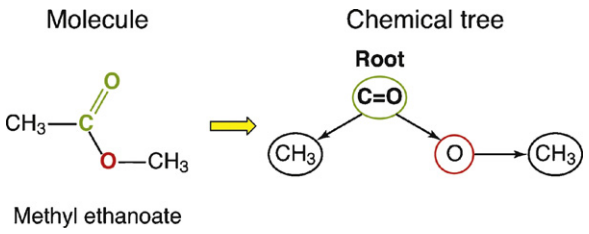

Fig. 1. Chemical structure and corresponding chemical tree of methyl ethanoate.

particular in the form of labeled rooted ordered trees), which are a vehicle of much more fruitful information than the traditional flat vectors of descriptors used by traditional QSPR/QSAR approaches. Moreover, RNN models have the ability to adaptively encode the input structures by learning from the given structure–property training examples. The learning algorithm allows the model for tuning the free parameters of the encoding process, while using a constructive (cascade-correlation based) approach that progressively and automatically adds the hidden units (HU) of the hidden layer during the training phase (see Refs. [1–4]). The use of an adaptive model avoids the need for an a priori definition of descriptors or of a similarity measure for structured data.

In order to achieve this goal, our RNN uses an encoding recursive process that mimics the morphology of each different input hierarchical structure. For each structure, the neural model encodes the substructures according to the molecular topology and to the content of each vertex label. Finally, the code developed by the model is mapped to the property values by the output part of the neural network.

Chemical compounds are represented as labeled rooted ordered trees by a 2D graph that can be obtained from their structural formula. Each low molecular weight molecule is partitioned into a limited number of atom groups and a priority scale is defined among groups [5]. Basically, each group corresponds to a tree vertex and each bond corresponds to an edge. To have a unique correspondence between graph and molecular structure, the total order on each vertex subtree is defined according to the priority scale and the root of the tree is fixed on the group with the highest priority (see Fig. 1).

To represent polymeric structures, we devised a model in which the 2D graph of the repeating unit, once partitioned into the appropriate atom groups, is rooted in an additional super-source vertex, "Start" (see Fig. 2). This group, besides closing one side of the repeating unit, can convey information on the average characteristics of macromolecules. In our experiments, "Start" accounted for the main chain stereoregularity recorded as molar fraction of *r* dyads [7–10]. A "Stop" group closes the other end of the repeating unit.

As explained in the introduction, the main focus of this paper is on the representation of the cyclic groups into hierarchical structures. For this purpose, different methods were devised and validated by their application to the prediction of two physical–chemical properties: the melting point (Tm) of ionic liquids,

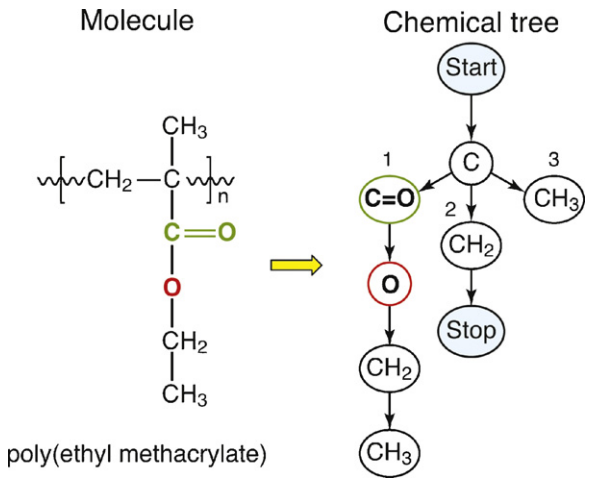

Fig. 2. Chemical structure and corresponding chemical tree of poly(ethyl methacrylate). The numbers indicate the children order.

namely pyridinium bromides, and the glass transition temperature (Tg) of (meth)acrylic polymers. For each property, two data sets were used, one being a subset of the other, reaching a total of four. Their size, characteristics and conventional name are given in Table 1.

Data set IL1 contained Tm of 117 pyridinium bromides, in which the only cyclic structures are the pyridinium ion and differently substituted phenyl rings. They were represented in the prediction experiment by rooting the chemical tree in the pyridinium ring, and by describing each cycle as a single vertex with either six (*Py*) or five (*Phenyl*) children, in order to account for all possible substitutions [9]. It must be stressed that the children are constituted by the chemical groups directly bonded to atoms constituting the ring skeleton (see for example Fig. 3A).

Data set PA1 deal with the Tg of 271 acrylic and methacrylic polymers, in which the only occurring cyclic moiety is the phenyl ring, either mono- or di-substituted at the 2–4 positions. As in the previous case, we represented this group as a single vertex (*Phenyl*) with three children (see for example Fig. 4A). The position of each child corresponds to the ring location of the attached group (1 = *ortho*, 2 = *meta*, 3 = *para*) [4,10].

The procedure of describing the whole cyclic structure as a single group has been named *group* representation. This approach has the computational advantage of generating rather compact trees: on average PA1 phenyl compounds are made of 14.7 vertices and span 8.4 vertices in depth. On the other hand, each cycle type has a different label, which needs enough sampling for the RecNN to learn it. It is hence suitable only for data sets containing largely sampled cycle types. When a greater ring variety is present, more general representation methods are needed. They all tackle the sampling issue by decomposing each cycle into atom groups occurring in most structures in order to limit the number of group labels with poor sampling.

Table 1
Summary of the used data sets for the validation of cyclic compound representations.

Data set	Chemical compounds	Target property	Number of compounds[a]	Cyclic moieties[b]
IL1	Pyridinium bromides	Tm	117 (117)	Phenyl (11), pyridinium (117)
PA1	(Meth)acrylic polymers	Tg	271 (110)	Phenyl (110)
IL2	Pyridinium bromides	Tm	126 (126)	Phenyl (11), pyridinium (126), pyridine (4), morpholine (1), cyclohexene (2)
PA2	(Meth)acrylic polymers	Tg	339 (178)	Phenyl (129), cycloalkyl (20), naphthyl (2), bornyl (4), adamantyl (3), other condensed cycles (2), dioxane (22), other heterocycles (8)

[a] The number in parentheses indicates the number of compounds in which a cyclic moiety is present.
[b] The number in parentheses indicates the number of compounds in which the moiety is present.

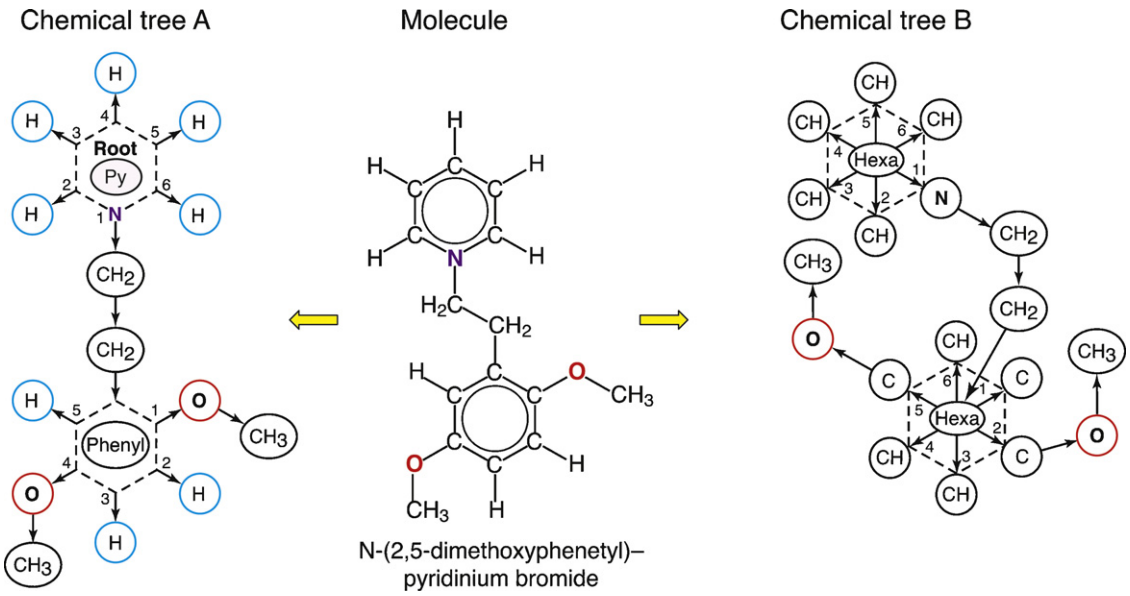

Fig. 3. Chemical structure (center) and the corresponding chemical trees of N-(2,5-dimethoxyphenetyl)pyridinium bromide: (A) each ring structure is represented as a single specific label (*group* representation); (B) all cycles are represented by a vertex (*Hexa*) with six children, which are the atom groups of the ring skeleton (*template* representation). The numbers indicate the children order.

The IL2 data set is made of the same compounds of IL1, added of 9 molecules containing also other cycles, such as pyridine, morpholine and cyclohexene. Although these cycles are set apart by the content of double bonds and/or by the presence of atoms different from carbon (oxygen, nitrogen), they all have the same size (6 vertices). We exploited this common feature by representing them all with a generic "six-edged-cycle template" group (*Hexa*) with 6 children, which are the atom groups constituting the ring skeleton (see for example Fig. 3B). This representation avoids the introduction of a great number of group labels with poor sampling by composing the cycle with atom groups occurring in almost all structures. The tree root is positioned on the *Hexa* group corresponding to the pyridinium ion; its first child is the nitrogen atom (N), while the first child of the other rings is the atom closest to the root. The other positions are ordered clockwise, the highest priority group having the lowest possible position number [6]. In this approach the cycle skeleton forms a *template* used to represent various types of cycles. The same template can be used for all cycles sharing matching skeleton geometry. Hence, the *template* approach is a compromise between the very specific *group* representation and a more general one that is presented by following.

In addition to PA1 compounds, the PA2 data set included 68 polymers containing cyclic moieties differing not only in the constituent atoms, but also in size and shape, such as various cycloalkyls and condensed cycles (see Table 1) with a number of ring vertices ranging from 3 to 17. They were represented by breaking the cycle and by adding a "cut1" vertex at both ends of the

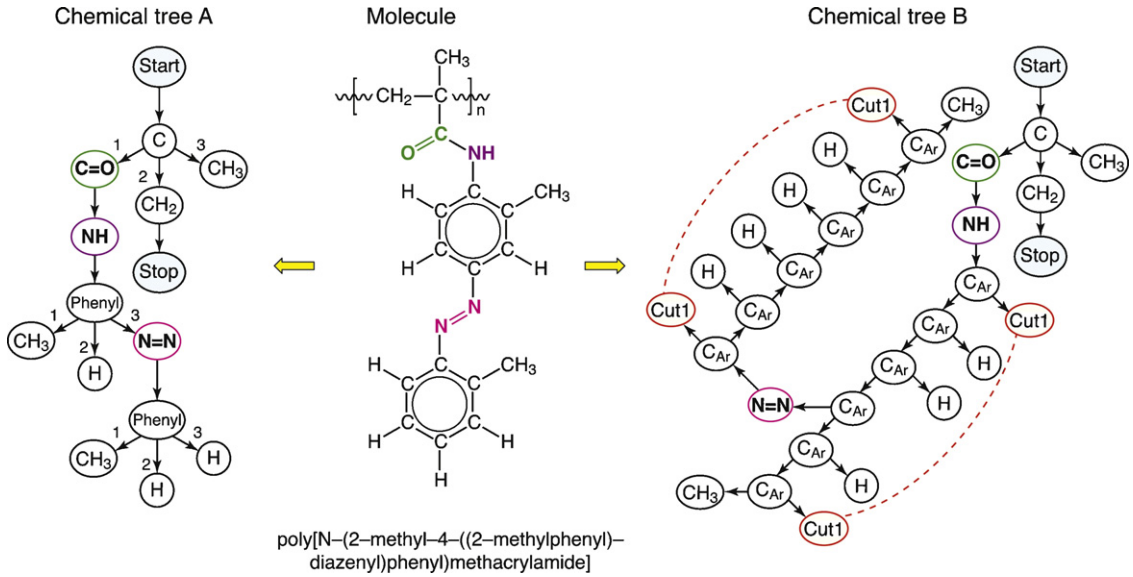

Fig. 4. Chemical structure and corresponding chemical tree of poly[N-(2-methyl-4-((2-methylphenyl)diazenyl)phenyl)methacrylamide]: (A) each cycle is represented as a single specific label (*group* representation); (B) cycles are opened and a "cut1" vertex is placed in the correspondence of the broken bond, here redrawn for clarity as a red dashed line (*cycle break* representation). In (A) the children order is indicated by numbers. In (B) the numbers are omitted to avoid overcrowding and the priorities (set according to UniqueSMILES) follow the same left-to-right (counterclockwise) order. C_{Ar} stands for aromatic carbon.

broken bond (see for instance Fig. 4B). In the case of condensed cycles, more than one bond must be cut and other labels are used (“cut2”, “cut3”, etc.). Identical labels match atom groups connected by the same broken bond. When rings are not condensed and no spiro moiety is present, the same “cut” number can be used repeatedly within a single molecule. Standard molecular representation formats such as, UniqueSMILES and InChI [11–17] are used to decide which ring bond to break and to assign the children order. This method, named *cycle break* representation, generates deeper trees. If applied to the phenyl compounds of PA1, the resulting graphs contain an average number of 28 vertices, with an average depth of 14.2 vertices. As a consequence, this representation requires a greater computational effort. On the other hand, the flexibility of this method allows for describing any cycle type regardless of its sampling, given that its constituting fragments are present in the data set. The generality of the *cycle break* representation is supported by the existence of standard molecular representation systems such as, UniqueSMILES and InChI that can describe any type of cyclic molecular structure by breaking some edges in a cyclic graph while maintaining the topological information.

To better compare the effect of the representation system on the prediction results, the *cycle break* representation was also tested on PA1 data set [10].

3. Experiments

Five experiments were carried out to investigate the prediction of two molecular physical properties, i.e. the melting point (Tm) and the glass transition temperature (Tg), over four data sets by using different representation methods. For each experiment the results are summarized by an ensemble averaging over 16 (=2^4) different trials (i.e. training of the model), all with the same training/test split. Specifically, the number of recursive hidden units, the mean absolute error (MAE), the squared correlation coefficient (R^2) and the standard error deviation (S) are reported in Table 2. For each class of experiments the test set validation has been based on data that were not used for the development and training of the model. Since data sets IL1 and IL2 were taken from a work by Katritzky et al. [18] the same partition between training and test, with the latter being about one-third of the total, was maintained here. In data sets PA1 and PA2 the test set was built through a random selection of about one-fifth of all the compounds, though being representative of the chemical moieties present in the whole data set. The distribution of training and test target values throughout data sets IL2 and PA2 is plotted in Fig. 5. A plot of the results of experiment Tg1, for both training and test phase, is shown as an example in Fig. 6.

In experiment Tm1 [9] the IL1 data set was split into training and test sets of 80 and 37 molecules, respectively. The occurring cyclic structures, pyridinium ion and variously substituted phenyl rings, were described by the *group* representation as explained in Section 2. Experiment Tm2 [6] was carried out on the IL2 data set

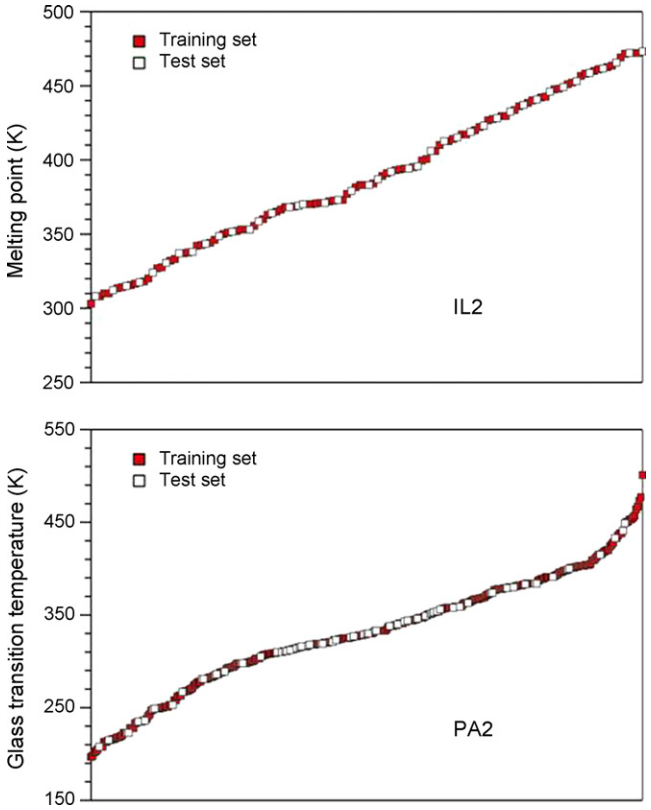

Fig. 5. Distribution of training and test target values in IL2 (top) and PA2 (bottom) data sets.

divided in 84 and 42 molecules for training and testing, respectively. The *template* representation was used for cyclic moieties. For either datasets, the RNN model achieved almost the same performance and the recorded MAE and S values are similar to those obtained by other literature QSPR methods that employ pre-defined molecular descriptors [18,19]. The more detailed structural information introduced in Tm2 by the *template* representation brings about an extension of the input space that is not compensated by a proportional increase of molecular data set size. Nonetheless, the result accuracy recorded for Tm2 is not worse than that of Tm1. Of course, this explicit decomposition of cycles has the clear advantage of conveying more useful information while allowing for increasing the class of molecules that can be represented in the data set.

Experiment Tg1 [10] was performed on data set PA1, subdivided into 217 and 54 polymers for training and testing, respectively. The phenyl rings contained in 110 of the data set compounds were described with the *group* representation. Tg2 was carried out on PA2, split into 272 training set and 67 test set compounds. The many different cycle types were represented by

Table 2
Average prediction results obtained by application of different representation models to the investigated data sets.

Exp.	Data set	Cycle represent.	Number of compounds		HU	Results					
			Training	Test		Training set			Test set		
						MAE (K)	R^2	S (K)	MAE (K)	R^2	S (K)
Tm1	IL1	*Group*	80	37	6	11.0	0.92	13.9	25.0	0.62	29.6
Tm2	IL2	*Template*	84	42	14	10.4	0.93	12.9	25.1	0.60	30.4
Tg1	PA1	*Group*	217	54	17	8.3	0.97	11.2	15.6	0.84	21.1
Tg2	PA2	*Cycle break*	272	67	32	7.1	0.98	9.5	18.5	0.80	24.0
Tg1b	PA1	*Cycle break*	217	54	21	8.3	0.97	11.0	15.7	0.85	20.4

HU = (recursive) hidden units; MAE = mean absolute error; R^2 = squared correlation coefficient; S = standard deviation; K = Kelvin.

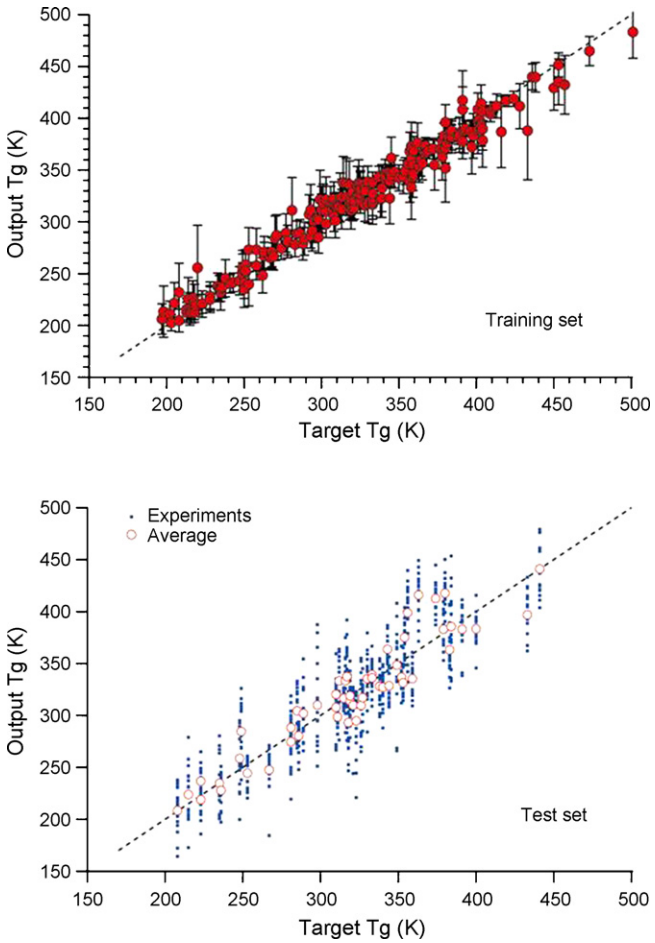

Fig. 6. Results of experiment Tg1. Top: training set; the circles are averages over the 16 trials, the error bars show the standard deviation. Bottom: test set; the small dots show the results of each trial, whose average is indicated by circles.

the *cycle break* method in accordance with the cuts and children order of UniqueSMILES. In both cases, the recorded MAE and *S* values were again comparable to those obtained by most ad hoc literature methods for polymer property prediction [20–26]. However, the Tg2 test set resulted in an increase of both MAE and *S* values, as compared with Tg1, of almost 3 Kelvin (K). This behavior can be attributed to either the representation change or the greater complexity of PA2 data set.

To shed light on this point, a further experiment, named Tg1b, was run on PA1 by using the *cycle break* representation (again according to UniqueSMILES) [10]. Only a slight change of MAE and *S* was detected as compared to Tg1. This result suggests that the observed error increase should be attributed rather to the greater complexity of PA2 data set as compared to PA1 than to the more demanding prediction task required by a more general representation design. However, the comparison of the results relevant to cyclic and acyclic compounds, respectively, indicates that the type of representation does affect the RNN performance. The behavior of the *group* representation (Tg1) is very similar for the two classes, their MAE and *S* being set apart by only 1 K. Both MAE and *S* values of cyclic compounds were instead more than 4 K greater than those of acyclic ones when using the *cycle break* representation (Tg1b). Though not yet fully understood, this behavior may be explained by considering that the structures generated for both cyclic and acyclic compounds have more or less the same size when using the *group* representation; on the other hand, *cycle break* gives rise to much deeper trees for cyclic structures.

4. Conclusions

The reported experiments demonstrate the RecNN flexibility in the treatment of molecular structured data of different types. In particular, the labeled tree representation can be exploited to treat very different molecular structures by finding a balance between structural detail and molecular sampling in each investigated data set. The designed cycle representations span from a very specific one (*group*), dedicated to a restricted class of molecules, to renditions of increasing generality (*template* and *cycle break*). Comparison of the results obtained with different cycle representations on identical or similar data sets highlights a very limited error increase in going from a specific technique to more general ones. The first option might however prove useful for restricted classes of compounds, as it provides more accurate results at less computational effort, due to the use of simpler structures.

This variety of helpful choices emphasizes the adaptability introduced by the RNN approach to the QSPR based on structures. This flexibility allows the designer for tuning the level of structural detail to the characteristics of the investigated molecular data set while using the same computational approach.

Acknowledgement

The financial support by MIUR Cofin Project and by Fondazione Cassa di Risparmio di Pisa 2006/172 Project is gratefully acknowledged.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2008.12.001.

References

[1] A. Micheli, A. Sperduti, A. Starita, A.M. Bianucci, A novel approach to QSPR/QSAR based on neural networks for structures, Studies in Fuzziness and Soft Computing 120 (2003) 265–296 (Soft Computing Approaches in Chemistry).
[2] A.M. Bianucci, A. Micheli, A. Sperduti, A. Starita, Application of cascade correlation networks for structures to chemistry, Appl. Int. J. 12 (2000) 117–146.
[3] A. Micheli, A. Sperduti, A. Starita, A.M. Bianucci, Analysis of the internal representations developed by neural networks for structures applied to quantitative structure–activity relationship studies of benzodiazepines, J. Chem. Inf. Comput. Sci. 41 (2001) 202–218.
[4] A. Micheli, A. Sperduti, A. Starita, A. Bianucci, Design of new biologically active molecules by recursive neural networks, in: Proceedings of the International Joint Conference on Neural Networks, vol. 4, 2001, pp. 2732–2737.
[5] L. Bernazzani, C. Duce, A. Micheli, V. Mollica, A. Sperduti, A. Starita, M.R. Tiné, Predicting physical chemical properties of compounds from molecular structures by recursive neural networks, J. Chem. Inf. Model. 46 (2006) 2030–2042.
[6] R. Bini, C. Chiappe, C. Duce, A. Micheli, R. Solaro, A. Starita, M.R. Tiné, Ionic liquids: prediction of their melting points by a recursive neural network model, Green Chem. 10 (2008) 306–309.
[7] C. Duce, A. Micheli, R. Solaro, A. Starita, M.R. Tiné, Prediction of chemical–physical properties by neural networks for structures, Macromol. Symp. 234 (2006) 13–19.
[8] C. Duce, A. Micheli, A. Starita, M.R. Tiné, R. Solaro, Prediction of polymer properties from their structure by recursive neural networks, Macromol. Rapid Commun. 27 (2006) 712–716.
[9] C. Bertinetto, R. Bini, C. Chiappe, C. Duce, A. Micheli, R. Solaro, A. Starita, M.R. Tiné, Recent advances in the representation of molecular structures for RecNN–QSPR analysis, in: T. Simos, G. Maroulis (Eds.), Lecture Series on Computer and Computational Sciences, vol. 7, Brill Academic Publishers, Leiden, 2006 , pp. 1352–1355.
[10] C. Bertinetto, C. Duce, A. Micheli, R. Solaro, A. Starita, M.R. Tiné, Prediction of the glass transition temperature of (meth)acrylic polymers containing phenyl groups by recursive neural network, Polymer 48 (2007) 7121–7129.
[11] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[12] http://www.daylight.com/smiles/index.html.
[13] H.L. Morgan, Generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service, J. Chem. Doc. 5 (1965) 107–113.
[14] D. Weininger, A. Weininger, J.L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, J. Chem. Inf. Comput. Sci. 29 (1989) 97–101.

[15] B.D. McKay, Practical graph isomorphism, Congressus Numerantium 30 (1981) 45–87.
[16] A. McNaught, The IUPAC international chemical identifier: InChI—a new standard for molecular informatics, Chem. Int. 28 (2006) 12–14.
[17] S.E. Stein, H.R. Heller, D.V. Tchekhovskoi, The IUPAC Chemical Identifier—Technical Manual, http://www.iupac.org/inchi/.
[18] A.R. Katritzky, A. Lomaka, R. Petrukhin, R. Jain, M. Karelson, A.E. Visser, R.D. Rogers, QSPR correlation of the melting point for pyridinium bromides, potential ionic liquids, J. Chem. Inf. Comput. Sci. 42 (2002) 71–74.
[19] G. Carrera, J. Aires-de-Sousa, Estimation of melting points of pyridinium bromide ionic liquids with decision trees and neural networks, Green Chem. 7 (2005) 20–27.
[20] J. Bicerano, Prediction of Polymer Properties, Marcel Dekker, New York, 2002.
[21] A.R. Katritzky, S. Sild, V.S. Lobanov, M. Karelson, Quantitative structure–property relationship (QPSR) correlation of glass transition temperatures of high molecular weight polymers, J. Chem. Inf. Comput. Sci. 38 (1998) 300–304.
[22] R. Garcia-Domenech, J.V. de Julián-Ortiz, Prediction of indices of refraction and glass transition temperatures of linear polymers by using graph theoretical indices, J. Phys. Chem. B 106 (2002) 1501–1507.
[23] S.I. Joyce, D.J. Osguthorpe, J.A. Padgett, G.J. Price, Neural network prediction of glass–transition temperatures from monomer structure, J. Chem. Soc., Faraday Trans. 91 (1995) 2491–2496.
[24] B.G. Sumpter, D.W. Noid, On the use of computational neural networks for the prediction of polymer properties, J. Therm. Anal. 46 (1996) 833–851.
[25] C.W. Ulmer I.I., D.A. Smith, B.G. Sumpter, D.I. Noid, Computational neural networks and the rational design of polymeric materials: the next eneration polycarbonates, Comput. Theor. Polym. Sci. 8 (1998) 311–321.
[26] B.E. Mattioni, P.C. Jurs, Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks, J. Chem. Inf. Comp. Sci. 42 (2002) 232–324.