Available online at www.sciencedirect.com

ScienceDirect

Polymer 48 (2007) 7121–7129

polymer

www.elsevier.com/locate/polymer

Prediction of the glass transition temperature of (meth)acrylic polymers containing phenyl groups by recursive neural network

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 Informatics, University of Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy*

Received 30 May 2007; received in revised form 26 September 2007; accepted 26 September 2007
Available online 4 October 2007

Abstract

A recursive neural network QSPR model that can take directly molecular structures as input was applied to the prediction of the glass transition temperature of 277 poly(meth)acrylates. This model satisfactorily predicted the chemical–physical properties of high and low molecular weight acyclic compounds. However, side-chain benzene rings are present in about one half of the selected polymers. In order to render cyclic structures, the molecular representation through hierarchical structures was extended by two methods, named *group* and *cycle breaking*, respectively. The latter approach exploits standard unique molecular description systems, i.e. Unique SMILES and InChI. In all cases the prediction was very good, with 15–16 K mean absolute error and 19–21 K standard deviation. This result confirms the robustness of our method with respect to the inclusion of different structures. Moreover, the good performance of the *cycle breaking* representation paves the way for the investigation of data sets that contain a variety of poorly sampled cyclic structures.

Keywords: Glass transition temperature; Poly(meth)acrylates; Recursive neural network

1. Introduction

The ability to predict the physical–chemical properties of polymeric materials from molecular structure has got great importance in designing polymers. The applications of most of the familiar polymers with relatively simple repeating unit structures have reached their limits, so that the chemical structures of polymers suitable for advanced applications have increased in complexity. Their synthesis and experimental characterization have become more and more expensive and time consuming. Therefore it is increasingly necessary to develop predictive methods to evaluate candidates for specific applications. In this framework, efforts have been spent on the development of Quantitative Structure–Property Relationship (QSPR) techniques. The aim of QSPR is to find an appropriate function, which given a proper representation of a molecule can predict a selected property. A property that has often been used as a benchmark for new prediction methods is the glass transition temperature (T_g) because of the availability of a large number of experimental data. The T_g is also of great technological significance, since it determines the utilization limits of polymeric materials. With respect to the prediction of this property, a great number of QSPR methods are available [1–16] and a more detailed description is provided in Ref. [17]. Summarizing, they can be classified into two main classes: group additive property (GAP) methods and systems that use molecular descriptors.

Because of several issues, both methods are not really suited for predicting the properties of very different classes of chemical compounds. 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 (RNN) methods that take labelled hierarchical structures, such

* Corresponding author. Tel.: +39 050 2219 284; fax: +39 050 2219 260.
E-mail address: rosola@dcci.unipi.it (R. Solaro).

0032-3861/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2007.09.043

as rooted trees, as input, allowing us for a variably sized representation of molecules.

Successful application of the RNN model was achieved in predicting the pharmacological activity of substituted benzodiazepines [18–21], the boiling points of linear and branched alkanes [19–21], and the prediction of standard free energy of solvation in water of organic compounds [22–24]. More recently, the method was extended to macromolecules by investigating the T_g of a set of acyclic (meth)acrylic polymers [17,25]. The representation of each polymer was based on the 2D structure of their repeating unit, which was capped at both ends by two fictitious groups named "Start" and "Stop". The "Start" group contains also information regarding the average polymer structure. This RNN model correctly predicted the dependence of the target property on both monomer structure and stereoregularity of a series of acyclic polymers.

In the present paper we applied the RNN method to the prediction of the T_g of an extended polymer data set including both acyclic and cyclic structures such as phenyl, biphenyl, and azophenyl groups. In particular, we focused specifically in assessing methods to represent cycles in polymers while preserving the basic approach founded on hierarchical structures. This approach allows for making a direct comparison with the original strategy adopted for acyclic compounds. The description of phenyl moieties was carried out by two techniques, named *group* and *cycle breaking* representation, respectively. Both of them were applied to the same data set in order to have a better comparison of the results, although the characteristics of the *cycle breaking* representation make it suitable for data sets containing a larger variety of cyclic structures. Our aim was also to show the adaptability provided by the adopted structure representation, even when constrained to a hierarchical form. This flexibility allows for describing molecular structures at different detail levels, including the information on the occurrence of cycles.

2. Method

A detailed description of the adopted RNN model was already reported elsewhere [20,21,26]. This RNN is a generalization of the well-known feed-forward neural network that *directly* deals with variable-size structured data. This data type cannot be treated right away by feed-forward neural networks. The graphical objects used to describe chemical structures are labelled trees, a subclass of DPAG (Directed Positional Acyclic Graph) in which a finite out-degree k is defined, i.e. k is the maximum number of edges leaving a vertex, or the number of children of a vertex. If $T(r)$ is a tree rooted in the vertex r (e.g. the "Start" vertex in Fig. 1, which has a unique child "C"), the *sub-tree* $T(v)$ is the tree rooted in v induced by the descendants of v (e.g. the sub-tree rooted in "C" in Fig. 1). Each children of a vertex v is characterized by its position, which distinguishes the 1st child, 2nd child, etc., up to the kth child of v.

In labelled trees there are labels attached to each vertex. The set of label symbols depends upon the devised chemical groups. The chemical symbols are represented by numerical vector labels based on a "1-of-n" coding scheme for categorical data, which has the basic aim to discriminate among different symbols (see Section 3 for details).

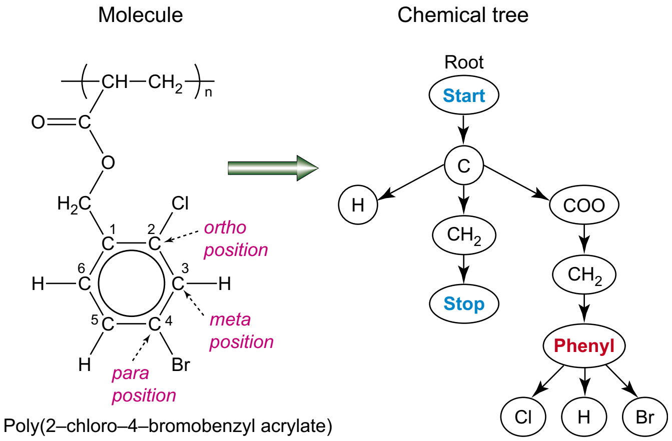
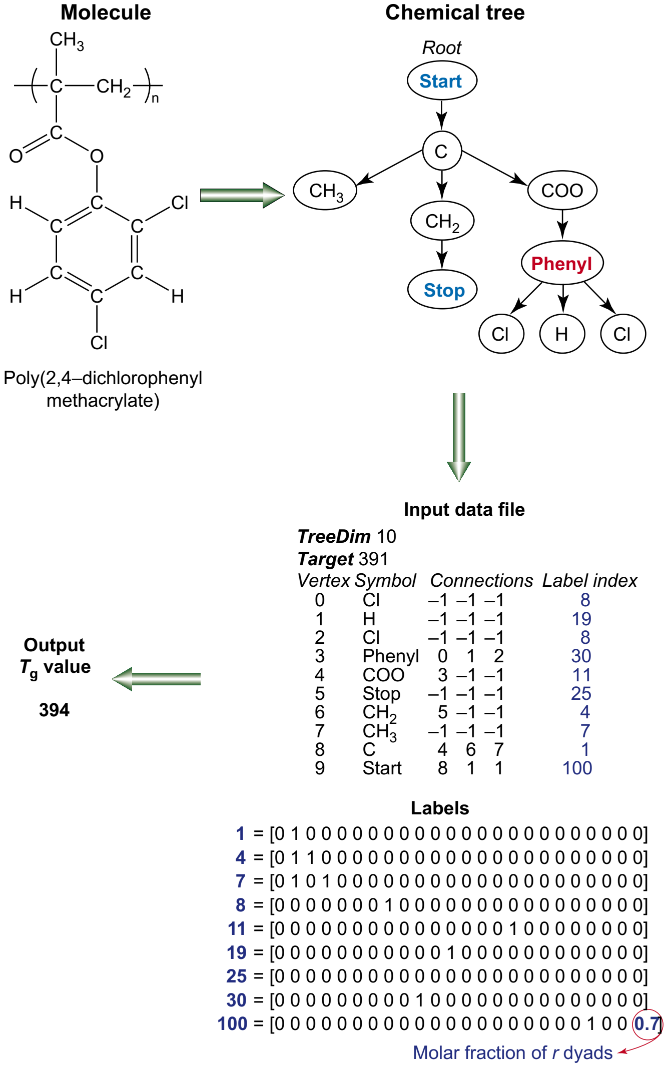


Fig. 1. Group representation of poly(2-chloro-4-bromobenzyl acrylate).

The RNN exploits a recursive encoding process, which mimics the morphology of each input structure. For each vertex of the tree, the model computes a numerical code by using information of both the vertex numerical label and, recursively, the code of the sub-tree descending from the current vertex. This process computes a code for the whole molecular structure, which is able to consider both vertex labels and the structure topology. For further details see Refs. [18,20–22,24,27]. Besides computational details, it is important to stress here that by this encoding process the RNN can directly treat an input in the form of a labelled tree (e.g. the chemical trees shown in Figs. 1 and 2 with vertexes labelled by numerical vectors): the RNN visits and encodes the tree vertex-by-vertex through a recursive process. The code is then mapped to the output property value by the same RNN model. Hence, the RNN is able to realize a direct mapping between the chemical input tree and the output value (T_g values are represented by real numbers). Apart from details on the specific molecular representation introduced in Section 3, this direct mapping is graphically shown in Fig. 2. Fig. 2 also shows the implementation used to store the chemical tree into the input file, which is based on a connection table of the structure. Note that the connection table is not used as an input vector. Rather it conveys the connections among vertexes and their sub-trees that are used by the RNN recursive visiting process.

The encoding and mapping free parameters of the neural network are adapted to the task through the learning algorithm on the basis of training examples. By this process, the RNN models a direct and *adaptive* relationship between molecular structures and target properties. In particular, we use a constructive approach to realize the RNN architecture (Recursive Cascade Correlation). The process is incremental, meaning that neural Hidden Units (HU) are progressively added until the errors among the outputs of the training examples and their target values are below a tolerance determined by the operator.

The major advantage introduced by RNN is that the encoding of the input structured representations can be learnt



Input data file

```
TreeDim 10
Target 391
Vertex Symbol   Connections  Label index
   0    Cl      –1 –1 –1          8
   1    H       –1 –1 –1         19
   2    Cl      –1 –1 –1          8
   3    Phenyl   0  1  2         30
   4    COO      3 –1 –1         11
   5    Stop    –1 –1 –1         25
   6    CH2      5 –1 –1          4
   7    CH3     –1 –1 –1          7
   8    C        4  6  7          1
   9    Start    8  1  1        100
```

Labels

```
  1 = [0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
  4 = [0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
  7 = [0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
  8 = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 11 = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0]
 19 = [0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0]
 25 = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
 30 = [0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
100 = [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0.7]
```

Fig. 2. Structure, chemical tree, input data file, and numerical labels associated with poly(2,4-dichlorophenyl methacrylate). The input data file contains the dimension of the tree (number of vertexes), the value of the target property, and the connection table. Column 1 of the connection table reports the order number identifying a specific vertex, which corresponds to the group indicated in column 2; columns 3–5 indicate the presence of a "child" identified by its order number (−1 means no child); column 6 reports the index of the associated numerical label.

according to the given QSPR task. Hence, compared to traditional QSPR approaches, the RNN can automatically generate by learning the specific structural descriptors (numerical code) for the particular task to be solved. As a result, no *a priori* definition/calculation and/or selection of input properties are needed.

3. Molecular representation

The adopted chemical tree representation is an extension of the one used in previous works. Molecular graphs were built by splitting the compounds into atomic groups that constitute the vertexes of the tree [24]. A label was assigned to each group, and priority rules which univocally determined the tree root and the order of the sub-trees were set to have a biunique correspondence between graph and chemical structure. This representation was later extended to describe polymers [17]: each repeating unit was decomposed using almost the same atomic groups, labels and priority rules, but the tree root was positioned on an additional super-source vertex called "Start", which could hold also information about the overall polymer structure (molecular weight, stereoregularity, etc.). The other end of the unit was capped by another group called "Stop", with the only purpose of closing the structure.

The representation is very flexible and allows for defining the molecular fragmentation with the selected detail level. The used atomic groups are: C, C aryl, C≡C, CH_2, CH_2–CH_2, CH_3, H, C=O, COO, CF_2, CF_3, F, Cl, Br, I, CN, N, NH, NH_2, N=N, NO_2, NO_3, O, OH, S, S=O, SO_2, Start, Stop, Phenyl and cut1. The last two are involved in the representation of phenyl groups that will be discussed in this section. Most fragments correspond to common chemical groups, whereas Start, Stop and cut1 are fictitious groups. A numerical label is associated with each group symbol, discriminating among different symbols by the means of a "1-of-*n*" coding scheme. Besides this, sharing of "1" among different labels represents the similarity of chemical groups (for instance CH_2 and CH_3 in Fig. 2). Moreover, the "Start" label conveys information on main chain stereoregularity. See Fig. 2 as an example reporting the detailed data representing an input chemical tree, and Ref. [25] for other instances related to polymeric compounds.

The key innovation of this paper is the assessment of methods that allow for describing polymer containing cyclic moieties. Among the possible ones, the choice was restricted to techniques that would not overset the model used so far. Two methods, which were named *group* and *cycle breaking* representation, respectively, have been used. The first one consists of associating each cyclic group with a molecular vertex [28]. The benzene ring is the only cyclic structure present in our data set; it was represented by the "phenyl" group as shown in Fig. 1. The unsubstituted benzene ring is constituted of six carbon atoms arranged in a hexagonal cycle; each carbon atom is also linked to a hydrogen atom outside the cycle. Phenyl group carbons are numbered as follows: number one is assigned to the carbon atom connected to the polymer structure and the remaining carbons are numbered consecutively, clockwise or counterclockwise in order to give the lowest possible number to substituted carbon atoms. Phenyl rings containing atoms other than hydrogen linked to ring carbons 2–4 are said to be *ortho*, *meta*, and *para* substituted, respectively. All phenyl groups in our data set were either monosubstituted or disubstituted in *ortho*–*meta*, *ortho*–*para* or *meta*–*para* positions. In other words, there was no need to refer to 5- and 6-position. Therefore, we decided to assign this group as an out-degree $k = 3$, that is, the "phenyl" group has only three children. The order of the children corresponded to their rank in the ring (1st child = *ortho*, 2nd child = *meta*, 3rd child = *para*). This ordering allows for discriminating compounds differing only because of substituent position. Simplicity is the main advantage of this representation: it generates rather compact trees that are more easily computed by the RNN. However, the RNN needs enough sampling of each cyclic structure in order to learn it.

The second method (*cycle breaking*) exploits the possibility of deriving a connected hierarchical structure directly by breaking some edges in cyclic graphs, while maintaining the topological information such as the occurrence of cycles in the original graph. *Cycle breaking* is supported by current standard molecular representation formats. This method consists of ideally breaking the cycle at a certain point. The resulting structure is then written as a tree after placing a ‘‘cut1’’ group at both sides of the broken bond (Fig. 3). In the case of condensed cycles, more than one bond must be cut and other labels must be used (‘‘cut2’’, ‘‘cut3’’, etc.). Atoms connected by the same broken bond are matched by identical labels. When rings are not condensed and no spiro moiety (where two cycles share just one atom) is present, the same cut number can be used repeatedly within a single molecule. Accordingly, no more fragments other than ‘‘cut1’’ are needed to represent the polymers of our data set, although several samples contain two or three rings. This representation differs in many aspects from the previous one. In most cases, it leads to rather deep trees. On the other hand, the generality of the cycle cutting approach allows for dealing with any cyclic structure (Fig. 4) independently of its sampling, given that its constituting fragments are enough represented in the data set.

The issue of representation uniqueness gets really important in cycle cutting. One has to single out which bond must be cut, and which order must be given to the sub-trees. Indeed, graphs associated with cyclic structures can become quickly

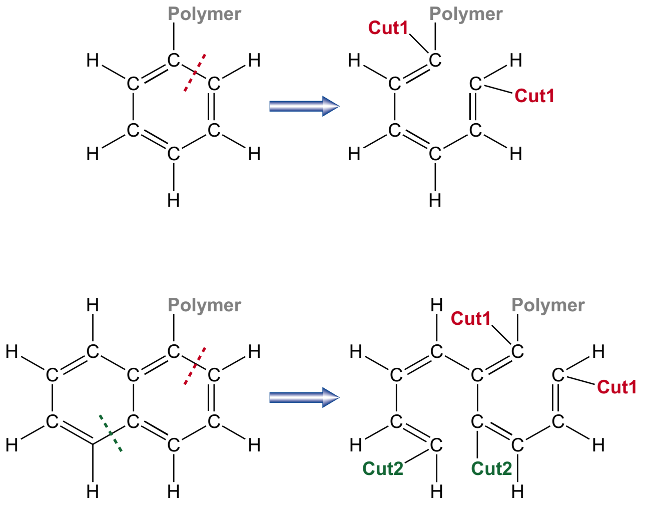


Fig. 3. Schematic representation of cycle cutting of polymer structures containing one cycle (top) and two condensed cycles (bottom).

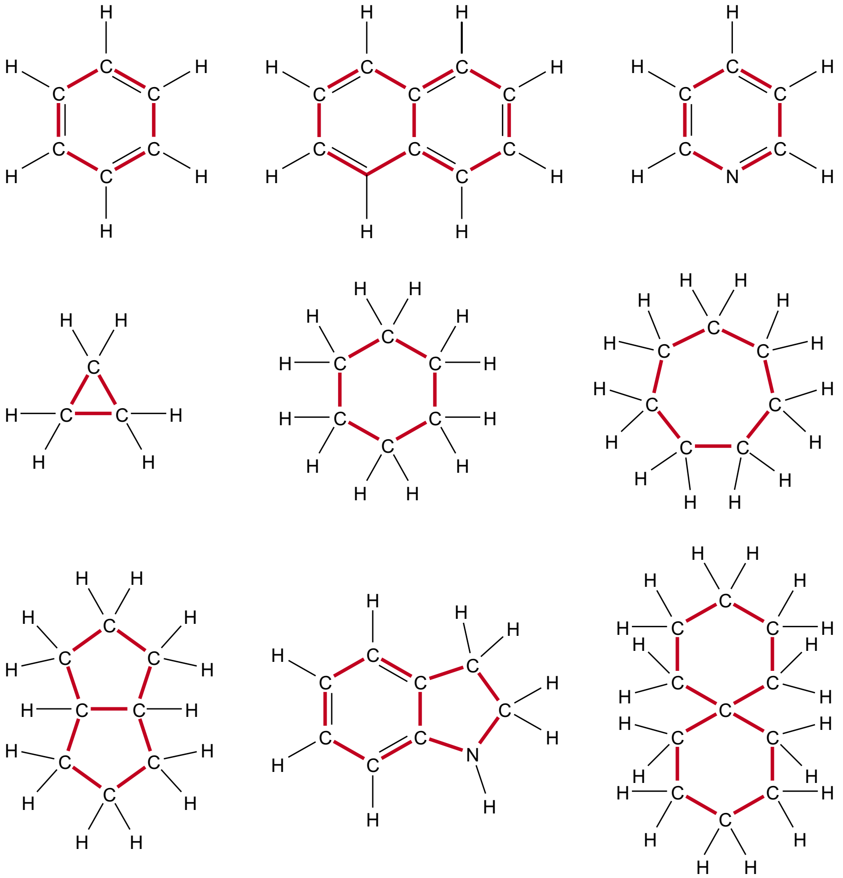


Fig. 4. A few examples of cycles that can be represented by the cycle cutting approach: six-member aromatic cycle (top left), condensed six-member aromatic cycles (middle top), six-member aromatic cycle containing a nitrogen heteroatom (top right), saturated cycles containing a different number of vertexes (middle), condensed aliphatic cycles (bottom left), mixed aromatic–aliphatic cycles containing a nitrogen heteroatom (middle bottom), and a spiro form (right bottom).

very complicated. Accordingly, we decided not to face the problem directly and to exploit standard molecular representation models. In particular, we relied on Unique SMILES [29,30] and InChI [31–33], two well-known representation models that produce a unique hierarchical structure, usually written in string form, for each chemical compound.

SMILES (Simplified Molecular Entry Line System) is a molecular representation system developed by Daylight that translates molecules into strings. Canonical Unique SMILES representation can be obtained by an algorithm by Weininger [34,35]. InChI (International Chemical Identifier) is a method recently developed by IUPAC. None of these systems supports an explicit representation of polymers [30,33], but we overcame this limitation easily. The repeating unit was given as input to either standard method with the artificial addition of a long enough aliphatic chain to the atom that we chose as tree root (Fig. 5). Both standards tend in fact to start the graph from the longest non-branched aliphatic chain.

Since it was not possible to decide beforehand which one of the standard systems performed better in our application, we tested them both. Moreover, we did not know whether to give up completely our previous priority rules or not. Accordingly, the standard method was applied to the whole structure in some experiments, whereas in other cases our priority rules and standard methods were applied to the linear and the cyclic parts of the molecule, respectively.

4. Experiments

The T_gs of 110 poly(meth)acrylic esters containing phenyl rings [36–75] were added to the data set of 167 acyclic poly(meth)acrylates used in the last three experiments in Ref. [17]. Whenever disagreeing data were found, preference was given to sources in which the polymer synthesis was specified, tacticity was determined by NMR, T_g was measured by DSC (with a heating rate as low as possible), molecular weight was as close as possible to 200.000 MU. Information on main chain stereoregularity was stored in the ‘‘Start’’ label as explained in Ref. [17]. Experimental NMR tacticity data were used whenever available; in other cases, the molar fraction of r dyads was set at 1 for samples indicated as syndiotactic, 0 for isotactic samples, and 0.6 or 0.7 for atactic polyacrylates or polymethacrylates, respectively. Table 1 in Supplementary data lists T_g, tacticity and literature source

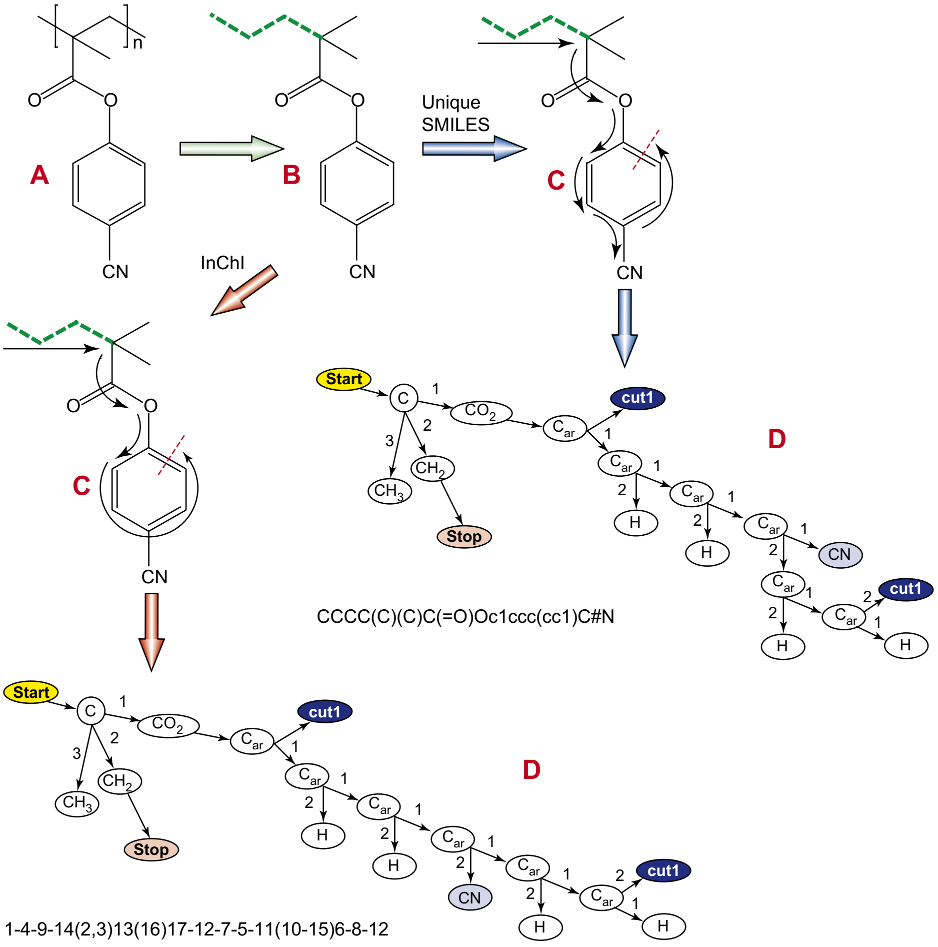


Fig. 5. Tree generation for poly(4-cyanophenyl methacrylate). (A) Input structure; (B) the bonds across the parentheses are erased, and a linear aliphatic chain (thick dashed bonds), which has to be longer than any other chain present in the rest of the compound, is added to the side where we want the tree to start from; (C) a standard representation system (either Unique SMILES or InChI) is used to process the ‘‘molecule’’ obtained in B, defining branches and disconnections. The corresponding trees and the resulting encoding strings are shown in (D).

for all polymers containing cyclic moieties, whereas data relevant to acyclic samples are reported in Ref. [17].

The whole polymer set was divided into disjoint training and test sets for learning and validation processes, respectively. Test samples are representative of the different functional groups, molecular size, and topology of the investigated repeating units. A ‘‘guess’’ set was created to include the samples whose molecular features are scarcely represented in the data set or whose target is highly uncertain. The polymers that were included in the guess set are poly(acrylic acid), poly(*N*-secbutylacrylamide), poly(*N*-terbutylmethacrylamide) and syndiotactic samples of poly(isopropyl methacrylate), poly(ethyl methacrylate) and poly(isobutyl methacrylate). The final data set consisted of 217, 54 and 6 samples in the training, test and guess set, respectively. The maximum error tolerance for the learning process was set at 60 K for all the experiments, since this value is comparable with the literature data spread of the target property. The target T_g ranged from 197 to 501 K and from 208 to 441 K in the training set and test set, respectively.

Five experiments were performed and their results are listed in Table 1. In Exp. 1 the benzene rings were described through the *group* representation, in Exps. 2–5 through the *cycle breaking* one. The difference among different experiments consisted in the adopted conventions and standard representation systems:

Exp. 1 – group representation for the cyclic part, our priority rules for the linear one;
Exp. 2 – Unique SMILES for the cyclic part, our priority rules for the linear one;
Exp. 3 – Unique SMILES for the whole compound but the polymer main chain that cannot be represented by the standard system;
Exp. 4 – InChI for the cyclic part, our priority rules for the linear one;
Exp. 5 – InChI for the whole compound but the polymer main chain.

For each experiment, the complete list of training, test, and guess sets is given in Supplementary data, where the target T_g, the molar fraction of *r* dyads, the mean calculated output, and the relative standard deviation, σ, over 16 trials are reported for each polymer sample.

5. Results

The random initialization of the RNN connection weights can lead to different outcomes because of the use of a stochastic gradient-based technique to solve a least mean square problem. In order to have a significant result, in each experiment 16 trials were carried out for the RNN simulation and the results were averaged over the different trials. For each compound we computed the Absolute Average Error (AAE) and the standard deviation (σ) over the 16 trials. On the whole data set we computed the Mean Absolute Error (MAE, which is calculated over all AAEs of the data set), the maximum absolute error (MAX), the correlation coefficient (R) and the standard deviation (S). These data are reported in Table 1 together with the number of RNN hidden units (HU) and the number of samples (N). The results [17] obtained for a restricted data set that does not contain cyclic structures (Exp. A) are also reported for comparison.

Exp. A and Exp. 1 show very similar MAE and S, although the data set of Exp. 17 is much more heterogeneous because of the addition of aromatic samples. The MAE is even lower, although part of this improvement can be attributed to the transfer of poly(ethyl methacrylate) and poly(isobutyl methacrylate), which have extrapolated targets and showed very high AAE, from test to guess sets. The prediction is better for acyclic compounds (MAE = 15.11 K) than for aromatic ones (MAE = 16.02 K). This is understandable, since the latter polymers contain chemical groups that are absent in the first group. The needed HU are 17, only two more than the previous experiment.

However, it must be stressed that some of the atomic groups used in Exp. A to describe the molecules were changed in subsequent experiments. Indeed, some preliminary runs were carried out to test different molecular fragmentations, exploiting the flexibility of the structure representation concerning the information details. It was observed that fragment compressions, such as two consecutive ‘‘CH_2’’ into ‘‘$CH_2{-}CH_2$’’ and ‘‘C=O’’ and ‘‘O’’ into ‘‘COO’’, improved the prediction (MAE decreased by ≈2 K) and reduced the computational load (the RNN needed about 13 HU less). In the latter case, we transmitted chemical information to the RNN. Indeed, COO is a well-known group and this unification helps the RNN to understand that the ester group is something more than just the sum of one carbonyl group and one oxygen

Table 1
Average RNN results recorded for training and test sets

Exp.	Training set						Test set				
	N	HU	MAE (K)	MAX (K)	*R*	*S* (K)	*N*	MAE (K)	MAX (K)	*R*	*S* (K)
A	137	15	8.26	42.17	0.9867	11.14	26	16.42	53.58	0.9385	21.07
1	217	17	8.36	44.88	0.9847	11.23	54	15.62	74.60	0.9177	21.18
2	217	21	8.63	40.68	0.9843	11.40	54	14.91	58.75	0.9307	19.51
3	217	21	8.34	38.10	0.9853	11.04	54	15.79	56.16	0.9235	20.45
4	217	22	8.25	36.53	0.9857	10.88	54	14.73	58.02	0.9351	18.90
5	217	23	7.96	37.72	0.9857	10.90	54	15.54	66.40	0.9241	20.37

N = number of samples; HU = number of RNN hidden units calculated as the average of the number of hidden units over 16 trials; MAE = mean absolute error; MAX = max absolute error; *R* = correlation coefficient; *S* = standard deviation. Exp. A refers to data reported for Exp. 4 in Ref. [17].

atom. On the other hand, the CH_2–CH_2 compression is a rather arbitrary unification that does not have any particular chemical meaning. It does, however, help the RNN by reducing the graph size and hence the computational effort. As expected, AAE and σ of compounds containing long aliphatic chains took most advantage from this unification. These two compressions introduced one new group each in the fragment set, but affected a large number of samples. Other fragmentation variants that differentiated amine from amide nitrogen, and mobile from non-polar hydrogen turned out to be less effective. Very likely, these variants produced a more complicated fragment set, but only a few compounds were affected thus giving rise to sampling issues.

All experiments that use cycle cutting (Exps. 2–5) gave very similar results, in terms of MAE, R and S to those obtained with the "phenyl" group representation (Exp. 17). The only clear difference is the number of HU, which slightly increased from 17 to 21–23. This resemblance was not necessarily obvious, since the two representations are quite different from each other. Indeed, the fragment set has been modified ("Phenyl" has been removed, "C aryl" and "cut1" have been added) and the molecular graphs have very different size. Moreover, the behaviour of Unique SMILES and InChI is often in contrast with the older priority rules. These rules were set up mainly on chemical basis, whereas standard systems make cutting and branching choices on purely syntactical/graphical criteria and do not account for RNN issues. For instance, they tend to produce graphs with a very long chain and short side chains, in which the distinctive part of the molecule (e.g. a functional group) is often far from the root. The RNN instead takes computational advantages in handling balanced structures, with branches of about the same length.

The results of Exps. 2–5 are similar to those of Exp. 1 not only in average values, but also in terms of individual outputs. We can compute the function $\Gamma(a,b)$, defined as:

$$\Gamma(a,b) = \sum_{i=1}^{n} \frac{\left[A_{\mathrm{out}}(i)_{\exp.a}\right] - \left[A_{\mathrm{out}}(i)_{\exp.b}\right]}{n}$$

where n is the number of samples in the data set and A_{out} is the averaged output over 16 trials. This function expresses the average output difference between two experiments. $\Gamma(2,1)$, $\Gamma(3,1)$, $\Gamma(4,1)$ and $\Gamma(5,1)$ have values of 8.1, 7.9, 9.1, and 9.5 K, respectively. These values are rather low (less than half of the S value), indicating that the individual outputs do not depend much on the representation method. This finding reveals the flexibility of the RNN technique, as the input can be given in different ways without heavily affecting the results.

Differences are small also among experiments adopting Unique SMILES (Exps. 2 and 3) and InChI (Exps. 4 and 5). Exp. 4 has a slightly better outcome, but all MAEs are contained within about 1 K and all $\Gamma(a,b)$ values ($2 \le a,b \le 5$) are between 4.6 and 8.8 K. It must be stressed, though, that the samples undergoing representation changes from one experiment to the other one are only about one fourth or less of the total data set. All other polymers were described by exactly the same graph. The learning ability of our neural network is demonstrated by the good model accuracy, as shown in Fig. 6 (see also Supplementary data tables).

In order to clarify the RNN learning, the test set predictions of Exp. 1 were analyzed to highlight the occurrence of particular trends among molecules differing by only one feature. The examined trends were: compounds having the same substituent(s) in different position(s) ("substitution" trend) of the benzene ring; compounds obtained by single group replacement, like: CH_2 → O, CH_3 → F, CH_3 → Cl ("replacement" trend); compounds differing by the length of the hydrocarbon chain between the aromatic ring and the acrylic ester group ("chain length" trend). These trends give information on the RNN behaviour in conditions of reduced variable number. It is worth noting that the trends are analyzed for analytical purposes only, as our model takes into account all structural features at once.

The RNN correctly reproduced the experimental trends in the output values of most test samples, though with some exceptions. Many of them concern the "substitution" trend, very likely because the sampling was inadequate to properly train the RNN. The T_g experimental order usually is *para* > *ortho* > *meta*, but often becomes *ortho* > *para* > *meta* or *ortho* > *meta* > *para*. The computed T_g (333 K) of poly(3-chlorophenyl acrylate) is larger than 330 K, the experimental T_g of the *para* compound. It, however, follows the correct order with respect to the outputs of the *ortho* and *para* compounds in the training set, which are 336 and 339 K, respectively. Poly(biphenyl-4-yl acrylate) exhibits the same behaviour. The poly(methyl acryloyloxybenzoate) series is well reproduced (*para* > *ortho* > *meta*), whereas a very similar one, poly(ethyl acryloyloxybenzoate), is fitted in the training set with a wrong order (*para* > *meta* > *ortho*). The RNN

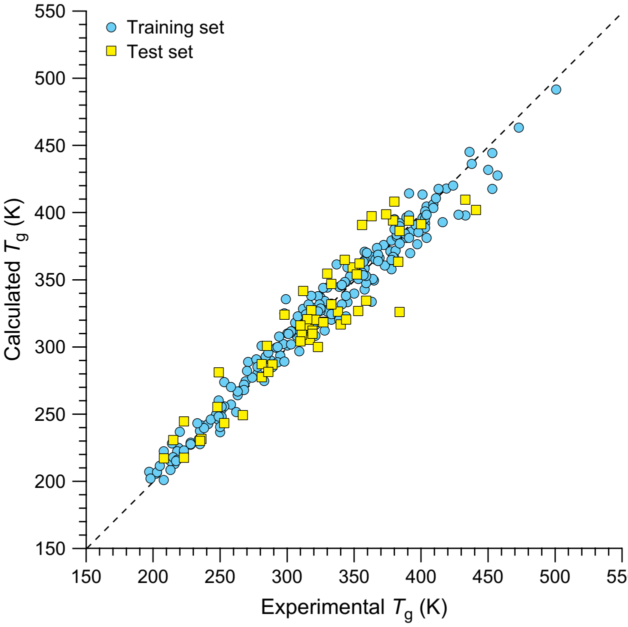


Fig. 6. Plot of computed vs. experimental T_gs of (meth)acrylic polymers (Exp. 4).

assigned the anticipated *para* > *ortho* > *meta* trend to the series of poly(methacryloyloxybenzoic acid), whereas the experimental trend was *ortho* > *meta* > *para*. This behaviour can be attributed tentatively to the presence of a carboxylic acid group that is scarcely sampled in the data set.

The "replacement" trend was correctly predicted for poly(3-methylphenyl methylacrylate) and poly(3-chlorophenyl methylacrylate) but not for poly(2,4-dimethylphenyl methacrylate) and poly(2,4-dichlorophenyl methacrylate).

In most cases, the experimental T_g dependence on spacer length was correctly reproduced. On the other hand, poly(4-(4-(4-nitrophenyldiazenyl)phenoxy)butyl acrylate) and poly(4-(phenyldiazenyl)phenyl methacrylate) are given a wrong order in their series. Poly(benzyl 4-(methacryloyloxy)benzoate), poly(6-(4-((4-(dimethylamino)phenyl)diazenyl)phenoxy)hexyl acrylate), poly(2-(4-methoxybiphenyl-4′-oxy)ethyl acrylate) and poly(2-(2-(2-(((4-cyanophenyl)diazenyl)phenoxy)ethoxy)-ethoxy)ethyl methacrylate) follow the expected trend only with respect to the training output values of the other compounds of their respective series.

Exps. 2–5 reproduced most of the above trends, including the *ortho–meta–para* trend, although the information of the substituent position is carried in a very different way (edge order in the case of *group* representation, position on the chain in the case of *cycle cutting*). This behaviour is another element that supports the RNN flexibility. In a few cases, however, there are some differences. The output values of the poly(chlorophenyl acrylate) series are in the wrong order. On the other hand, poly(2,4-dimethylphenyl methacrylate) and poly(2,4-dichlorophenyl methacrylate), which were in the wrong order in the first experiment, are now correctly reproduced. The output of poly(8-((4′-((*S*)-2-methylbutoxy)biphenyl-4-yl)oxy)octyl acrylate), whose target value can be in a class of its own because of its low molecular weight, fits into the experimental trend in exps 3 and 4, but not in exps 2 and 5.

It must be stressed, however, that the target values among the members of a series often differ by only 10 K or less, a value that is much smaller than the training error tolerance.

As already observed, the RNN results are not particularly affected by the choice of the standard representation system and very few samples show a clearly different output as a consequence of the different descriptions given by InChI and Unique SMILES. However, two cases are particularly interesting. The first one is the case of poly(3-methoxyphenyl methacrylate) and poly(4-methoxyphenyl methacrylate): the order of their T_g (*para* > *meta*) is respected in all experiments. Their output values are very close in Exps. 2 and 3 (which use Unique SMILES), whereas in Exps. 4 and 5 (which make use of InChI) they are separated by about 30 K, which is more correct. Unique SMILES represented these two compounds in the same way, with the methoxy group prior to the rest of the phenyl ring, whatever its position. Instead, InChI set a higher priority for the methoxy group when it was in 3-position and a lower one when it was in 4-position.

The second case involves poly(4-cyanophenyl acrylate) and poly(4-cyanobenzyl acrylate). Unique SMILES gives priority to the cyano group in both polymers, while InChI always gives it to the aromatic ring. As a result, the output values of the experiments using InChI are 20–30 K lower than the ones using Unique SMILES. On the other hand, their AAEs are about the same, since one method overestimates and the other one underestimates the output values by more or less the same amount.

6. Conclusions

This work constitutes a further advance in the treatment of polymers through structure-based predictive methods. In particular, we have exploited different possibilities in the representation of cycles through hierarchical structures, describing them either in a single group form or with methods derived from standard formats. In all cases we have shown that the inclusion of cyclic moieties in a RNN–QSPR study is feasible and does not affect the predictive accuracy. The results of the experiments reported in this paper are indeed comparable to those obtained on a data set containing only acyclic compounds [17]. The mean average errors and the standard deviations are comparable also among experiments that use different representation types. This confirms the robustness of the method with respect to the introduction of different typologies of data. Moreover, the reported results are also quite satisfactory in comparison with those of different literature methods.

The RNN flexibility allows for choosing a molecular representation by finding a balance between structural detail and sampling in each investigated data set. The two representation methods adopted in the present investigation, *group* and *cycle breaking*, have indeed different sampling requirements. The first one is best suited for more homogeneous and specialized data sets, whereas *cycle breaking* can treat a larger variety of structures. The reported experiments show that both representations are effective; the designer is left the freedom to choose between the simplicity and computational advantage of *group* representation and the generality of the other technique.

The good results afforded by the *cycle breaking* representation open the way to the investigation of data sets that contain a greater variety of cyclic moieties with poor sampling, thus fully exploiting the potential of this description method.

Acknowledgments

This research was performed with the financial support of the Italian Ministry of University and Research (MIUR).

Appendix. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.polymer.2007.09.043.

References

[1] Van Krevelen DW. Properties of polymers. 2nd ed. New York: Elsevier; 1976.
[2] Hopfinger AJ, Koehler MG, Pearlstein RA. J Polym Sci Part B Polym Phys 1988;26:2007–28.

[3] Koehler MG, Hopfinger AJ. Polymer 1989;30:116–26.
[4] Bicerano J. Prediction of polymer properties. 3rd ed. New York: Marcel Dekker; 2002. Revised and expanded.
[5] Camelio P, Cypcar CC, Lazzeri V, Waegel B. J Polym Sci Part A Polym Chem 1997;35:2579–90.
[6] Cypcar CC, Camelio P, Lazzeri V, Mathias LJ, Waegel B. Macromolecules 1996;29:8954–9.
[7] Gao H, Harmon JP. J Appl Polym Sci 1997;64(3):507–17.
[8] Katritzky AR, Sild S, Lobanov VS, Karelson M. J Chem Inf Comput Sci 1998;38:300–4.
[9] Garcia-Domenech R, de Juliàn-Ortiz JV. J Phys Chem B 2002;106:1501–7.
[10] Joyce SJ, Osguthorpe DJ, Padgett JA, Price GJ. J Chem Soc Faraday Trans 1995;91:2491–6.
[11] Sumpter BG, Noid DW. J Therm Anal 1996;46:833–51.
[12] Ulmer II CW, Smith DA, Sumpter BG, Noid DI. Comput Theor Polym Sci 1998;8:311–21.
[13] Askadskii AA. Polym Sci USSR 1966;9:471–87.
[14] Askadskii AA, Slonimskii GL. Polym Sci USSR 1971;13:2158–60.
[15] Porter D. Group interaction modeling of polymer properties. New York: Dekker; 1995.
[16] Mattioni BE, Jurs PC. J Chem Inf Comput Sci 2002;42:232–40.
[17] Duce C, Micheli A, Solaro R, Starita A, Tiné MR. J Mat Chem, in press.
[18] Micheli A, Sperduti A, Starita A, Bianucci AM. J Chem Inf Comput Sci 2001;41:202–18.
[19] Bianucci AM, Micheli A, Sperduti A, Starita A. In: Sztandera LM, Cartwright HM, editors. Soft computing approaches in chemistry. Heidelberg: Springer-Verlag; 2003. p. 265–96.
[20] Bianucci AM, Micheli A, Sperduti A, Starita A. Appl Intell J 2000; 12:117–46.
[21] Micheli A. Ph.D. thesis. TD-13/03, Department of Computer Science, University of Pisa; 2003.
[22] Bernazzani L, Duce C, Micheli A, Mollica V, Sperduti A, Starita A, et al. TR-04-16. Department of Computer Science, University of Pisa; 2004.
[23] Duce C. Ph.D. thesis. Department of Chemistry and Industrial Chemistry, University of Pisa; 2005.
[24] Bernazzani L, Duce C, Micheli A, Mollica V, Sperduti A, Starita A, et al. J Chem Inf Model 2006;46:2030–42.
[25] Duce C, Micheli A, Starita A, Tiné MR, Solaro R. Macromol Rapid Commun 2006;27:712–6.
[26] Sperduti A, Starita A. IEEE Trans Neural Networks 1997;8:714–35.
[27] Duce C, Micheli A, Solaro R, Starita A, Tiné MR. Recursive neural networks for quantitative structure–property relationship analysis of polymers. In: Simos T, Maroulis G, editors. Lecture series on computer and computational sciences, vol. 4. Leiden: Brill Academic Publishers; 2005. p. 1546–9.
[28] Micheli A, Sperduti A, Starita A, Bianucci AM. Proc Int Joint Conf Neural Networks 2001;4:2732–7.
[29] Weininger D. J Chem Inf Comput Sci 1988;28:31–6.
[30] <http://www.daylight.com/smiles/index.html>.
[31] McKay BD. Congressus Numerantium 1981;30:45–87, <http://cs.anu.edu.au/~bdm/papers/pgi.pdf>.
[32] McNaught A. Chem Int 2006;28(6):12–4.
[33] Stein SE, Heller SR, Tchekhovskoi DV. Nimes Int Chem Inf Conf Proc 2003:131–143, see <http://www.iupac.org/inchi/>.
[34] Morgan HL. J Chem Doc 1965;5:107–13.
[35] Weininger D, Weininger A, Weininger JL. J Chem Inf Comput Sci 1989; 29:97–101.
[36] Krause S, Gormley JJ, Roman N, Shetter JA, Watanabe WH. J Polym Sci Part A Polym Chem 1965;3:3573–86.
[37] Diaz-Calleja R, Riande E, San Romàn J. Macromolecules 1991;24:1854–8.
[38] Pizzirani G, Magagnini PL. Chim Ind (Milan) 1968;50:1218–21.
[39] Baccaredda M, Magagnini PL, Pizzirani G, Giusti P. J Polym Sci Part B Polym Phys 1971;9:303–10.
[40] Butler K, Thomas PR, Tyler GJ. J Polym Sci 1960;48:357–66.
[41] Petrovich-Djakov DM, Filipovich JM, Vrhovac LP, Velickovic JS. J Therm Anal 1993;40:741–6.
[42] Pilcher SC, Ford WT. J Polym Sci Part A Polym Chem 2001;39: 519–24.
[43] Ekstrin FA, Gurevich KL, Marinin VG, Kalinin AI, Kuksenkova NS. Trudy Khim Khim Tekhnol 1970;2:172–5.
[44] Velickovic JS, Filipovic JM, Plavsic MB, Petrovic-Djakov DM, Petrovic ZS, Budinski JK. Polym Bull 1991;27:331–6.
[45] Kihira Y, Sugiyama K. J Macromol Sci Phys 1987;B26(2):227–36.
[46] Alberda Van Ekenstein GOR, Altena HJH, Tan YY. Eur Polym J 1989;25(2):111–5.
[47] Chetyrkina GM, Sokolova TA, Koton MM. Vysokomol Soedin Vsesoyuz Khim Obsch Mendeleeva 1959;1:248–53.
[48] Yuki H, Hatada K, Nijnomi T, Hashimoto M, Oshima J. Polym J 1971;2:629–39.
[49] Vasquez B, Gurruchaga M, Goni I, Narvarte E, San Romàn J. Polymer 1995;36(18):3467–72.
[50] Rottink JBH, Te Nijenhuis K, Addink R, Mijs WJ. Polym Bull 1993; 31:221–8.
[51] Li M, Zhou E, Xu J, Yang C, Tang X. Polym Bull 1995;35:65–72.
[52] Shibaev VP, Kostromin SG, Plate NA. Eur Polym J 1982;18:651–9.
[53] Yilmaz F, Kasapoglu F, Hepuzer Y, Yagci Y, Toppare L, Grillo Fernandes E, et al. Des Monomers Polym 2005;8:223–36.
[54] Pugh C, Percec V. ACS Polym Prep 1986;27(1):366–8.
[55] Bai S, Zhao Y. Macromolecules 2002;35:9657–64.
[56] Haitjema HJ, Buruma R, Alberda van Ekenstein GOR, Tan YY, Challa G. Eur Polym J 1996;32(12):1447–55.
[57] Diaz-Calleja R, Riande E, San Romàn J. J Phys Chem 1992;96:6843–8.
[58] Cristofolini L, Berzina T, Fontana MP, Konovalov O. Mol Cryst Liq Cryst Sci Tech Sec A Mol Cryst Liq Cryst 2002;375:689–99.
[59] Chiellini E, Galli G, Cioni F, Dossi E, Gallot B. J Mater Chem 1993;3(10):1065–73.
[60] Chien LC, Boyden MN, Waltz AJ, Shenouda IG, Citano CM. Mol Cryst Liq Cryst 1998;317:273–85.
[61] Licea-Claverie A, Rogel-Hernandez E, Salgado-Rodriguez R, Lopez-Sanchez JA, Castillo LA, Cornejo-Bravo JM, et al. Macromol Symp 2004;207:193–215.
[62] Finkelmann H, Koldehoff J, Ringsdorf H. Angew Chem Int Ed Engl 1978;17(12):935–6.
[63] Koehler W, Robello DR, Willand CS, Williams DJ. Macromolecules 1991;24:4589–99.
[64] Altomare A, Ciardelli F, Mellini L, Solaro R. Macromol Chem Phys 2004;205:1611–9.
[65] Altomare A, Ciardelli F, Ghiloni MS, Solaro R. Gazz Chim Ital 1997;127:143–9.
[66] Altomare A, Ciardelli F, Lima R, Solaro R. Chirality 1991;3:292–8.
[67] Altomare A, Andruzzi L, Ciardelli F, Mader M, Tirelli N, Solaro R. Macromol Symp 1999;137:33–46.
[68] Altomare A, Ciardelli F, Faralli G, Solaro R. Macromol Mater Eng 2003;288:679–92.
[69] Soykan C, Erol I. J Polym Res 2004;11:53–63.
[70] Nanjundan S, Sreekuttan Unnithan C, Jone Selvamalar CS, Penlidis A. React Funct Polym 2005;62:11–24.
[71] Han YK, Dufour B, Wu W, Kowalewski T, Matyjaszewski K. Macromolecules 2004;37:9355–65.
[72] Jone Selvamalar CS, Krithiga T, Penlidis A, Nanjundan S. React Funct Polym 2003;56:89–101.
[73] Matsuzaki K, Kanai T, Yamawaki K, Samre Rung KP. Makromol Chem 1973;174:215–23.
[74] Sin SL, Gan LH, Hu X, Tam KC, Gan YY. Macromolecules 2005; 38:3943–8.
[75] Niezette J, Desreux V. Makromol Chem 1971;149:177–83.