Inter*net* Electronic Journal of Molecular Design

February 2002, Volume 1, Number 2, Pages 80–93

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Alexandru T. Balaban on the occasion of the 70th birthday
Part 2

Guest Editor: Mircea V. Diudea

Neural Network Modeling of Melting Temperatures for Sulfur–Containing Organic Compounds

Julian Kozioł

Department of Physical Chemistry, Rzeszów University of Technology, Powstańców Warszawy Ave. 6, P.O. Box 85, 35–041 Rzeszow, Poland

Received: December 5, 2001; Revised: January 31, 2002; Accepted: February 15, 2002; Published: February 28, 2002

Citation of the article:
J. Kozioł, Neural Network Modeling of Melting Temperatures for Sulfur–Containing Organic Compounds, *Internet Electron. J. Mol. Des.* **2002**, *1*, 80–93, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Neural Network Modeling of Melting Temperatures for Sulfur–Containing Organic Compounds[#]

Julian Kozioł*

Department of Physical Chemistry, Rzeszów University of Technology, Powstańców Warszawy Ave. 6, P.O. Box 85, 35–041 Rzeszow, Poland

Received: December 5, 2001; Revised: January 31, 2002; Accepted: February 15, 2002; Published: February 28, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (2), 80–93

Abstract

Motivation. Searching for a comprehensive numerical description of the chemical structure and for methods that enable to develop effective and credible QSPR (quantitative structure–property relationships) models represent significant challenges for the contemporary theoretical chemistry. Among these methods artificial neural networks (ANN) appears to be very promising in obtaining models that convert structural features into different properties of chemical compounds.

Method. Two different models relating structural descriptors to melting temperatures of sulfur containing organic compounds are developed using ANN. A new set of molecular descriptors is evaluated to determine their suitability for QSPR studies. Using two data sets containing 150 sulfides and 226 sulfones, ANN trained with the back propagation and conjugated gradient algorithms are able to predict the melting temperatures with good accuracy.

Results. The results obtained show a good predictive ability for the ANN models, giving R^2_{cv} equal to 0.880 and 0.794 for the sulfides and sulfones, respectively.

Conclusions. The QSPR studies described in this paper provide strong evidence that the tested structural descriptors are useful and effective for the ANN modeling of the melting temperatures of sulfides and sulfones.

Keywords. QSPR; molecular descriptors; artificial neural networks; melting temperature; sulfide; sulfone.

Abbreviations and notations

ANN, artificial neural network	QSPR, quantitative structure–property relationships
IPS, intelligent problem solver	SA, sensitivity analysis
t_m, melting temperature	SNN, Statistica Neural Networks
PER, prediction error	

1 INTRODUCTION

The numerical description of the molecular structure using structural invariants with the aim of modeling the physicochemical properties of chemical compounds has gained increased importance over last decade, became one of most explored areas of research in computational chemistry [1].

[#] Dedicated on the occasion of the 70th birthday to Professor Alexandru T. Balaban.

* Correspondence author; phone: +48-17-865-1822; fax: +48-17-854-9830; E-mail: koziol@prz.rzeszow.pl.

Special interest in obtaining reliable models able to estimate different properties of known structures of not yet synthesized chemical molecules is connected with wide application of combinatorial chemistry tools. Also experimental determination of properties of newly synthesized chemical compounds may be sometimes not cost effective or even impossible due to a lack or instability of available material.

Recently, besides traditional methods of computing the properties of chemical compounds, various statistical methods such as multiply linear regression, cluster analysis, principal component analysis and partial least–squares regression have been applied to the QSPR studies [2–4]. For the prediction of physical properties, high–quality models, usually based on predictive equations obtained using linear regression techniques, were used to correlate structural parameters with observed properties [5–7]. Currently, neural networks, representing general nonlinear methods, were used with encouraging success in development of various QSPR models [8–25]. Artificial neural networks (ANN) are well–suited to describe structure–property models. Moreover, ANN is able to consider not only particular structure characteristics, but also interrelations and interdependences between mutually influencing structural features. Therefore, they can be easily adapted for processing larger vectors of structural data formed by a set of descriptors.

A set of indices converting structural features into a multicomponent vector of numerical values, scaled in the range of 0.1 to 0.9, was proposed recently [25]. The basic assumption of this coding scheme is the treatment of each molecule as a linear structure with linear, branched, or cyclic substituents. The described method is useful for estimating the boiling temperatures of hydrocarbons, nitrogen and oxygen containing compounds [22,25] and melting temperatures of amides [25]. However, it was not applied to compounds with other types of heteroatoms. The work presented here extends this model for sulfur containing compounds. In this study two sets of sulfides and sulfones have been investigated, obtaining ANN models with good predictive performance in modeling their melting temperatures.

2 MATERIALS AND METHODS

2.1 Chemical Data

Chemical structures and melting temperatures of 150 sulfides and 226 sulfones were selected from [26]. The data sets contain different types of structures: aliphatic (linear and branched), cyclic and/or aromatic. The melting temperatures used in this study were expressed in units of °C. Some of the selected compounds had only a single reported value, while others had melting temperature ranges. For the purposes of this work, it was necessary to obtain a single value for each compound. In such cases, only compounds with temperature range limited to 3 °C were selected and the mean of the melting range was then used as the melting temperature.

2.2 Generation of Structural Descriptors

Structural descriptors encode the constitutional, topological and geometrical characteristics of the molecular structure of investigated compounds. For the present study it has been decided to complete the existing set of descriptors with equation transforming polycyclic substructures into numerical values. At the beginning it was assumed that newly introduced descriptor should characterize the size of the rings forming the polycyclic fragments in a molecule, mutual dislocation, number of bonds between carbon–carbon atoms and carbon–heteroatoms, unsaturation degree and aromaticity. The basic concept applied in the formulation of the new descriptor comes from the idea of distance–based topological indices introduced by Wiener [27], Hosoya [28] Trinajstić [29] and Balaban [30–34].

The starting point in calculating the polycyclic substructure index is summation of neighborhood matrix elements. This matrix is formed on the base of a polycyclic graph. The polycyclic graphs are the graph representations of linear, branched or cyclic chains of rings, including poliphenylenes, spiranes, bridged ring systems and condensed polycyclic fragments of the molecule. The polycyclic index is defined as the sum of distances (neighborhoods) between all pairs of vertices of the respective polycyclic graph G modified in the following manner. Let G be the polycyclic graph representation of the ring skeleton of the substructure of an organic molecule. Let G possess k vertices (rings) labeled by v_1, v_2, ..., v_k and l edges labeled by e_1, e_2, ... , e_l. The distance between vertices v_i and v_j, called neighborhood N_{ij}, is:

$$N_{ij} = nr_{ij} - \frac{\Delta nb_{ij}}{\sum nb_{ij}} - \frac{nca_{ij}}{nar_i} \tag{1}$$

where nr_{ij} is the neighborhood range between the considered rings i and j, Δnb_{ij} is the difference between numbers of bonds connecting via two different and shortest paths the pair of considered rings (i,j), Σnb_{ij} is the sum of bonds connecting the pair of rings (i,j), nca_{ij} is the number of atoms common for rings (i,j), and nar_i is the total number of atoms in i–th ring. In Figure 1, the calculation of N_{ij} is demonstrated for the hydrogen–depleted graphs of anthracyl and fenanthryl substituents. Diagonal elements of the created matrices show the number of atoms NA forming consecutive rings in the polycyclic system.

The special treatment was applied according to biphenyl fragment where bond connecting phenylene rings is formally attributed as two–membered ring. Other examples of polycyclic fragments have been passed over because they do not appear in the collected sets of chemicals. The main component of the proposed index is the sum S of matrix elements: neighborhoods N_{ij} between all pairs of vertices of the underlying polycyclic substructure graph and membership NA of all cycles forming this molecular fragment:

$$S = \sum_{i}^{n} NA + \sum_{ij,\ i \neq j}^{n} N_{ij} \tag{2}$$

The following element of the elaborated index is the unsaturation degree *UI* of the multi–ring substituent:

$$UI = \frac{1}{2} \cdot \left[2 \cdot \left(n_{IV} + 1 \right) - n_I + n_{III} \right] \tag{3}$$

where n_I, n_{III}, n_{IV} represent the number of mono–, three–, and four–valent atoms forming the polycyclic fragment of the molecule. The comparative analysis of bonds longitude (singular, aromatic, double and triple) in different types of polycyclic compounds shows that the mean relative bonds shrinkage from single into double bond is –0.122. For this reason the sum of matrix graph elements is diminished according to this coefficient:

$$S_U = S - 0.122UI \tag{4}$$

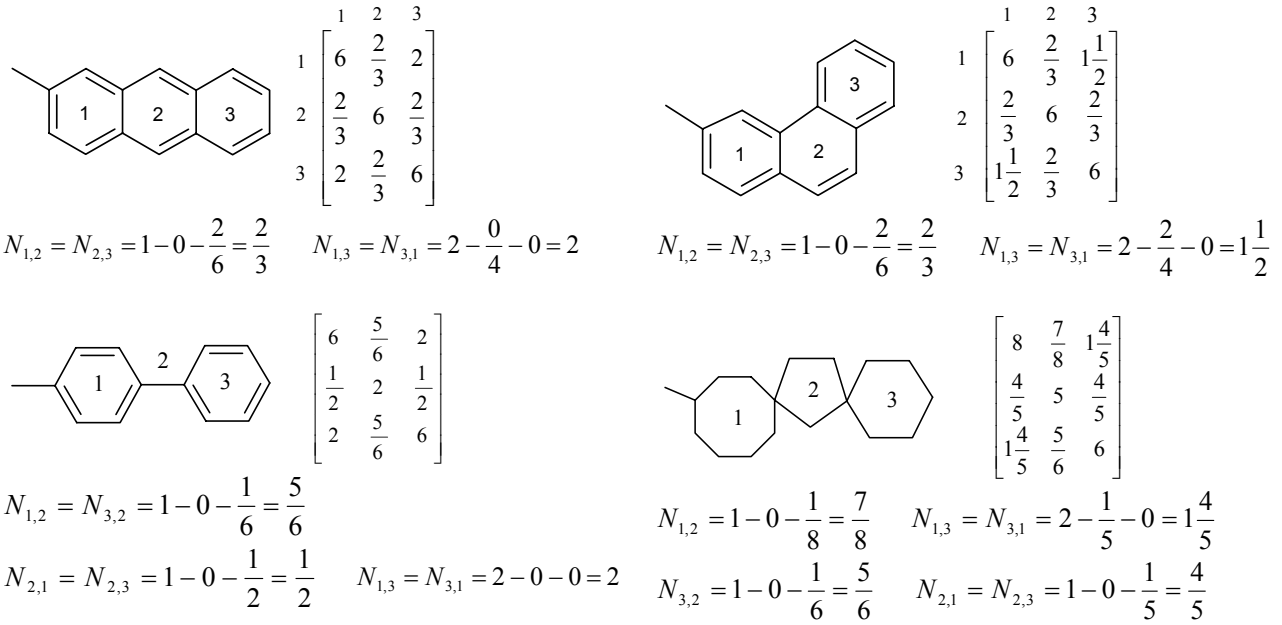


Figure 1. The calculation of N_{ij} for different types of polycyclic substituents.

In order to identify the aromatic rings system, the algorithm described in [35] has been applied. The set of rings fulfilling the preliminary aromaticity condition are preprocessed with the aim of revealing the combination(s) of aromatic rings possessing a maximal number of common atoms. If the obtained combination is a singular ring or a condensed system and the total number of delocalized electrons attributed to all atoms forming such combination fulfils the Hückel's rule (is $4n+2$, where n = 0, 1, 2, ...) then this combination of rings is aromatic. The aromaticity degree *AD* is:

$$AD = n\pi ea \cdot \frac{nab}{nb} \tag{5}$$

where $n\pi ea$ is the number of π electrons in aromatic part of the polycyclic substituent, *nab* is the number of aromatic bonds, and *nb* is the total number of bonds in the polycyclic fragment.

The aromatic character of a chemical compound reduces the molecular volume. A comparison of the calculated molar volume for different types of aromatic and non–aromatic polycyclic compounds shows a relative decreasing of this property with a coefficient of –0.18. The aromaticity degree was used as a third component of the polycyclic index reducing the sum of matrix graph elements *S*:

$$S_{AD} = S - 0.18AD \tag{6}$$

Finally, the elaborated components of the polycyclic index *PI* converts the considered structural features into numerical values scaled in the range 0.1 to 0.9, with *S* scaled down with the coefficient 0.01. Additionally, in order to differentiate the values obtained with the novel descriptor in relation to previously introduced indices describing monocyclic fragments, the shifted value of 0.3 was added:

$$PI = 0.3 + \left[0.01\left(\sum_{i}^{k} NA + \sum_{ij, i \neq j}^{l} N_{ij} - 0.122UI - 0.18AD \right) \right] \tag{7}$$

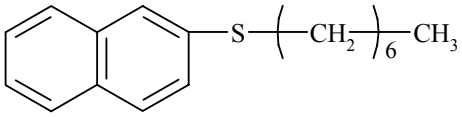
Table 1. Structural Features (Descriptors) Set

No.	Descriptor	No.	Descriptor
1	Number of C atoms in a molecule	32	Type, number and location of saturated side substituents connected to the ring
2	Number of C atoms in the main chain		
3–8	Number of heteroatoms: N, O, S, P, F, Cl	33	Type, number and location of unsaturated side substituents connected to the ring
9	Polycyclic index		
10	Total number of atoms in the molecule (without H)	34	Type, number and location of side substituents with heteroatoms connected to the ring
11	Geometric isomerism (*E*/*Z*) in the main chain		
12	Number of cyclic fragments	35	Indicators of cumulated, coupled or aromatic unsaturated bonds systems
13	Number of substitutents connected to main chain		
14–19	Type of substituents composed by C and H atoms	36	Unsaturation index of cyclic fragments
20	Average distance between tertiary and quaternary C atoms in aliphatic part of compounds (measured in number of bonds)	37	Number and location of S atoms in the main chain,
		38	Location of heteroatoms in the main chain
		39–40	Number and location of S atoms as a branches of the main chain (–S–)
21–24	Number and location of tertiary and quaternary C atoms in the main chain	41–42	Number and location of S atoms as a branches of the main chain connected via carbon atoms (–CH_2–S–, –CH_2–SO_2–, etc.)
25–28	Number and location of double and triple bonds in the molecule structure		
29	Location of cyclic substituents connected to the main chain	43	Average distance between carbon atoms with multiple bonds and S atoms
30	Number and location of double bonds in cyclic substituent	44–46	Number and location of S atoms in a substituents of the cyclic fragments
31	Location of substituents connected to cyclic fragments of molecule (cyclic substituents)	47–56	Structural descriptors representing molecular features analogous to 37–46, describing the presence of oxygen atoms

In Table 1 the whole pool of selected features is presented, grouped under three main headings, namely elementary composition (1–8, 10), construction of molecule (9, 11–36) and the way of connection of heteroatoms, *i.e.* sulphur (37–46) and oxygen (47–56). Using the equations described in [25] and the newly elaborated polycyclic index, the structures of the investigated compounds were coded into a 46–component vector of numerical values for sulfides and a 56–component vector for sulfones. The melting temperature t_m was scaled with the formula:

$$t_{m\,sc} = (t_m + 200)/1000 \quad (8)$$

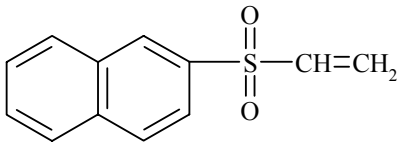
Two examples of numerical representation for heptyl–2–naphtyl sulfide and 2–naphthylvinyl sulfone together with scaled value of the melting temperature are presented in Figure 2.



x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}
0.27	0.17	0	0	0.1	0	0	0	0.406	0.28	0	0.1	0.1	0	0	0	0

x_{18}	x_{19}	x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	x_{30}	x_{31}	x_{32}	x_{33}	x_{34}
0	0	0	0	0	0	0	0	0	0	0	0.225	0.1584	0	0	0	0

x_{35}	x_{36}	x_{37}	x_{38}	x_{39}	x_{40}	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	y
0.3	0.5	0.1	0.2125	0	0	0	0	0.16	0	0	0	0.234



x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
0.22	0.12	0	0.2	0.1	0	0	0	0.406	0.25	0.5	0.1	0.1	0	0	0	0	0	0

x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	x_{30}	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
0	0	0	0	0	0	0	0.1	0.102	0.433	0.1564	0	0	0	0	0.3	0.5	0.1	0.3

x_{39}	x_{40}	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}	x_{48}	x_{49}	x_{50}	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}	x_{56}	y
0	0	0	0	0.136	0	0	0	0	0	0.2	0.1389	0	0	0.1962	0	0	0	0.294

Figure 2. Numerical representation of heptyl-2-naphtyl sulfide and 2-naphthylvinyl sulfone.

Other vectors of descriptors representing the structures of all investigated compounds together with the scaled values of melting temperatures are placed in the file Sulfid_D.txt for sulfides and file Sulfon_D.txt for sulfones (see the supplementary material).

2.3 Computer Software

All computations were performed on an IBM PC–type microcomputer, running under the Windows 98 operating system. The artificial neural networks computations were carried out with the network simulation program Statistica Neural Network [36]. Data manipulation and interpretation of the obtained results was carried out by means of Microsoft Excel v. 97.

2.4 Neural Networks

In this study multilayer feedforward networks were used. The architecture of multilayer networks consists of an input layer, one hidden layer and an output layer. The input layer contains one node for each structural index. The output layer has one node generating the scaled estimated value of the melting temperature. It is known that in the hidden layer learning and approximation

occurs. The number of hidden neurons needs to be sufficient to ensure that the information contained in the data utilized for training the network is adequately represented. On the other hand the small number of collected examples (possible to select from available data sources) limited the complexity of the networks. For this reason only networks with two processing layers and two nodes in hidden layer were considered. The starting networks architectures were determined applying the automatic optimization procedure available in Statistica Neural Network v 4.0 package programs, named Intelligent Problem Solver (IPS) [36]. The IPS program was forced to search optimal networks according to the above limits. The candidate networks architectures (with best performance characteristics) generated by IPS were retained for further learning and testing of their predictive ability: 40:2:1 for sulfides, and 44:2:1 for sulfones. In both networks, the dimensionality of the input layer corresponds to the number of descriptors having non–zero values for all compounds. These descriptors were chosen as valid input variables. The sigmoid function was used for the processing neurons in the hidden layer and the linear one in output neuron. The networks were preliminary trained for a period of 50 epochs by the standard back propagation procedure and then the learning process was continued over a dozen cycles with a conjugated gradient algorithm.

2.5 Reduction of Structural Descriptors

The next experiment on the sulfides and sulfones compounds was to determine whether a reduced set of descriptors could provide similarly effective or better models. The selection of the optimal set of input variables for both types of investigated compounds has been carried out on the base of sensitivity analysis SA available in Statistica Neural Network.

To perform the selection of variables new sets of networks (five 40:2:1 for sulfides and five 44:2:1 for sulfones) were trained using the IPS procedure and applying random subdivision of the entire sets of examples (in proportion 4:1) into training and verification set. All the multilayer neural networks (with linear output neuron) were examined separately. Comparing the sets of unimportant variables proposed by the SA procedure twenty input variables common for all five sets generated for sulfides were removed. The second half of input variables (considered as important for melting temperature prediction) was retained for further processing. These highly important variables for sulfides are: 2, 5, 9, 10, 12, 13, 14, 16, 17, 21, 25, 26, 30, 31, 32, 35, 38, 39, and 43 (see Table 1). In the same way twenty–two descriptors were selected from the pool of indices applied for the sulfones processing: 1, 2, 4, 5, 9, 11, 12, 13, 15, 21, 27, 28, 30, 35, 36, 37, 38, 39, 40, 49, 50, and 53 (see Table 1).

The reduced data sets, containing a 20 components vector of numerical values for sulfides and 22 components for sulfones, were used for the final selection of the optimal network, which was performed applying the IPS procedure once again. Because of the collected sets of examples are relatively small according to the size of input vectors the cases were reassigned randomly to only training and cross–validation sets in the proportion of 4:1. The best networks from the preliminary

test optimized by the IPS automatic procedure (with the lower training and verification mean square errors figures) were retained for further optimization: 20:2:1 for sulfides and 22:2:1 for sulfones, respectively.

For the final optimization the conjugated gradient algorithm was used applying the leave–20%–out procedure. In this procedure 20% of the objects were selected out one after another, whereas for every selection the model was build up with remaining 80% examples. Next, this model was used to predict the melting temperature for selected molecules. Joined results of the melting temperature estimation gave information on the prediction ability and on model quality for the selected training and prediction sets. Each time the training was stopped when the root mean square error averaged over the training set had reached minimum value. Depending on particular network structure and training set, this occurred after about 380 epochs for sulfides and 400 epochs for sulfones. To avoid over–training of the neural network, the output error between the seen and those expected values has been calculated as well as for training and cross–validation set examples. Training was stopped (before the training error has reached mentioned above value) when the *RMS* error obtained for the control data was lowest. According to earlier experience [22,25], where the best ANN models were developed applying sigmoid function in all processing neurons, the final attempts of improvement was carried out replacing the linear activation function with sigmoid in the output neuron. The weights optimization of the modified ANN was performed using previous assignment of investigated cases into training and cross–validation sets and applying conjugate gradients learning algorithm. The learning process was continued over period of about 200 epochs for sulfides and 220 for sulfones. The linear networks generated by IPS, as a least squares linear model [36], have been retained for comparison purposes, with the structure 40:1 for sulfides and 44:1 for sulfones. These and the multilayer final network structures are presented in Tables 2 and 3.

2.6 Statistical Parameters

When the optimization process of all investigated models was completed, the output data obtained for both sets of examples has been decoded to their normal values expressed in °C:

$$t_m = t_{m\,sc} \times 1000 – 200 \tag{9}$$

In the next step we have evaluated the ANN models generated with the pool of elaborated structure descriptors. The statistical quality of the ANN results for both training and cross–validation sets was evaluated using the following parameters: squared correlation coefficient R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^c)^2}{\sum_{i=1}^{n}(y_i - y_0)^2} \tag{10}$$

average recognition and prediction errors *AER*:

$$AER = \frac{1}{n}\sum_{i=1}^{n}(y_i^c - y_i) \tag{11}$$

average absolute error *AAE*:

$$AAE = \frac{1}{n}\sum\left|y_i^c - y_i\right| \tag{12}$$

and standard deviation *SD*:

$$SD = \sqrt{\frac{n\sum_{i=1}^{n}(y_i^c - y_i)^2 - \left[\sum_{i=1}^{n}(y_i^c - y_i)\right]^2}{n^2}} \tag{13}$$

In these equations y_i represents the experimental target value (t_m) for the *i-th* compound, y_0 denotes the associated mean and y_i^c represents the calculated melting temperature using the ANN model, and *n* indicates number of examples in training and cross–validation sets.

3 RESULTS AND DISCUSSION

The goal of this paper was to investigate whether previously elaborated structural parameters, completed with newly proposed polycyclic index would be useful to predict the melting temperature of sulfur containing organic compounds. The structures of investigated sulfides and sulfones are stored as ISIS Draw files deposited in the supplementary material, in the archive files Sulfide.zip and Sulfone.zip, respectively. Numerical representations (vectors of structural indices) obtained in the coding phase are collected in the files Sulfid_D.txt and Sulfon_D.txt also deposited in the supplement to this article. The cross–validation set examples labeled by the numbers identifying each compound together with experimental and predicted melting temperatures are placed in the files SulfidPR.txt and SulfonPR.txt. The statistical results of the ANN modeling of sulfides and sulfones melting temperatures are presented in Tables 2 and 3.

Table 2. Statistics of ANN for Sulfides: Linear and Two-Layers for Calculating Melting Temperatures of Sulfides, with Linear Output Neuron (lon) and Sigmoid Activation Function for the Output Neuron (sfon)

Statistics	40:1		20:2:1 (lon)		20:2:1 (sfon)	
	training	cross-valid.	training	cross-valid.	training	cross-valid.
AER	0	2.36	-0.14	-5.87	0.21	0.072
AAE	17.84	24.90	16.22	21.69	14.80	15.20
SD	26.41	40.11	24.32	28.11	21.57	21.93
R^2	0.843	0.589	0.852	0.807	0.883	0.880

Linear regression of the melting temperature against 40 structural features describing sulfides molecules, using linear network, is summarized by the respective statistics in the first pair of columns in Table 2. The acceptable performance of the linear model obtained for training set examples is opposed to a poor prediction results which is seen by lower R^2 (0.589 *vs*. 0.843) as well

as greater *AAE* (24.9 *vs*. 17.84 °C), and the standard deviation 40.11 *vs*. 26.41 °C. This result reveals that coded structural functionalities of the sulfides correlate with their melting temperatures in a nonlinear way.

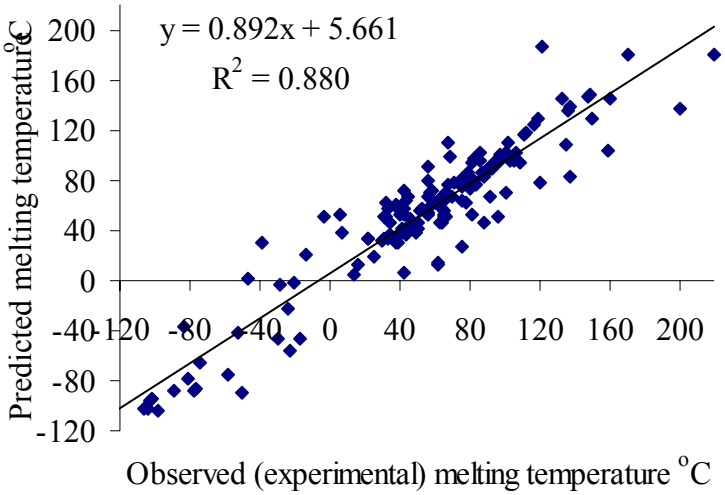


Figure 3. Predicted melting temperatures of sulfides versus experimental data.

The comparison of the statistical parameters obtained for melting temperature calibration and prediction, using the linear regression model developed for sulfones, is presented in Table 3. The calibration melting temperature for sulfones gave far worse statistical parameters then those obtained for sulfides. First of all, the higher *AAE* value (20.36 °C) has shown the worse adaptability of linear regression model for sulfones structure–property relationship modeling task. The predictions for the sulfones melting temperatures for cross–validation set examples gave the following statistics: *AER* –1.72 °C, *AAE* 30.27 °C, *SD* 39.47 °C and R^2 0.469, demonstrating the overall inaccuracy of the liner model.

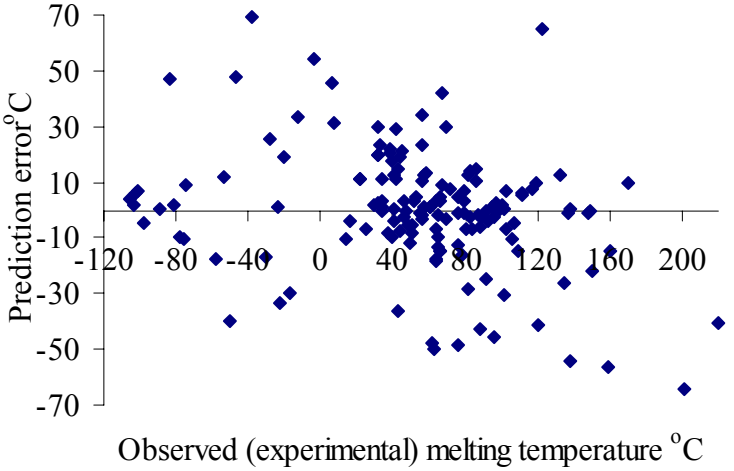


Figure 4. Distribution of the prediction error (*PER*) versus experimental melting temperatures for sulfides.

The statistical results of the multilayer ANN listed in Tables 2 and 3 show that significant improvement is obtained by using nonlinear models in comparison with the linear network. The final network architectures with best performance used for estimation of the melting temperatures for sulfides have 20 inputs and 2 neurons in the hidden layer. Therefore, the nonlinear models have a comparable number of adjustable parameters (connections between neurons) with the linear one. The statistics presented in the Table 2 for the network with sigmoid activation function in the output neuron indicate noticeable better predictive ability than for the network with linear output neuron.

The predicted versus observed melting temperatures for the sulfides cross–validation examples are displayed in Figure 3. The distribution of points along the regression line is quite good and no extreme outliers are seen. Obtained predictions for investigated groups of compound fit well to experimental data with the high correlation coefficient of $R = 0.938$. The calculated parameters of the regression equation in Figure 3 has a slope equal to 0.892 and an intersect of 5.66.

Table 3. Statistics of ANN for Sulfones: Linear and Two-Layers for Calculating Melting Temperatures of Sulfides, with Linear Output Neuron (lon) and Sigmoid Activation Function for the Output Neuron (sfon)

Statistics	44:1		22:2:1(lon)		22:2:1(sfon)	
	training	cross-valid.	training	cross-valid.	training	cross-valid.
AER	-1.71	-1.72	-7.74	-0.65	-0.25	-1.43
AAE	20.36	30.27	22.03	21.63	18.63	18.93
SD	25.39	39.47	25.98	26.80	23.26	23.66
R^2	0.763	0.469	0.752	0.735	0.800	0.794

The distribution of prediction errors *PER* for sulfides, over the experimental value range, is shown in Figure 4. The error for each tested compound was calculated as $PER = t_{m\ pr} - t_{m\ exp}$ where $t_{m\ pr}$ is the estimated melting temperature and $t_{m\ exp}$ the experimental value. The greatest overestimation of melting point is observed for the tested compounds (see the file SulfidPR.txt): **96** tetrakis(methylthio) methane, **203** 2,2–bis(tolyl–4–thio)–1,1–di–4–tolyl ethene, **219** 3–methylbut–2–enylphenyl sulfide, and greatest underestimation for the **230** tetrakis(tolyl–4–thio) ethene, **139** tetrakis(phenylthio) methane, and **120** methylphenyl sulfide. All enumerated compounds (except the last one) were predicted with considerable errors by the remaining ANN models. Most of them contain two or four sulfur atoms placed symmetrically in the structure. This list of outliers suggests that compounds with symmetrically located substituents connected with sulfur atoms to the central part of a molecule need more elaborated structural descriptors.

As can be seen from the Table 3, the main statistics characterizing modeling results for sulfones are slightly inferior to those obtained for sulfides. The statistical indices show that the networks with sigmoid activation function in the output neuron (last column of Table 3) behave better then the networks with a linear output neuron.

The correlation of the experimental and estimated melting temperatures using this network model is presented in Figure 5. Compared to the results obtained for sulfides, we have obtained a poorer fit for predictions to experimental data, as can be seen from smaller correlation factor $R =$

0.891 as well as greater *SD* equal to 23.66 °C.

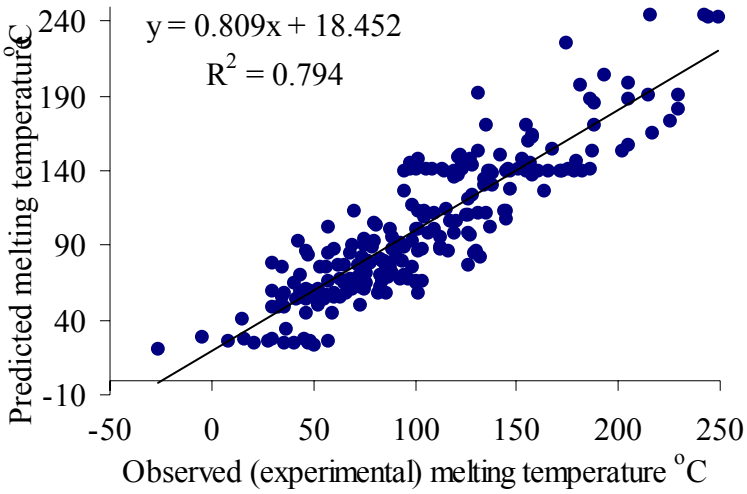


Figure 5. Predicted melting temperatures of sulfones versus experimental data.

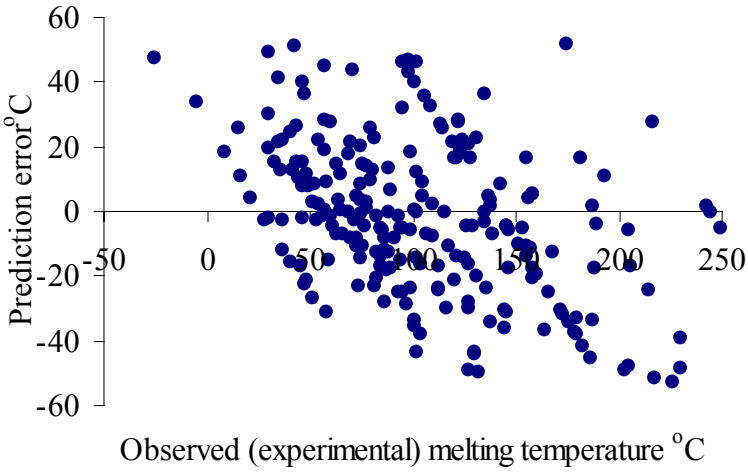


Figure 6. Distribution of the prediction error *PER* versus experimental melting temperatures for sulfones.

The calculated prediction error *PER* against experimental melting temperature of sulfones is plotted in Figure 6. The largest error of melting temperature are observed for **176** tris(tolyl–4–sulfonyl) ethane (60.79 °C), **61** methylsulfonylphenylsulfonyltolyl–4–sulfonyl methane (51.71 °C), **29** phenyl–1,1–dimethylpropyl sulfone (49.45 °C), **230** divinyl sulfone (47.95 °C), **22** ethylsulfonylmethyl–sulfonylphenylsulfonyl methane (–52.42 °C), **34** 2–methylsulfonylethyl–2–phenylsulfonylethyl sulfone (–51.08 °C), **220** methyl–2,4,6–trimethylphenyl sulfone (–49.58 °C) **82** 1–naphthyl–2–sulfonyl–2–tolyl–4–sulfonyl ethane (–48.96 °C) (see the file SulfonPR.txt). A visual inspection revealed no structural explanation for the reason why these compounds were not fit well with the rest of the sulfones. Five of them are highly sulfonated compounds containing two and three sulfonyl groups in the molecule.

4 CONCLUSIONS

The QSPR studies described in this paper provide strong evidence that the tested structural descriptors are useful and effective for this goal. They are representing particular structural features that can be related to the melting temperatures of sulfur containing chemicals. The melting temperatures of sulfides and sulfones, comprising various types of structures (aliphatic: linear and branched, cyclic: alicyclic and aromatic) have been successfully predicted using artificial neural networks. The results of this work show that a feed–forward multilayer neural network can be easily trained to model the melting temperatures of sulfides and sulfones. The ANN models are highly empirical, but well adapted to dealing with complicated relationships which are observed between structure and chemicals properties.

Acknowledgment

The author wishes to thank Prof. Z.S. Hippe for his support and helpful discussion.

Supplementary Material

The molecular files containing the structure for all sulfides and sulfones used in the QSPR models are deposited as an archive in the files Sulfide.zip and Sulfone.zip, respectively. The structural descriptors are collected in the files Sulfid_D.txt and Sulfon_D.txt. The prediction results obtained for cross–validation using the best neural models are placed in the files SulfidPR.txt and SulfonPR.txt.

5 REFERENCES

[1] A. R. Katritzky, M. Karelson, and V. S. Lobanov, QSPR as a Means of Predicting and Understanding Chemical and Physical Properties in Terms of Structure, *Pure Appl. Chem.* **1997**, *69*, 245–249.
[2] R. E. Aries, D. P. Lidiard, and R. A. Spragg, Principal Component Analysis, *Chem. Br.* 1991, pp. 821–824.
[3] S. Wold, PLS for Multivariate Linear Modelling; in: *Chemometric Methods in Molecular Design*, Ed. H. van de Waterbeemd, VCH, Weinheim, Germany, 1995, pp.195–218.
[4] S. Wold and M. Sjöström, Chemometrics, Present and Future Success, *Chemom. Intell. Lab. Syst.* **1998**, *44*, 3–14.
[5] P. C. Jurs, S. L. Dixon, and L. M. Egolf in: *Chemometric Methods in Molecular Design*, Ed. H. van de Waterbeemd VCH, Weinheim, Germany, 1995, p. 15.
[6] A. T. Balaban (Ed.), *From Chemical Topology to Three-Dimensional Geometry*, Plenum, New York, 1997.
[7] A. R. Katritzky, *Understanding How Chemical Structure Determines Physical Properties*, http://ark2.chem.ufl.edu/research/qspr_2000/QSPR_files.
[8] L. M. Egolf and P. C. Jurs, Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 616–625.
[9] M. E. Sigman and S. S. Rives, Prediction of Atomic Ionization Potentials I–III Using an Artificial Neural Network, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 617–620.
[10] A. A. Gakh, E. G. Gakh, B. G. Sumpter, and D. W. Noid, Neural Network–Graph Theory Approach to the Prediction of Physical Properties of Organic Compounds, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 832–839.
[11] A. T. Balaban, S. C. Basak, T. Colburn, and G. D. Grunwald, Correlation Between Structure and Normal Boiling Points of Haloalkanes C_1–C_4 Using Neural Networks, *J. Chem. Inf. Comput. Sci.*, *34*, 1118–1121.
[12] D. Cherqaoui and D. Villemin, Use of a Neural Networks to Determine the Boiling Points of Alkanes, *J. Chem. Soc., Faraday Trans.* **1994**, *90*, 97–102.
[13] T. H. Fisher, W. P. Petersen, and H. P. Lüthi, A New Optimisation Technique for Artificial Neural Networks Applied to Prediction of Force Constants of Large Molecules, *J. Comput. Chem.* **1995** *16*, 923–936.
[14] L. H. Hall and C. T. Story, Boiling Point and Critical Temperature of a Heterogeneous Data Set: QSAR with Atom Type Electrotopological State Indices Using Artificial Neural Networks, *J. Chem. Inf. Comput. Sci.* **1996**; *36*, 1004–1014.
[15] L. Vera, M. E. Guzman, and P. A. Ortega, Redes Neuronales y Semejanza Cuantica: Applicacion a Los Isomeros

de Octano, *Bol. Soc. Chil. Quim*. **1997**, *42*, 341–348.
[16] T. Suzuki, R–U. Ebert, and G. Schüürmann, Development of Both Linear and Nonlinear Method to Predict the Liquid Viscosity at 20 °C of Organic Compounds, *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1122–1128.
[17] O. Ivanciuc, The Neural Network MolNet Prediction of Alkane Enthalpies, *Anal. Chim. Acta* **1999**, *384*, 271–284.
[18] S. Arupjyoti and S. Iragavarapu, New Electrotopological Descriptor for Prediction of Boiling Points of Alkanes and Aliphatic Alcohols Through Artificial Neural Network and Multiple Linear Regression Analysis, *Comput. Chem*. **1998**, *22*, 515–522.
[19] R. C. Schweitzeri and J. B. Morris, The Development of a Quantitative Structure Property Relationship (QSPR) for the Prediction of Dielectric Constant Using Neural Networks, *Anal. Chim. Acta* **1999**, *384*, 285–303.
[20] J. Tetteh, T. Suzuki, E. Metcalfe, and S. Howells, Quantitative Structure–Property Relationships for the Estimation of Boiling Point and Flash Point Using a Radial Basis Function Neural Network, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 491–507.
[21] E. S. Goll and P. C. Jurs, Prediction of the Normal Boiling Points of Organic Compounds from Molecular Structures with a Computational Neural Network Model, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 974–983.
[22] J. Kozioł, Application of Artificial Neural Networks for Prediction of Phase Transition Temperature of Organic Compounds, *Proc. of Int. Conf.: Progress in Computing of Physical Properties*, 18–20 Nov. Warsaw, Poland, 1999.
[23] G. Espinosa, D. Yaffe, Y. Cohen, A. Arenas, and F. Giralt, Neural Network Based Quantitative Structural Property Relations (QSPRs) for Predicting Boiling Points of Aliphatic Hydrocarbons, *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 859–879.
[24] I. V. Tetko, V. Yu. Tanchuk, and A. E. P. Villa, Prediction of n–Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E–State Indices, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1407–1421.
[25] J. Kozioł, Neural Network Modeling of Physical Properties of Chemical Compounds, Int. J. Quantum Chem. **2001**, *84*, 117–126.
[26] Beilstein Handbuch der Organishen Chemie, Vierter Auflange, Springer–Verlag, Berlin, 1958.
[27] H. Wiener, Structural Determination of Parafin Boiling Points, *J. Am. Chem. Soc*. **1947**, *69*, 17–20.
[28] H. Hosoya, On Some Counting Polynomials in Chemistry, *Discr. Appl. Math*. **1988**, *19*, 239–257.
[29] D. Plavšić, S. Nikolić, N. Trinajstić, and Z. Mihalić, On the Harary Index for the Characterization of Chemical Graphs, J. Match. Chem. **1993**, *12*, 235–250.
[30] A. T. Balaban, Highly Discriminating Distance–Based Topological Index, *Chem. Phys. Lett*. **1982**, *89*, 399–404.
[31] A. T. Balaban, Topological Indices Based on Topological Distances in Molecular Graphs, *Pure Appl. Chem*. **1983**, *55*, 199–206.
[32] O. Ivanciuc, T.–S. Balaban, and A. T. Balaban, Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices, *J. Math. Chem*. **1993**, *12*, 309–318.
[33] O. Ivanciuc and A. T. Balaban, Design of Topological Indices. Part 8. Path Matrices and Derived Molecular Graph Invariants, *MATCH* (*Commun. Math. Chem*.) **1994**, *30*, 141–152.
[34] O. Ivanciuc, T. Ivanciuc, D. J. Klein, W. A. Seitz, and A. T. Balaban, Wiener Index Extension by Counting Even/Odd Graph Distances, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 536–549.
[35] Z. Hippe, Artificial Intelligence in Chemistry. Structure Elucidation and Simulation of Organic Reactions, PWN/Elsevier, Warsaw/Amsterdam, 1991, pp. 215–219.
[36] Statistica Neural Networks v. 4.0, http://www.statsoft.com/stat_nn.html.

Biographies

Julian Kozioł is assistant professor of Analytical Chemistry at the Department of Physical Chemistry, Rzeszów University of Technology, Poland.