Prediction of high weight polymers glass transition temperature using RBF neural networks

Antreas Afantitis, Georgia Melagraki, Kalliopi Makridima, Alex Alexandridis, Haralambos Sarimveis*, Olga Iglessi-Markopoulou

School of Chemical Engineering, National Technical University of Athens, 9, Heroon Polytechniou Str., Zografou Campus, Athens 15780, Greece

Received 9 September 2004; accepted 4 November 2004
Available online 5 January 2005

Abstract

A novel approach to the prediction of the glass transition temperature (T_g) for high molecular polymers is presented. A new quantitative structure–property relationship (QSPR) model is obtained using Radial Basis Function (RBF) neural networks and a set of four-parameter descriptors, ∑MV_α_β_γ_δ(R_n), L_f, ΔX_SB and ∑PEI. The produced QSPR model (R^2=0.9269) proved to be considerably more accurate compared to a multiple linear regression model (R^2=0.8227).

Keywords: RBF neural network; QSPR; Glass transition temperature

1. Introduction

Determination of the physical properties of organic compounds based on their structure is a major research subject in computational chemistry. Quantitative structure–property relationship (QSPR) correlations have been widely applied for the prediction of such properties over the last decades [1–3]. A breakthrough has occurred in this field with the appearance of artificial neural networks (ANNs).

The glass transition is the most important transition and relaxation that occurs in amorphous polymers. It has a significant effect on the properties and processing characteristics of this type of polymers [4]. The glass transition (T_g) is difficult to be determined because the transition happens over a comparatively wide temperature range and depends on the method, the duration and the pressure of the measuring device [5,6]. Besides these difficulties, the experiments are costly and time consuming.

In the past, numerous attempts have been made to predict T_g for polymers by different approaches. According to Katrinzky et al. [7] there are two kinds of approaches, the empirical and the theoretical. Empirical methods correlate the target property with other physical or chemical properties of the polymers, for example, group additive properties (GAP) [8]. The most widely referenced model of the theoretical estimations produced by Bicerano [6] combines a weighted sum of structural parameters along with the solubility parameter of each polymer. In his work, a regression model was produced for 320 polymers but no external data set compounds were used to validate this model.

Cameilio et al. [9] calculated the parameters of 50 acrylates and methacrylates with molecular mechanics and correlated them with T_g. Katrizky et al. [10] introduced a model for 22 medium molecular weight polymers using four parameters. Following this work, Katrinzky et al. [7] and Cao and Lin [11] obtained two separate models for 88 un-cross-linked homopolymers including polyethylenes, polyacrylates, polymethacrylates, polystyrenes, polyethers, and polyoxides. The models were used as predictors of the molar glass transition temperatures [7] (T_g/M) and glass transition temperatures [11]. Joyce et al. [12] used neural networks for the prediction of T_g based on monomer structure of polymers. Another approach with neural network was proposed by Sumpter and Noid [13] using...
the repeating unit structure as representative of the polymer. Finally Juris and Mattioni \[14\] obtained a QSAR model which predicts \(T_g\) values for a diverse set of polymers.

An ANN-based modeling method could produce a more accurate QSAR model compared to linear methods, since it has the ability to approximate the possible non-linear relationships between structural information and properties of compounds during the training process. The resulting model can generalize the knowledge among homologous series without need for theoretical formulas \[6\]. In this work we explore these neural network capabilities, by introducing a new QSAR model for the prediction of \(T_g\) values that is based on the RBF architecture. The database consists of 88 un-cross-linked homopolymers and contains the experimental values of \(T_g\) and the values of the following descriptors: \(L_{P}\), \(\Delta X_{SB}\) and \(\sum\) PEI. All the data are taken from Cao and Lin \[11\].

**2. Modeling methodology**

In this section we present the basic characteristics of the RBF neural network architecture and the training method that was used to develop the QSAR neural network models.

**2.1. RBF network topology and node characteristics**

RBF networks consist of three layers: the input layer, the hidden layer and the output layer. The input layer collects the input information and formulates the input vector \(\hat{x}\). The hidden layer consists of \(L\) hidden nodes, which apply non-linear transformations to the input vector. The output layer delivers the neural network responses to the environment. A typical hidden node \(l\) in an RBF network is described by a vector \(\hat{x}_l\), equal in dimension to the input vector and a scalar width \(\sigma_l\). The activity \(v_l(x)\) of the node is calculated as the Euclidean norm of the difference between the input vector and the node center and is given by:

\[
v_l(x) = \|x - \hat{x}_l\| \tag{1}
\]

The response of the hidden node is determined by passing the activity through the radially symmetric Gaussian function:

\[
f_l(x) = \exp\left(-\frac{v_l(x)^2}{\sigma_l^2}\right) \tag{2}
\]

Finally, the output values of the network are computed as linear combinations of the hidden layer responses:

\[
\hat{y}_m = g_m(x) = \sum_{l=1}^{L} f_l(x)w_{lm}, \quad m = 1, \ldots, M \tag{3}
\]

where \(\{w_{1, m}, w_{2, m}, \ldots, w_{L, m}\}\) is the vector of weights, which multiply the hidden node responses in order to calculate the \(m\)th output of the network.

**2.2. RBF network training methodology**

Training methodologies for the RBF network architecture are based on a set of input–output training pairs \((x(k); y(k))\) \((k=1,2,\ldots,K)\). The training procedure used in this work consists of three distinct phases:

(i) Selection of the network structure and calculation of the hidden node centers using the fuzzy means clustering algorithm \[15\]. The algorithm is based on a fuzzy partition of the input space, which is produced by defining a number of triangular fuzzy sets on the domain of each input variable. The centers of these fuzzy sets produce a multidimensional grid on the input space. A rigorous selection algorithm chooses the most appropriate knots of the grid, which are used as hidden node centers in the produced RBF network model. The idea behind the selection algorithm is to place the centers in the multidimensional input space, so that there is a minimum distance between the center locations. At the same time the algorithm assures that for any input example in the training set, there is at least one selected hidden node that is close enough according to a distance criterion. It must be emphasized that opposed to both the \(k\)-means \[16\] and the \(c\)-means clustering \[17\] algorithms, the fuzzy means technique does not need the number of clusters to be fixed before the execution of the method. Moreover, due to the fact that it is a one-pass algorithm, it is extremely fast even if a large database of input–output examples is available.

(ii) Following the determination of the hidden node centers, the widths of the Gaussian activation function are calculated using the \(p\)-nearest neighbour heuristic \[18\]

\[
\sigma_l = \left(\frac{1}{p} \sum_{i=1}^{p} \|\hat{x}_l - \hat{x}_i\|^2\right)^{1/2} \tag{4}
\]

where \(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_p\) are the \(p\) nearest node centers to the hidden node \(l\). The parameter \(p\) is selected, so that many nodes are activated when an input vector is presented to the neural network model.

(iii) The connection weights are determined using linear regression between the hidden layer responses and the corresponding output training set.

**3. Results and discussion**

The data set of 88 polymers was divided into a training set of 44 polymers, and a validation set of 40 polymers, while 4 polymers were rejected as outliers. The selection of the compounds in the training set was made according to the structure of the polymers, so that representatives of a wide range of structures (in terms of the different branching
and length of the carbon chain) were included. The polymers in the training set and validation sets along with the collected from the literature [11] experimental glass transition temperatures are presented in Tables 1 and 2, respectively.

Structural parameters for the 84 polymers were calculated by the equations provided in the literature [11]. Two sets of descriptors were formulated. The first one (set 1) includes four parameters \( \sum MV(\text{ter}) (R_{ij})\), \( L_{F} \), \( \Delta X_{SB} \) and \( \sum PEI \), while the second one (set 2) incorporates only three parameters \( \sum MV(\text{ter}) (R_{ij})\), \( \sum PEI \) and \( \Delta X_{SB} \). \( \Delta X_{SB} \) is related to the polarity of the repeating unit, while dipole of the side group depends on \( \sum PEI \) [11]. These two parameters express the intermolecular forces of the polymers. \( \sum MV(\text{ter}) (R_{ij})\) expresses the no free rotation part of the side chain and \( L_{F} \) (free length) expresses the bond count of the free rotation part of side chain [11]. The four descriptors are very attractive because they can be calculated easily, rapidly and they have clear physical meanings.

The RBF training method described in Section 2 was implemented using the Matlab computing language in order to produce the ANN models. It should be emphasized that the method has been developed in-house, so no commercial packages were utilized to build the neural network models. For comparison purposes, a standard multivariate regression

### Table 1

<table>
<thead>
<tr>
<th>A/A</th>
<th>Name</th>
<th>( T_{g(K)\text{exp}} )</th>
<th>( T_{g(K)\text{train (set 1 ANN)}} )</th>
<th>( T_{g(K)\text{train (set 2 ANN)}} )</th>
<th>( T_{g(K)\text{train (set 1 linear)}} )</th>
<th>( T_{g(K)\text{train (set 2 linear)}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Poly(ethylene)</td>
<td>195</td>
<td>198.5551</td>
<td>198.5575</td>
<td>206.2141</td>
<td>180.7988</td>
</tr>
<tr>
<td>2</td>
<td>Poly(1-pentene)</td>
<td>220</td>
<td>218.7587</td>
<td>221.2788</td>
<td>235.0911</td>
<td>232.7334</td>
</tr>
<tr>
<td>3</td>
<td>Poly(5-methyl-1-hexene)</td>
<td>268</td>
<td>268.3445</td>
<td>263.7424</td>
<td>266.4187</td>
<td>302.5932</td>
</tr>
<tr>
<td>4</td>
<td>Poly(3,3-dimethylbutyl methacrylate)</td>
<td>279</td>
<td>278.0939</td>
<td>273.3399</td>
<td>250.3024</td>
<td>247.1930</td>
</tr>
<tr>
<td>5</td>
<td>Poly(1,1-dichloroethylene)</td>
<td>302</td>
<td>291.4458</td>
<td>281.6227</td>
<td>295.7158</td>
<td>280.9432</td>
</tr>
<tr>
<td>6</td>
<td>Poly(2-methoxyethyl methacrylate)</td>
<td>306</td>
<td>304.5211</td>
<td>301.0940</td>
<td>307.6720</td>
<td>308.3655</td>
</tr>
<tr>
<td>7</td>
<td>Poly(2-chloroacrylate)</td>
<td>315</td>
<td>319.0651</td>
<td>317.5529</td>
<td>359.6010</td>
<td>365.0133</td>
</tr>
<tr>
<td>8</td>
<td>Poly(2-propyl methacrylate)</td>
<td>320</td>
<td>320.1201</td>
<td>319.8753</td>
<td>358.7604</td>
<td>364.9096</td>
</tr>
</tbody>
</table>

For comparison purposes, a standard multivariate regression
Table 2
Validation set

<table>
<thead>
<tr>
<th>A/A</th>
<th>Name</th>
<th>(T_{g(K),exp}[7])</th>
<th>(T_{g(K),pred\ (set\ 1\ ANN)} R^2)</th>
<th>(T_{g(K),pred\ (set\ 2\ ANN)} R^2)</th>
<th>(T_{g(K),pred\ (set\ 1\ linear)} R^2)</th>
<th>(T_{g(K),pred\ (set\ 2\ linear)} R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Poly(ethylacrylate)</td>
<td>228</td>
<td>225.7773</td>
<td>206.1942</td>
<td>254.3056</td>
<td>232.2911</td>
</tr>
<tr>
<td>2</td>
<td>Poly(cyclopentylethylene)</td>
<td>348</td>
<td>358.7344</td>
<td>343.5276</td>
<td>333.7406</td>
<td>312.7605</td>
</tr>
<tr>
<td>3</td>
<td>Poly(acrylic acid)</td>
<td>379</td>
<td>370.7699</td>
<td>383.7025</td>
<td>329.0515</td>
<td>303.8972</td>
</tr>
<tr>
<td>4</td>
<td>Poly(ethyl acrylate)</td>
<td>251</td>
<td>260.9209</td>
<td>246.7095</td>
<td>258.6331</td>
<td>259.2738</td>
</tr>
<tr>
<td>5</td>
<td>Poly(acrylonitride)</td>
<td>378</td>
<td>345.0173</td>
<td>371.8758</td>
<td>313.8227</td>
<td>286.6382</td>
</tr>
<tr>
<td>6</td>
<td>Poly(styrene)</td>
<td>373</td>
<td>371.7688</td>
<td>347.9344</td>
<td>346.6853</td>
<td>326.8437</td>
</tr>
<tr>
<td>7</td>
<td>Poly(3-chiroloystrene)</td>
<td>363</td>
<td>384.5075</td>
<td>389.0822</td>
<td>368.3181</td>
<td>351.7191</td>
</tr>
<tr>
<td>8</td>
<td>Poly(4-methylstyrene)</td>
<td>374</td>
<td>374.1514</td>
<td>372.7100</td>
<td>361.5876</td>
<td>344.9300</td>
</tr>
<tr>
<td>9</td>
<td>Poly(propylene)</td>
<td>233</td>
<td>226.4469</td>
<td>187.9298</td>
<td>262.2846</td>
<td>231.5684</td>
</tr>
<tr>
<td>10</td>
<td>Poly(ethoxyethylene)</td>
<td>254</td>
<td>225.3849</td>
<td>228.6502</td>
<td>252.0064</td>
<td>247.9495</td>
</tr>
</tbody>
</table>

Table 3
Summary of the results produced by the different methods

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Method</th>
<th>Training set</th>
<th>Validation set</th>
<th>(R^2_{train})</th>
<th>(R^2_{pred})</th>
<th>Figure</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Set 1</td>
<td>Neural network</td>
<td>44</td>
<td>40</td>
<td>0.9968</td>
<td>0.9269</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Set 2</td>
<td>Neural network</td>
<td>293</td>
<td>290.0164</td>
<td>285.9807</td>
<td>289.8661</td>
<td>305.7211</td>
</tr>
<tr>
<td>3</td>
<td>Set 1</td>
<td>Linear</td>
<td>300</td>
<td>271.9207</td>
<td>316.5163</td>
<td>345.6684</td>
<td>327.9476</td>
</tr>
<tr>
<td>4</td>
<td>Set 2</td>
<td>Linear</td>
<td>304</td>
<td>298.8250</td>
<td>345.7275</td>
<td>265.9810</td>
<td>272.7775</td>
</tr>
<tr>
<td>5</td>
<td>Set 1</td>
<td>Cross-validation, neural network</td>
<td>333</td>
<td>337.5040</td>
<td>338.5281</td>
<td>321.8972</td>
<td>312.9290</td>
</tr>
<tr>
<td>6</td>
<td>Set 2</td>
<td>Cross-validation, neural network</td>
<td>348</td>
<td>321.6752</td>
<td>251.2710</td>
<td>244.1416</td>
<td>300.7819</td>
</tr>
<tr>
<td>7</td>
<td>Set 1</td>
<td>Cross-validation, linear</td>
<td>378</td>
<td>372.8332</td>
<td>369.3446</td>
<td>377.9002</td>
<td>366.2196</td>
</tr>
<tr>
<td>8</td>
<td>Set 2</td>
<td>Cross-validation, linear</td>
<td>378</td>
<td>372.8332</td>
<td>369.3446</td>
<td>377.9002</td>
<td>366.2196</td>
</tr>
</tbody>
</table>
method for producing linear models was also utilized. Both neural networks and linear models were trained using the 44 individuals in the training set and were tested on the independent validation set consisting of 40 examples. The models produced by multiple linear regression on the two sets of descriptors are shown next:

$$T_g (K) = 0.3617 \sum \text{MV}_{\text{ter}}(R_{\text{ter}}) - 10.3254L_F$$

$$+ 159.7984 \Delta X_{SB} + 9.3931 \Sigma \text{PEI} + 206.2141$$

(5)

$$T_g (K) = 0.4394 \sum \text{MV}_{\text{ter}}(R_{\text{ter}}) + 167.2681 \Delta X_{SB}$$

$$+ 2.8929 \Sigma \text{PEI} + 180.7988$$

(6)

The RBF models generated using the two sets of descriptors consisted of 34 and 25 hidden nodes, respectively. RBF models are more complex compared to the linear models and are not shown in the paper for brevity, but can be available to the interested reader. The produced ANN QSPR models for the prediction of glass transition temperature, proved to be more accurate compared to multiple linear regression models using both sets of descriptors as shown in Table 3, where the results are summarized. More detailed results can be found in Tables 1 and 2 where the estimations of the two modeling techniques for the training examples and the predictions for the validation examples are depicted in an example-to-example basis. There are four columns of results in the two tables corresponding to the two modeling methodologies and the two sets of descriptors. Figs. 1–4 show the experimental glass transition temperatures vs. the predictions produced by the neural network and the multiple regression techniques in a graphical representation format.

To further explore the reliability of the proposed method we also used the leave-one-out cross-validation method on the full set of the available data (excluding the outliers). The results are summarized in Table 3 and are shown in Figs. 5 and 6, where again the superiority of the neural network methodology over the multiple linear regression method is clear. It should be mentioned, that contrary to the aforementioned results, there is a decrease in the $R^2$ statistic in both modeling methodologies when the three-descriptor set is utilized. However, the $R^2$ statistic for the neural network methodology using the second set of descriptors is still high, meaning that the respective neural network model is reliable.

Summarizing the results presented in this work we can make the following observations:

(i) The modeling procedures utilized in this work (separation of the data into two independent sets and leave-one-out cross-validation) illustrated the accuracy of the produced models not only by calculating their fitness on sets of training data, but also by testing the predicting abilities of the models.

(ii) We showed that using the neural network methodology we can still have a reliable prediction, when the descriptor $L_F$ is dropped. Therefore, a three-descriptor ANN model can be used for the prediction of the glass transition temperature at
the expense of the increased complexity of the model compared to the simple structure of a linear model.

4. Conclusions

The results of this study show that a practical model can be constructed based on the RBF neural network architecture for a set of 84 high molecular weight polymers. The most accurate models were generated using four descriptors and resulted in the following statistics: $R^2_{\text{set 1}} = 0.9968$ for the training data, $R^2_{\text{set 1}} = 0.9269$ for the validation data and $R^2_{\text{set 1, CV}} = 0.9269$ for the cross-validation method. We showed that using the neural network approach, we can further reduce the number of descriptors from four to three and still produce a reliable model. The neural network models are produced based on the special fuzzy means training method for RBF networks that exhibits small computational times and excellent prediction accuracies. The proposed method could be a substitute to the costly and time-consuming experiments for determining glass transition temperatures or to the approximate empirical equations with limited reliability.

Acknowledgments

A. Af. wishes to thank the A.G. Leventis Foundation for its financial support.

References