Inter*net* Electronic Journal of Molecular Design

May 2003, Volume 2, Number 5, Pages 315–333

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 9

Guest Editor: Jun–ichi Aihara

Neural Network Modeling of Refractive Indexes of Phosphorus–Containing Organic Compounds

Julian Kozioł

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

Received: February 21, 2003; Revised: April 10, 2003; Accepted: April 15, 2003; Published: May 31, 2003

Citation of the article:
J. Kozioł, Neural Network Modeling of Refractive Indexes of Phosphorus–Containing Organic Compounds, *Internet Electron. J. Mol. Des.* **2003**, *2*, 315–333, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Neural Network Modeling of Refractive Indexes of Phosphorus–Containing Organic Compounds[#]

Julian Kozioł*

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

Received: February 21, 2003; Revised: April 10, 2003; Accepted: April 15, 2003; Published: May 31, 2003

Internet Electron. J. Mol. Des. **2003**, *2* (5), 315–333

Abstract

Motivation. One of the most intensively explored areas of contemporary computational chemistry is searching for a comprehensive numerical description of chemical structures and for methods that enable to develop efficient and credible QSPR (quantitative structure–property relationships) models. Among these methods artificial neural networks (ANN) turned out to be a very promising methodology in obtaining models converting structural descriptors into different properties of chemicals.

Method. Five different models relating structural descriptors to refractive indexes of phosphorus containing organic compounds have been developed using ANN. A newly elaborated set of molecular descriptors is evaluated to determine their usefulness for QSPR studies. Using a data set containing 180 phosphates and diphosphates, ANN trained with the back propagation and conjugated gradient algorithms are able to predict the refractive index with relatively high accuracy.

Results. The results obtained show good predictive ability for the ANN models, giving the average prediction error of 0.24% and R^2_{cv} equal to about 0.99.

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 phosphates refractive index.

Keywords. QSPR; quantitative structure–property relationships; molecular descriptors; artificial neural networks; refractive index; phosphate; diphosphate.

Abbreviations and notations

ANN, artificial neural network	QSPR, quantitative structure–property relationships
IPS, intelligent problem solver	RI, refractive index
PER, prediction error	SA, sensitivity analysis
PMI, polymethylene index	SNN, Statistica Neural Networks

1 INTRODUCTION

An estimation of physicochemical properties values for chemical substances, particularly organic compounds, has gained an important role and became one of the most explored areas of the research in computational chemistry [1,2]. It is caused by the permanent need of physical and chemical data

[#] Dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday.

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

for rapidly developing branches of contemporary chemistry and relative fields like medicine, environmental protection etc. Wide application of combinatorial chemistry tools produced special interest in obtaining reliable models, which can estimate different properties of the known structures but not yet synthesized chemical molecules. Also experimental determination of properties for newly synthesized chemical compounds may encounter obstacles coming from insufficient quantity or instability of the available material. As a consequence, many methods using quantitative structure–property relationships (QSPR) have been proposed to estimate the physicochemical properties of these compounds enclosing theoretical and practical aspects of the model development; structural parameterization methods [3–8], structure descriptor selection [9–11], statistical methods [12–18] and available programs [19–25].

Over last decades, besides classical methods of computing the properties of chemical compounds, various statistical methods as multiply linear regression, cluster analysis and partial least–squares have been used for QSPR studies [26,27]. Currently, neural networks, representing general nonlinear methods, were used with encouraging success to correlate structural parameters with the observed properties [28–50]. Artificial Neural Networks (ANN) are well–suited to describe structure–property relations. Moreover, ANN may consider not only particular structure characteristics, but also interrelations and interdependences between mutually influencing structural descriptors. Therefore, ANN can be easily adapted for processing large vectors of structural data formed by various descriptors.

A set of indices in the form of an algebraic equation converting structural descriptors into a multicomponent vector of numerical values, scaled in the range of 0.1 to 0.9, was proposed [48]. The key feature of this coding scheme is the treatment of each molecule as a linear structure with linear, branched, and/or cyclic substituents. The elaborated coding method is useful for estimating the boiling points of hydrocarbons, nitrogen and oxygen containing compounds [42,48], the refractive index of amines [48], the melting points of amides [48], sulfides and sulfones [49]. However, it was not applied to compounds with other types of heteroatoms. The work described here extends this model for the phosphorus containing compounds.

Among experimentally determined properties of different types of phosphorus compounds, the refractive indexes are probably the most widely available and precisely measured data described in accessible data collections. Therefore, refractive index appears to be very convenient for the verification of practical applicability of developed descriptors for QSPR modeling.

Besides, the refractive index *n* is one of the most important optical properties that is frequently employed to characterize organic compounds in laboratory practice and in material science to evaluate the applicability of materials for various purposes. The refractive index is strictly associated with other significant molecular properties, particularly the molar refraction, polarizability, dielectric constant, etc.

Over the years numerous methods for estimation these of high importance properties from chemical structures have been developed. Two major approaches have been applied for this aim. The first was the use of molecular group contribution methods for the estimation of the molar refraction [51–53]. All these methods have been developed from experimental data. The difficulty of this approach is connected with the definition of a constituent set for groups and by the necessity to compute the contribution of each group from a statistically significant number of molecular structures where the respective group is present. Obtained group contribution schemes are restricted to molecules containing only the fragments present in the calibration set. Also, this method is limited to compounds containing structural functionalities for which the group contributions are available. Beside this, possible interactions between different groups present in the molecule would lead to the non–additivity of a property.

Second, the QSPR approach has gained nowadays an increasing importance overcoming limitations of the group contribution techniques. This undergo was successfully applied to estimate the molar refraction [54,55] of alkanes and alkyl substituted benzenes, the refractive index of group of diverse organic compounds [56] or both these properties for the alkyl hydroperoxides [57] using linear and nonlinear equations converting exclusively structure derivative indices into a needed property. Also, neural networks have been applied for a development of models of the refractive indices. A model based on seven topological descriptors obtained by Gakh *et al*. [30] reached an average error of 0.16% for predictions of the refractive index of hydrocarbons. Other reported models based on ANN having 5:5:1 architecture [58] were able to predict the refractive indexes of 55 alkanes with a relative standard deviation (RSD) of 0.11% and 66 alkanes [59] with an average RSD of 0.13% and R^2 = 0.978 respectively. The refractive indexes of 133 alkanes were also estimated with the MolNet models developed by Ivanciuc [60]. The best model predictions, obtained for 25 alkanes separated in the test set, gave the correlation coefficient equal to 0.972 and the standard deviation of 0.0033. The ANN model-estimated refractive indexes of different amines, reported in [48], had an average error of 0.17% (R^2 0.973). In this paper ANN QSPR models for the prediction of the refractive index developed for a data set of 180 phosphates and diphosphates will be described. The obtained models, relied on information derived from the compound molecular structure only, have shown good predictive performance in estimation of the refractive index.

2 MATERIALS AND METHODS

2.1 Chemical Data

Chemical structures and refractive indexes of 159 phosphates and 21 diphosphates were selected from [61]. The data set contains different types of structures: aliphatic (linear and branched), cyclic and aromatic. All experimental refractive index data used in this study were measured at a wavelength of 589 nm. Some of the selected compounds had more than one RI reported value

measured at different temperatures, ranged between 17–35 ºC. For the purposes of this work, it has been decided to use refractive data obtained in the range limited to 20–25 ºC. In this way, 25 phosphates are characterised by two RI values determined at 20 ºC and at a second temperature value from the interval of over 20 to 25 ºC.

2.2 Generation of Structure Descriptors

The structure descriptors encode the elementary, topological and geometrical characteristics of the molecular structure of investigated compounds. For the needs of the present study it has been decided to complete the existing set of descriptors with a descriptor for the cyclic phosphates moieties (Figure 1) into numerical values. At the beginning it was assumed that the newly introduced descriptor should characterize: the size of a ring forming cyclic fragment of a molecule, the bonds length between carbon–oxygen and oxygen–phosphorus atoms and the unsaturation degree. The basic concept applied in the formulation of the new descriptor, named the polymethylene index (*PMI*), comes from the idea of cyclic substructures described in [49].

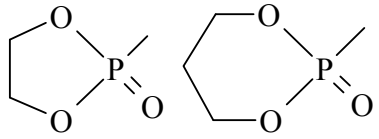


Figure 1. Substructures of different types of cyclic phosphates.

The starting point was to compute the size (S_R) of the ring. It has been defined as the sum of relative bond lengths (*rbl*) between two adjacent non–hydrogen atoms forming considered ring:

$$S_R = \sum_i rbl_i \tag{1}$$

The bond length is related to the average length of carbon–carbon bonds in cyclic hydrocarbons. The next element of the elaborated index is the unsaturation degree (*UI*) of the ring substructure defined according to:

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

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

$$S_U = S_R - 0.122UI \tag{3}$$

Finally, the elaborated components of the polymethylene index (*PMI*) have been arranged in the form of an algebraic equation converting considered structural feature into numerical value, scaled in the range 0.1 to 0.9. Therefore, the S_u value is scaled down with the coefficient 0.1:

$$PMI = 0.1(S_U - 0.122UI) \tag{4}$$

The whole pool of selected structural descriptors is presented in Table 1.

Table 1. The Set of Structural Descriptors

No.	Descriptor	No.	Descriptor
1	Number of C atoms in a molecule	32	Type, number and location of saturated side sub–stituents connected to the ring
2	Number of C atoms in the main chain		
3–8	Numbers of heteroatoms: N, O, S, P, F, Cl	33	Type, number and location of unsaturated side substituents connected to the ring
9	Cyclic esters (*PMI*) 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	Indicator of cumulated, coupled unsaturated bonds systems, including aromatic
13	Number of substitutents connected to main chain		
14–19	Types 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 O atoms in the main chain
		38	Location of oxygen atoms in the main chain
		39–40	Numbers and locations of heteroatoms as branches of the main chain
21–24	Number and location of tertiary and quaternary C atoms in the main chain	41–42	Numbers and locations of O atoms as a branches of the main chain connected via carbon atoms
25–28	Number and location of double and triple bonds in the molecule structure	43	Average distance between carbon atoms with multiple bonds and O atoms
29	Location of cyclic substituents connected to the main chain	44–46	Number and location of O atoms in a substituents of the cyclic fragments
30	Number and location of double bonds in cyclic substituents	47–56	Structural descriptors representing molecular descriptors analogous to 37–46, describing the presence of phosphorus atoms.
31	Location of substituents connected to cyclic fragments of molecule (cyclic substituents)		

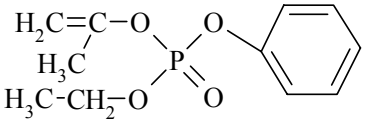
The structural descriptors collected in Table 1 are grouped under three main groups: the elementary composition (1–8, 10), the construction of a molecule (9, 11–36) and the way of heteroatoms connection: oxygen (37–46) and phosphorus (47–56). Using the equations described in [48] and the newly elaborated cyclic index, structures of investigated compounds were coded into a 56–component vector of numerical values. When a given structural feature was absent, adequate component of the code vector was zeroed. Also, the temperatures of the RI measurement, related to coded structures, was scaled according to the following formula:

$$T_{sc} = T/10 \tag{5}$$

and added as descriptor number 57.

An example of a numerical representation of the ethyl–isopropenyl–phenyl phosphate and 1,1–diethyl–2,2–dipropyl diphosphate structures together with the temperature of refractive index measurement and its experimental value are presented in Figure 2.

Other vectors of descriptors numerical values representing structures of all investigated compounds together with the value of refractive indexes are placed in the file *Phosphate_D.txt* (see the supplementary material).

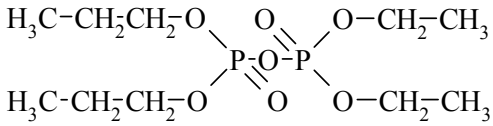


Ph_60

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}	x_{20}
1.1	0.4	0	0.4	0	0.1	0	0	0	3.1	0.5	0.1	0.03	0.005	0.5755	0	0	0	0	0.1

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}
0.1	0.025	0.1	0.05	0	0	0.1	0.00796	0.343	0.062	0	0	0	0	0.3	0.15	0.2	0.09163

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}	x_{57}	RI
0.1	0.01735	0	0	0.1722	0	0	0	0.1	0.3306	0	0	0	0	0.15714	0	0	0	2	1.4845



Ph_58

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}	x_{20}
0.1	0.5	0	0.7	0	0.2	0	0	0	4.3	0	0	0.4	0.0343	0.0793	0	0	0	0	0.1

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}	x_{39}
0	0	0.2	0.04091	0	0	0	0	0	0	0	0	0	0	0	0	0.3	0.09333	0.2

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}	x_{57}	RI
0.0295	0	0	0.0867	0	0	0	0.2	0.477	0	0	0	0	0.1	0	0	0	2	1.4219

Figure 2. Numerical representation of ethyl–isopropenyl–phenyl phosphate and 1,1–diethyl–2,2–dipropyl diphosphate.

2.3 Computer Software

All computations were performed on IBM PC–type microcomputer, running under the MS–Windows'98 operating system. The artificial neural networks computations were carried out with the network simulation program Statistica Neural Networks (SNN) [62]. Data manipulation and interpretation of the obtained results was carried out with Microsoft Excel v. 97.

2.4 Neural Networks

In this study, the linear and multilayer, feedforward networks were applied. 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 and the temperature at which the refractive index was obtained. The output layer has one node generating the estimated value of the investigated property. The input and output values were linearly scaled between 0 and 1 by the standard *Minimax* conversion function available in the SNN program.

Because the learning and approximation occurs mainly in the hidden layer, the number of hidden neurons needs to be sufficient to ensure that the information contained in the data utilized for the network training 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 the hidden layer were considered. The starting networks architectures were determined by applying an automatic optimization procedure available in Statistica Neural Network v 4.0 programs package, named Intelligent Problem Solver (IPS) [62]. The IPS program was forced to search optimal networks according to stated above limits. The sigmoid squashing function was applied for the processing neurons in hidden layers and the linear one in an output neuron. The candidate network architecture 40:2:1 (with the best performance characteristics) was retained for further learning and testing its predictive ability. The dimensionality of the input layer in the best network corresponds to the number of descriptors having non–zero values for all compounds. These descriptors were chosen as valid input variables. The network was preliminary trained for a period of 50 epochs by standard back propagation procedure and then the conjugated gradient algorithm was used over a dozen learning cycles.

The final attempt for improvement of the QSPR model was carried out by replacing the linear activation function with a sigmoid one in the output neuron. The parallel learning of both networks over a period of about 300 epochs gave a slight improvement of predictions obtained with the model with sigmoid function in the output neuron.

2.5 Reduction of Structural Descriptors

The next experiment on the phosphates and diphosphates 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 a sensitivity analysis (SA), available as a standard procedure in Statistica Neural Network program package. To perform the selection of variables a new set of five 40:2:1 networks was trained using IPS procedure and applying random subdivision of the entire set of examples (in proportion 4:1) into only training and verification sets. 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 phosphates were removed. The second half of input variables (considered as important for refractive indexes prediction) was retained for further processing. These highly active variables for the phosphates are: 1, 2, 9, 10, 11, 12, 24, 29, 34, 35, 36, 38, 39, 40, 43, 44, 45, 48, 53 and 57 (see Table 1). It should be noted that in the most cases, frequently inactive descriptors (*i.e.* equal to zero) were discarded.

The reduced data sets containing a 20–component vector of numerical values for phosphates and diphosphates were used for the final selection of the optimal network, which was performed applying the IPS procedure once again. Because the collected sets of examples are relatively small according to the size of input vectors, the cases were randomly reassigned only to training and

cross–validation sets in a 4:1 proportion. The best network (20:2:1) from the preliminary optimized by IPS automatic procedure (with the lower training and verification mean square errors values) was retained for further optimization. The searching of the best network procedure was repeated twice for the networks with successively reduced numbers of inputs by removing from selected pool of descriptors the variable having the lowest value of the sensitivity index in order: 39 and 43. The best networks with structures 19:2:1 and 18:2:1 have been retained for further optimization.

For the final optimisation the conjugated gradient algorithm was used applying leave–20%–out procedure. In this procedure one–fifth of the objects were selected out one after another, whereas for every selection the model was build up with remaining part of examples. Next, this model was used to predict the refractive index values for the selected compounds. Joined results of the refractive indexes estimation gave information on the prediction ability and on the model quality for the selected training and prediction sets. Each time the training results were saved when the root mean square error averaged over the training set had reached minimum value. Depending on particular network structure and the training set, this occurred after about 120 to 200 epochs.

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 the training and cross–validation set examples. Training was stopped (before the training error has reached the above mentioned value) when the *RMS* error obtained for the control data was the lowest. The linear network (with the structure 40:1) generated by IPS, as a least squares linear model [62], has been retained for comparison purposes. This and the multilayer final network structures are specified in Tables 2 and 3.

2.6 Statistical Parameters

When the optimisation process of all investigated models was completed, the output data obtained for both sets of examples has been stored in the Excel data form for further work out. In the next step of this investigation, the generated ANN models were evaluated. The statistical quality of the ANN modeling results for both the training and cross–validation sets was evaluated using the following parameters: squared correlation coefficient R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^c)^2}{\sum_{i=1}^{n}(y_i - y_0)^2} \tag{6}$$

average recognition and prediction errors *AER*:

$$AER = \frac{1}{n}\sum_{i=1}^{n}(y_i^c - y_i) \tag{7}$$

average absolute error *AAE*:

$$AAE = \frac{1}{n}\sum \left| y_i^c - y_i \right| \tag{8}$$

and errors standard deviation *SD*

$$SD = \sqrt{\frac{n\sum_{i=1}^{n}(y_i^c - y_i)^2 - \left[\sum_{i=1}^{n}(y_i^c - y_i)\right]^2}{n^2}} \tag{9}$$

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

3 RESULTS AND DISCUSSION

The main objective of this work was to verify the utility of previously elaborated structural descriptors, completed with the newly proposed index describing cyclic esters, for the prediction of the refractive index of phosphorus containing organic compounds. The structures of investigated phosphates and diphosphates are stored as ISIS Draw files deposited in supplementary material, in the archive file Phosphate.zip.

Numerical representations (vectors of structural indices values) obtained in the coding phase of this work, are collected in the file Phosphate_D.txt, also deposited in the supplement to this article. The statistical results of the ANN modeling of the phosphates and diphosphates refractive indexes obtained with different models are listed in Tables 2 and 3.

Table 2. Statistics of the Refractive Index Predictions with Linear Neural Network (LNN) Model

Statistics	*AER*	*AAE*	*SD*	R^2
Training set	0.0001	0.0333	0.0084	0.9680
Cross–validation	0.0001	0.0087	0.0118	0.9430

A linear regression of the refractive index against 40 structural descriptors characterizing phosphates molecules, using the linear network, is summarized by the respective statistics in the Table 2. The linear model performance characteristics obtained for the training set examples are clearly better then those obtained for prediction results, as can be seen from the higher R^2 = 0.968 versus R^2_{cv} = 0.943 as well as lower standard deviation 0.0084 *vs*. 0.0118.

Table 3. Statistics of Two–Layers Neural Networks Models for Calculating Refractive Indexes of Phosphates, With Sigmoid Activation Function in Output Neuron (sfon) and Linear Output Neuron (lon)

Statistics	40:2:1 (sfon)		20:2:1 (lon)		19:2:1 (lon)		18:2:1 (lon)	
	training	cross–val	training	cross–val	training	cross–val	training	cross–val
AER	0.0004	0	0.0003	0.0001	–0.0007	0	0	0.0002
AAE	0.0029	0.0033	0.0032	0.0036	0.0033	0.0037	0.0033	0.0038
SD	0.0040	0.0044	0.0043	0.0048	0.0045	0.0050	0.0047	0.0052
R^2	0.9926	0.9908	0.9912	0.9889	0.9906	0.9882	0.9894	0.9870

The statistical results of the multilayer ANN modeling listed in Table 3 have shown the significant improvement that has been obtained for refractive indexes prediction using nonlinear models comparing with the linear network. The comparison of the statistical parameters obtained for the refractive index calibration and prediction, using the multilayer neural networks, reveals the superiority of nonlinear models. Both calibration refractive index values and predicted during the cross–validation procedure gave better statistical parameters than those obtained with the linear network. The best neural model (40:2:1) gives the most accurate predictions with average absolute errors of 0.0029 and 0.0044 as well as the highest R^2 coefficients equal to 0.9926 and R^2_{cv} 0.9908 for the training and cross–validation phase, respectively. Predictions obtained for the investigated compounds with the model based on twice reduced pool of structural descriptors (network 20:2:1) are quite precise: *AAE* 0.0032, R^2 0.9912 and AAE_{cv} 0.0036, R^2_{cv} 0.9889. The predicted refractive index values along with the deviation from experimental values (using the 20:2:1 NN model) are listed in Table 4.

Table 4. Experimental and Calculated Refractive Index for 180 Organic Phosphates and Diphosphates

No	Comp Id	Phosphates	Exp *n*	Calc *n*	Δ*n*
1	Ph_1	monomethyl	1.4200 [e]	1.4127	–0.0073
2	Ph_2	monoethyl	1.4270 [e]	1.4155	–0.0115
3	Ph_3	dimethyl	1.4080 [e]	1.4122	0.0042
4	Ph_3	dimethyl	1.4070 [a]	1.4136	0.0066
5	Ph_4	trimethyl	1.3950 [e]	1.3968	0.0018
6	Ph_4	trimethyl	1.3964 [a]	1.3983	0.0019
7	Ph_5	ethyl methyl	1.4115 [a]	1.4180	0.0065
8	Ph_6	methyl propane–1,2–diyl	1.4250 [a]	1.4263	0.0013
9	Ph_7	ethyl dimethyl	1.4015 [a]	1.4007	–0.0008
10	Ph_7	ethyl dimethyl	1.3984 [c]	1.3997	0.0013
11	Ph_8	diethyl	1.4170 [a]	1.4194	0.0024
12	Ph_9	tetramethyl diphosphate	1.4121 [e]	1.4131	0.0010
13	Ph_9	tetramethyl diphosphate	1.4136 [a]	1.4144	0.0008
14	Ph_10	ethyl propane–1,2–diyl	1.4265 [a]	1.4260	–0.0005
15	Ph_11	isopropenyl dimethyl	1.4165 [a]	1.4192	0.0027
16	Ph_12	mono(3–methylbutyl)	1.4150 [a]	1.4233	0.0083
17	Ph_13	ethyl trimethyl diphosphate	1.4150 [a]	1.4140	–0.0010
18	Ph_14	trivinyl	1.4289 [a]	1.4301	0.0012
19	Ph_15	propyl propane–1,2–diyl	1.4290 [a]	1.4293	0.0003
20	Ph_16	triethyl	1.4043 [e]	1.4034	–0.0009
21	Ph_16	triethyl	1.4053 [a]	1.4048	–0.0005
22	Ph_17	dipropyl	1.4251 [a]	1.4250	–0.0001
23	Ph_18	1,2–diethyl 1,2–dimethyl diphosphate	1.4170 [a]	1.4145	–0.0025
24	Ph_19	1,1–diethyl 2,2–dimethyl diphosphate	1.4156 [e]	1.4143	–0.0013
25	Ph_19	1,1–diethyl 2,2–dimethyl diphosphate	1.4160 [a]	1.4155	–0.0005
26	Ph_20	allyl diethyl	1.4216 [e]	1.4277	0.0061
27	Ph_21	butyl propane–1,2–diyl	1.4312 [a]	1.4324	0.0012
28	Ph_22	isobutyl propane–1,2–diyl	1.4310 [a]	1.4307	–0.0003
29	Ph_23	diethyl isopropenyl	1.4158 [e]	1.4184	0.0026
30	Ph_23	diethyl isopropenyl	1.4200 [a]	1.4198	–0.0002
31	Ph_24	ethyl 2,2–dimethylpropanediyl	1.4435 [a]	1.4426	–0.0009
32	Ph_24	ethyl 2,2–dimethylpropanediyl	1.4388 [e]	1.4413	0.0025
33	Ph_25	ethyl 3–methylbutyl	1.4210 [e]	1.4258	0.0048

Table 4. (Continued)

No	Comp Id	Phosphates	Exp *n*	Calc *n*	Δ*n*
34	Ph_26	diethyl isopropyl	1.4038 [e]	1.4053	0.0015
35	Ph_27	dimethyl pentyl	1.4127 [e]	1.4051	–0.0076
36	Ph_28	triethyl methyl	1.4180 [a]	1.4225	0.0045
37	Ph_29	dimethyl phenyl	1.4887 [a]	1.4876	–0.0011
38	Ph_30	diallyl ethyl	1.4350 [a]	1.4376	0.0026
39	Ph_31	diethyl 1–methylenepropyl	1.4270 [a]	1.4247	–0.0023
40	Ph_32	cyclohexyl dimethyl	1.4417 [e]	1.4414	–0.0003
41	Ph_33	diethyl 1–methylpropenyl	1.4274 [a]	1.4240	–0.0034
42	Ph_34	diisobutyl	1.4320 [a]	1.4265	–0.0055
43	Ph_35	butyl diethyl	1.4110 [e]	1.4125	0.0015
44	Ph_35	butyl diethyl	1.4131 [a]	1.4138	0.0007
45	Ph_36	dibutyl	1.4337 [a]	1.4298	–0.0039
46	Ph_37	ethyl diisopropyl	1.4044 [e]	1.4062	0.0018
47	Ph_38	hexyl dimethyl	1.4180 [e]	1.4076	–0.0104
48	Ph_39	diethyl isobutyl	1.4074 [e]	1.4097	0.0023
49	Ph_40	tetraethyl diphosphate	1.4170 [e]	1.4182	0.0012
50	Ph_40	tetraethyl diphosphate	1.4222 [a]	1.4194	–0.0028
51	Ph_41	1,1–diisopropyl 2,2–dimethyl diphosphate	1.4165 [e]	1.4170	0.0005
52	Ph_42	1,1–dimethyl 2,2–dipropyl diphosphate	1.4199 [e]	1.4167	–0.0032
53	Ph_43	methylethanediyl phenyl	1.5068 [a]	1.5046	–0.0022
54	Ph_44	dimethyl m–tolyl	1.4910 [a]	1.4865	–0.0045
55	Ph_45	dimethyl p–tolyl	1.4898 [a]	1.4858	–0.0040
56	Ph_46	triallyl	1.4435 [e]	1.4401	–0.0034
57	Ph_46	triallyl	1.4500 [a]	1.4420	–0.0080
58	Ph_47	triisopropyl	1.4069 [a]	1.4101	0.0032
59	Ph_48	tripropyl	1.4136 [e]	1.4113	–0.0023
60	Ph_48	tripropyl	1.4165 [a]	1.4131	–0.0034
61	Ph_49	diethyl pentyl	1.4152 [e]	1.4156	0.0004
62	Ph_50	butane–1,3–diyl phenyl	1.5163 [e]	1.5085	–0.0078
63	Ph_51	diethyl phenyl	1.4773 [e]	1.4825	0.0052
64	Ph_51	diethyl phenyl	1.4761 [a]	1.4844	0.0083
65	Ph_52	3,5–dimethyl–phenyl dimethyl	1.4946 [a]	1.4875	–0.0071
66	Ph_53	ethyl bis(2–methylallyl)	1.4390	1.4379	–0.0011
67	Ph_54	bis(3–methylbutyl)	1.4375 [a]	1.4319	–0.0056
68	Ph_55	ethyl dibutyl	1.4168 [e]	1.4181	0.0013
69	Ph_56	dipentyl	1.4395 [a]	1.4349	–0.0046
70	Ph_57	dimethyl octyl	1.4236	1.4135	–0.0101
71	Ph_58	1,1–diethyl 2,2–dipropyl diphosphate	1.4212 [e]	1.4203	–0.0009
72	Ph_58	1,1–diethyl 2,2–dipropyl diphosphate	1.4219 [a]	1.4222	0.0003
73	Ph_59	1,1–diethyl 2,2–diisopropyl diphosphate	1.4175 [e]	1.4218	0.0043
74	Ph_60	ethyl isopropenyl phenyl	1.4845 [a]	1.4896	0.0051
75	Ph_61	diethyl m–tolyl	1.4814 [a]	1.4837	0.0023
76	Ph_62	diethyl o–tolyl	1.4812 [a]	1.4837	0.0025
77	Ph_63	dibutyl isopropenyl	1.4268 [a]	1.4359	0.0091
78	Ph_64	diisobutyl isopropenyl	1.4245 [a]	1.4267	0.0022
79	Ph_65	diallyl phenyl	1.4965 [e]	1.4874	–0.0091
80	Ph_66	diisopropyl phenyl	1.4684 [e]	1.4790	0.0106
81	Ph_67	diethyl 1–phenylethyl	1.4870 [a]	1.4835	–0.0035
82	Ph_68	tris(2–methylallyl)	1.4454 [e]	1.4444	–0.0010
83	Ph_69	tributyl	1.4220 [e]	1.4208	–0.0012
84	Ph_69	tributyl	1.4249 [a]	1.4217	–0.0032
85	Ph_70	triisobutyl	1.4173 [e]	1.4166	–0.0007
86	Ph_70	triisobutyl	1.4190 [a]	1.4175	–0.0015
87	Ph_71	diethyl octyl	1.4210 [e]	1.4247	0.0037
88	Ph_72	tetraisopropyl diphosphate	1.4170 [e]	1.4230	0.0060
89	Ph_72	tetraisopropyl diphosphate	1.4200 [a]	1.4248	0.0048

Table 4. (Continued)

No	Comp Id	Phosphates	Exp *n*	Calc *n*	Δ*n*
90	Ph_73	1,1–diethyl 2,2–dibutyl diphosphate	1.4245 [e]	1.4231	–0.0014
91	Ph_74	1,1–diisopropyl 2,2–dipropyl diphosphate	1.4210 [e]	1.4228	0.0018
92	Ph_75	tetrapropyl diphosphate	1.4248 [e]	1.4242	–0.0006
93	Ph_76	methyl diphenyl	1.5373 [e]	1.5303	–0.0070
94	Ph_76	methyl diphenyl	1.5320 [b]	1.5325	0.0005
95	Ph_77	butyl isopropenyl phenyl	1.4825 [a]	1.4873	0.0048
96	Ph_78	dipropyl m–tolyl	1.4779 [a]	1.4787	0.0008
97	Ph_79	diethyl 4–isopropylphenyl	1.4770 [a]	1.4809	0.0039
98	Ph_80	2–ethyl–4–methylpentyl 2,2–dimethylpropanediyl	1.4485 [a]	1.4496	0.0011
99	Ph_81	ethyl diphenyl	1.5318 [e]	1.5261	–0.0057
100	Ph_82	2–ethylhexane–1,3–diyl phenyl	1.5017 [e]	1.5114	0.0097
101	Ph_83	dibutyl phenyl	1.4689 [e]	1.4769	0.0080
102	Ph_83	dibutyl phenyl	1.4736 [a]	1.4777	0.0041
103	Ph_84	diethyl 3–tert–butylphenyl	1.4770 [e]	1.4790	0.0020
104	Ph_85	dibutyl hexyl	1.4263 [e]	1.4282	0.0019
105	Ph_86	diethyl decyl	1.4266 [e]	1.4297	0.0031
106	Ph_87	1,1–dibutyl 2,2–diisopropyl diphosphate	1.4235 [e]	1.4260	0.0025
107	Ph_88	isopropenyl diphenyl	1.5483 [a]	1.5385	–0.0098
108	Ph_89	allyl diphenyl	1.5214 [e]	1.5314	0.0100
109	Ph_90	diphenyl propyl	1.5249 [e]	1.5226	–0.0023
110	Ph_90	diphenyl propyl	1.5246 [a]	1.5255	0.0009
111	Ph_91	dibenzyl methyl	1.5308 [e]	1.5273	–0.0035
112	Ph_92	2–ethylhexyl methyl phenyl	1.4802 [e]	1.4785	–0.0017
113	Ph_93	4–(1,1–dimethylpropyl)phenyl diethyl	1.4838 [a]	1.4791	–0.0047
114	Ph_94	tripentyl	1.4320 [a]	1.4289	–0.0031
115	Ph_95	2–methylallyl diphenyl	1.5240 [a]	1.5331	0.0091
116	Ph_96	butyl diphenyl	1.5190 [e]	1.5190	0
117	Ph_97	isobutyl diphenyl	1.5188 [e]	1.5225	0.0037
118	Ph_98	dibenzyl ethyl	1.5285 [e]	1.5264	–0.0021
119	Ph_99	dipentyl phenyl	1.4715 [c]	1.4746	0.0031
120	Ph_100	dibutyl 1–phenylethyl	1.4756 [a]	1.4826	0.0070
121	Ph_101	dimethyl 4–(1,1,3,3–tetramethylbutyl) phenyl	1.4912 [a]	1.4819	–0.0093
122	Ph_102	bis(2–ethylhexyl)	1.4448 [a]	1.4384	–0.0064
123	Ph_102	bis(2–ethylhexyl)	1.4430 [e]	1.4373	–0.0057
124	Ph_103	dibutyl octyl	1.4296 [e]	1.4327	0.0031
125	Ph_104	tetrabutyl diphosphate	1.4296 [e]	1.4302	0.0006
126	Ph_105	pentyl diphenyl	1.5192 [e]	1.5157	–0.0035
127	Ph_106	2–methylbutyl diphenyl	1.5197 [e]	1.5176	–0.0021
128	Ph_107	3–methylbutyl diphenyl	1.5164 [e]	1.5189	0.0025
129	Ph_108	2,2–dimethylpropyl diphenyl	1.5132 [e]	1.5208	0.0076
130	Ph_109	diethyl 4–(1,1–dimethylpentyl)–phenyl	1.4812 [a]	1.4778	–0.0034
131	Ph_110	diethyl 4–heptylphenyl	1.4761 [a]	1.4778	0.0017
132	Ph_111	methyl dioctyl	1.4362 [e]	1.4369	0.0007
133	Ph_112	hexyl diphenyl	1.5131 [e]	1.5134	0.0003
134	Ph_113	2–methylpentyl diphenyl	1.5130 [e]	1.5147	0.0017
135	Ph_114	2–ethylbutyl diphenyl	1.5152 [e]	1.5165	0.0013
136	Ph_115	2,2–dimethylbutyl diphenyl	1.5118 [e]	1.5162	0.0044
137	Ph_116	dibenzyl butyl	1.5233 [e]	1.5203	–0.0030
138	Ph_117	butyl di–m–tolyl	1.5170 [a]	1.5188	0.0018
139	Ph_118	butyl octyl phenyl	1.4691 [e]	1.4726	0.0035
140	Ph_119	2–ethylhexyl butyl phenyl	1.4698 [e]	1.4727	0.0029
141	Ph_120	2–ethylhexyl 1–methylpropyl phenyl	1.4783 [e]	1.4710	–0.0073
142	Ph_121	2–ethylhexyl 2–methylpropyl phenyl	1.4720 [e]	1.4707	–0.0013
143	Ph_122	dibutyl decyl	1.4329 [e]	1.4361	0.0032
144	Ph_123	trihexyl	1.4340 [e]	1.4341	0.0001
145	Ph_124	heptyl diphenyl	1.5086 [e]	1.5098	0.0012

Table 4. (Continued)

No	Comp Id	Phosphates	Exp *n*	Calc *n*	Δ*n*
146	Ph_125	3–methylbutyl di–m–tolyl	1.5140 [a]	1.5176	0.0036
147	Ph_126	diethyl 4–nonylphenyl	1.4765 [a]	1.4770	0.0005
148	Ph_127	bis(2–methylallyl) 2–biphenyl	1.5331 [e]	1.5326	–0.0005
149	Ph_128	octyl diphenyl	1.5070 [e]	1.5083	0.0013
150	Ph_128	octyl diphenyl	1.5072 [a]	1.5098	0.0026
151	Ph_129	6–methylheptyl diphenyl	1.5076 [e]	1.5084	0.0008
152	Ph_130	2–ethylhexyl diphenyl	1.5080 [e]	1.5097	0.0017
153	Ph_131	2–ethylbutyl di–m–tolyl	1.5170 [a]	1.5147	–0.0023
154	Ph_132	butyl bis(3,5–dimethylphenyl)	1.5160 [a]	1.5147	–0.0013
155	Ph_133	1,1–diethyl 2,2–bis(2–ethylhexyl) diphosphate	1.4390 [e]	1.4345	–0.0045
156	Ph_134	2–allylphenyl diphenyl	1.5640 [e]	1.5616	–0.0024
157	Ph_135	tri–o–tolyl	1.5587 [e]	1.5572	–0.0015
158	Ph_135	tri–o–tolyl	1.5575 [a]	1.5595	0.0020
159	Ph_136	tri–m–tolyl	1.5553 [e]	1.5572	0.0019
160	Ph_137	nonyl diphenyl	1.5050 [e]	1.5049	–0.0001
161	Ph_138	diphenyl 3,5,5–trimethylhexyl	1.5057 [e]	1.5088	0.0031
162	Ph_139	2–ethylhexyl phenyl p–tolyl	1.5082 [e]	1.5080	–0.0002
163	Ph_140	bis(3,5–dimethylphenyl 3–methylbutyl	1.5140 [a]	1.5137	–0.0003
164	Ph_141	4–tert–butylphenyl diphenyl	1.5522 [e]	1.5523	0.0001
165	Ph_142	decyl diphenyl	1.5022 [e]	1.5019	–0.0003
166	Ph_143	2–butylhexyl diphenyl	1.5069 [e]	1.5061	–0.0008
167	Ph_144	octyl di–m–tolyl	1.5120 [a]	1.5044	–0.0076
168	Ph_145	2–ethylhexyl 1–methylheptyl phenyl	1.4687 [e]	1.4688	0.0001
169	Ph_146	bis(2–ethylhexyl) phenyl	1.4682 [e]	1.4690	0.0008
170	Ph_146	bis(2–ethylhexyl) phenyl	1.4750 [a]	1.4705	–0.0045
171	Ph_147	diethyl 4–dodecylphenyl	1.4750 [a]	1.4739	–0.0011
172	Ph_148	2,6–diallylphenyl diphenyl	1.5637 [e]	1.5588	–0.0049
173	Ph_149	bis(2–allylphenyl) phenyl	1.5422 [a]	1.5619	0.0197
174	Ph_150	tris[(R)–1–phenylethyl]	1.5498 [d]	1.5454	–0.0044
175	Ph_151	triphenetyl	1.5669 [e]	1.5592	–0.0077
176	Ph_152	dodecyl diphenyl	1.4987 [e]	1.4966	–0.0021
177	Ph_152	dodecyl diphenyl	1.5030 [a]	1.4985	–0.0045
178	Ph_153	2–butyloctyl diphenyl	1.4996 [e]	1.5002	0.0006
179	Ph_154	bis(3,5–dimethylphenyl) octyl	1.5110 [a]	1.5010	–0.0100
180	Ph_155	4–ethyl–1–isobutyloctyl butyl phenyl	1.4588 [e]	1.4682	0.0094
181	Ph_156	trioctyl	1.4403 [e]	1.4430	0.0027
182	Ph_157	tris(2–ethylhexyl)	1.4414 [a]	1.4403	–0.0011
183	Ph_158	tris(2,4,4–trimethylpentyl)	1.4395 [a]	1.4381	–0.0014
184	Ph_159	hexadecyl isopropyl phenyl	1.4633 [e]	1.4657	0.0024
185	Ph_160	bis[2–(2–methylallyl)–phenyl] phenyl	1.5647 [e]	1.5557	–0.0090
186	Ph_161	tris(3–phenylpropyl)	1.5404 [d]	1.5599	0.0195
187	Ph_162	trinonyl	1.4485 [e]	1.4433	–0.0052
188	Ph_163	tris(3,5,5–trimethylhexyl)	1.4420 [e]	1.4400	–0.0020
189	Ph_164	hexadecyl diphenyl	1.4934 [e]	1.4854	–0.008
190	Ph_165	2–allylphenyl bis(4–tert–butylphenyl)	1.5421 [e]	1.5488	0.0067
191	Ph_166	bis(2–allylphenyl) 2–biphenyl	1.5872 [e]	1.5792	–0.0080
192	Ph_167	tridecyl	1.4452 [e]	1.4488	0.0036
193	Ph_168	2–ethylhexyl 2–ethylhexane–1,3–diyl	1.4490 [e]	1.4512	0.0022
194	Ph_169	2–ethyl–2–butylpropanediyl phenyl	1.4998 [e]	1.5030	0.0032
195	Ph_170	isopropyl 2–methyl–2–propylpropanediyl	1.4447 [e]	1.4433	–0.0014
196	Ph_171	diethyl p–tolyl	1.4780 [a]	1.4837	0.0057
197	Ph_172	diethyl 3,5–dimethyl–phenyl	1.4830 [a]	1.4827	–0.0003
198	Ph_173	diethyl dodecyl	1.4335 [e]	1.4340	0.0005
199	Ph_174	ethyl 2–methyl–2–propylpropanediyl	1.4457 [e]	1.4429	–0.0028
200	Ph_175	2–ethylhexyl propane–1,2–diyl	1.4400 [e]	1.4384	–0.0016
201	Ph_176	1,1–diethyl diphosphate	1.4370 [e]	1.4257	–0.0113

Table 4. (Continued)

No	Comp Id	Phosphates	Exp *n*	Calc *n*	Δ*n*
202	Ph_177	1,2–diisopropyl diphosphate	1.4330 [e]	1.4261	–0.0069
203	Ph_178	1,2–dimethyl diphosphate	1.4250 [e]	1.4353	0.0103
204	Ph_179	1,2–dibutyl diphosphate	1.4310 [e]	1.4387	0.0077
205	Ph_180	phenyl diundecyl	1.4671 [a]	1.4626	–0.0045

[a, b, c, d, e] Temperatures of RI measurement: [a] 20, [b] 21, [c] 22, [d] 24, [e] 25 ºC.

The linear plot of predicted versus observed refractive indexes for the phosphates cross–validation examples is given in Figure 3. The distribution of points along the regression line is quite good and no extreme outliers are seen. The obtained predictions fit well to the experimental data with the high correlation coefficient of R_{cv} = 0.9944. The calculated parameters for the regression equations (Figure 3) shows a slope equal to 0.993 and an intersect of 0.01. The distribution for the prediction errors (*PER*) for phosphates and diphosphates (specified in Table 4) is presented in Figure 4. The error for each tested compound was calculated as $PER = RI_{pr} - RI_{exp}$ where RI_{pr} is the estimated refractive index and RI_{exp} is the experimental value.

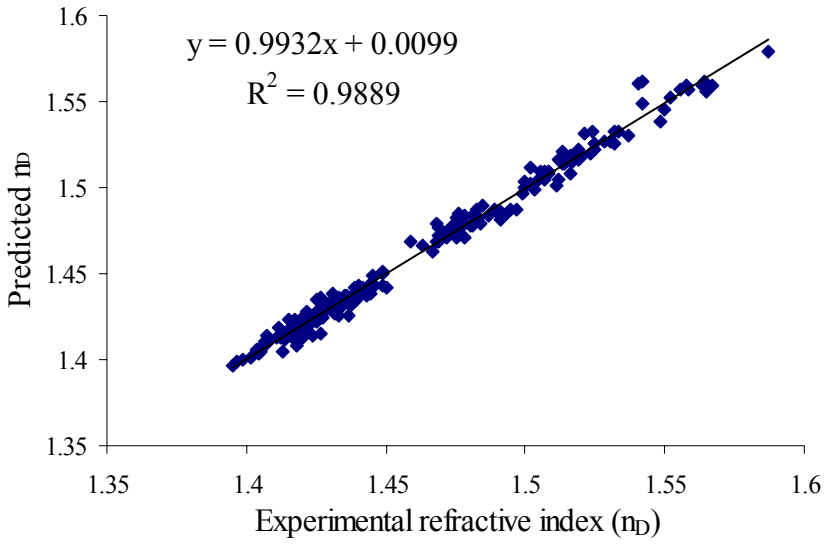


Figure 3. Predicted refractive indexes of phosphates versus experimental data.

The examination of the prediction error plot (beginning from compound number 1) indicates that refractive indexes values for compounds with lower molecular weight (aliphatic phosphates containing 1–8 carbon atoms) are often overestimated, although the observed prediction errors are relatively small. The next range spans examples numbers 40 until 100 and characterize compounds with 9 to 16 carbon atoms, most of diphospates and phosphates with mixed types of substituents: aliphatic, alicyclic and aromatic. The prediction errors for this group of compounds are distinctly higher and extending between –0.01 and 0.01. The following range of the error plot shows the second region of higher prediction accuracy. These predictions were obtained for examples

numbered 112–145 and representing compounds containing 18–22 carbon atoms, generally containing the phenyl and phenyl substituted phosphates. The last part of the plot represents prediction errors obtained for phosphates with 24–30 carbon atoms and various types of the substituents: aromatic (benzyl, phenyl with side aliphatic chains), aliphatic (long and/or highly branched chains, etc.). This part of plot indicates the broader span of prediction errors.

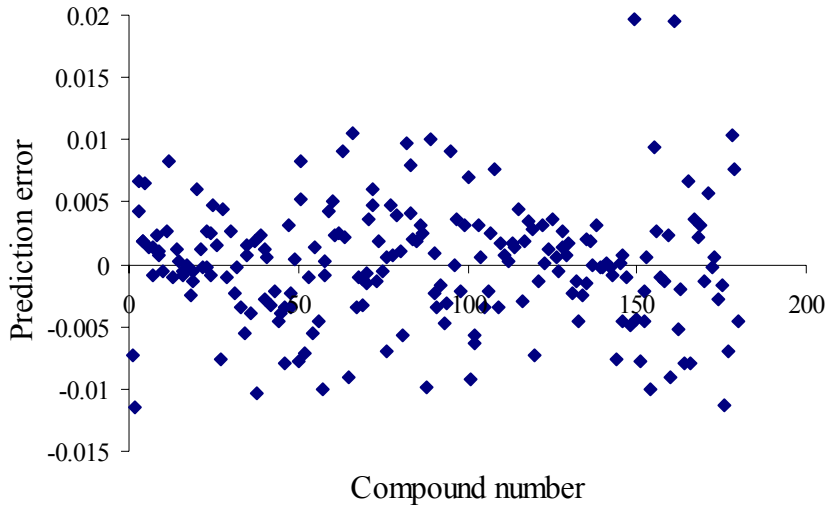


Figure 4. Distribution of the prediction error (*PER*) for the refractive indexes of 180 phosphates.

For the majority of investigated compounds the prediction of the refractive index of phosphates gives the error between ±0.012 with the mean error value of ±0.00357 although two outliers can be observed. The greatest overestimation of refractive index is obtained for the compounds (see Table 4) **149** phenyl–bis(2–allylphenyl) phosphate and **161** tris(3–phenylpropyl) phosphate. Both compounds were predicted with considerable errors (in the range of 0.0169–0.0197 for phosphate **149** and 0.0095–0.0195 for **161**) by the remaining ANN models. These compounds contain three phenyl groups with side chains or connected with phosphate group via three–membered aliphatic chain, that is, different than the large majority of investigated esters possessing phenylic groups in the molecule structure. In the case of compound **161** one can suggest the molecule symmetry as a possible cause of the prediction error, suggesting that compounds with symmetrically located substituents connected via oxygen with the phosphorus atom placed in the central part of a molecule need a more elaborated parameterization. The largest outlier is predicted with an error of 1.28%.

A wider examination of the prediction results for each of the tested networks has shown that some of substantial errors associated with investigated compounds occurred only for the above–discussed model. Others appeared among the test results, which had been obtained using the rest of the multilayer ANN models characterized in Table 3. This observation suggest that phosphates and diphosphates with symmetrically located constituents in relation to the center of molecule as well as the branched in neighborhood of functional group need more elaborated descriptors. Irrespective of

these outliers and the observed dispersion of prediction errors most of predictions (64%) are estimated with the inaccuracy less than the mean error value of ±0.0036. The performance of the model is satisfactory, with an average prediction error of 0.24%.

The next of models characterized in Table 3 and developed on the base of the network with 19 inputs presents slightly lower predictive ability then the model utilizing 20 variables described above. Adequate statistical characteristics both for the training and cross–validation phase (*AAE* 0.0033, R^2 0.9906 and AAE_{cv} 0.0037, R^2_{cv} 0.9882, respectively) show the greater inaccuracy of the obtained refractive index estimations.

The final network architecture (18:2:1) used for estimation of the refractive indexes values has the lowest predictive ability: *AAE* 0.0033, R^2 0.9894 and AAE_{cv} 0.0038, R^2_{cv} 0.987. It should be noticed that the predictions were obtained with more than twice reduced number of structural descriptors in comparison with the linear model. Although the number of adjustable parameters (weights of connections between neurons) is comparable with number of coefficients in the linear network, the obtained statistics show that the predictive power of nonlinear model remains better.

4 CONCLUSIONS

The main focus of this paper was to provide the evidence that the tested structural descriptors are useful and effective for QSPR modeling. They are representing particular structural descriptors that can be related to the refractive index of phosphorus compounds. The refractive index values of diverse phosphates and diphosphates, comprising various types of structures (aliphatic: normal and branched, cyclic: alicyclic and aromatic) for different temperatures in the range of 20–25 ºC have been successfully predicted using artificial neural networks. The prediction was possible solely on the basis of the molecular structure. The predicted values have an average error of 0.24% when compared with experimental values. The obtained model may be used with a high degree of confidence for practical prediction of the refractive index of organic phosphates.

Despite the demonstrated usefulness of elaborated descriptors, it should be emphasized that further investigations are necessary for selection of the most informative structural descriptors that enable a neural network to model structure–physical property relations for a wide group of compounds. The coding method requires a wide range of experiments in order to determine how the structural code behaves in modeling the physical properties of chemicals. The results of this work proved that a feed–forward, multilayer neural network can be easily trained to model the structure–properties relationship for the investigated group of organic compounds. These non–linear models can predict the refractive indexes of phosphates and diphosphates more accurately than the linear model.

Supplementary Material

The molecular files containing the structure geometry for all phosphates and diphosphates used in investigated QSPR models are deposited as an archive in the Phosphate.zip. Also numerical representations (vectors of indices values) obtained in the coding phase of this work, are collected in file: Phosphate_D.txt attached as a supplement to this article.

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. Katritzky, U. Maran, V. S. Lobanov, and M. Karelson, Structurally Diverse Quantitative Structure–Property Relationship Correlations of Technologically Relevant Physical Properties, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1–18.
[3] H. Hosoya, The Topological Index *Z* Before and After 1971, *Internet Electron. J. Mol. Des.* **2002**, *1*, 428–442, http://www.biochempress.com.
[4] M. Randić, Topological Indices; in: The *Encyclopedia of Computational Chemistry*, Eds. P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H. F. Schaefer III, and P. R. Schreiner, Wiley & Sons, London, 1998, p. 3018.
[5] N. Trinajstić, Chemical Graph Theory, 2nd revised ed., CRC Press, Boca Raton 1992, Chapter 4.
[6] O. Ivanciuc and J. Devillers, Algorithms and Software for the Computation of Topological Indices and Structure–Property Models; in: *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds. J. Devillers and A. T. Balaban, Gordon and Breach Science Publishers, Amsterdam, 1999, pp 779–804.
[7] M. Karelson, *Molecular Descriptors in QSAR/QSPR*, John Wiley & Sons, New York, 2000.
[8] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley – VCH, 2000.
[9] P. J. M. van Lannrhoven and E. H. L. Aarts, *Simulated Annealing: Theory and Applications*, Reidel, Dordrecht, 1987.
[10] D. B. Hibbert, Genetic Algorithms in Chemistry, *Chemom. Intell. Lab. Sys.* **1993**, *19*, *1*
[11] R. E. Aries, D. P. Lidiard, and R.A. Spragg, Principal Component Analysis, *Chem. Br.* 1991, pp. 821–824.
[12] N. R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley & Sons, New York, 1981.
[13] O. Strouf, *Chemical Pattern Recognition*, John Wiley & Sons, New York, 1986.
[14] R. H. Myers, *Classical and Modern Regression with Applications*, PWS–KENT Publishing Co, Boston, 1989.
[15] S. Wold, PLS for Multivariate Linear Modelling; in: *Chemometric Methods in Molecular Design*, Ed. H. van de Waterbeemd, VCH Publishers, Weinheim, Germany, 1995, pp.195–218.
[16] S. Wold and M. Sjöström, Chemometrics, Present and Future Success, *Chemometr. Intell. Lab. Sys.* **1998**, *44*, 3–14.
[17] 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.
[18] J. Zupan and J. Gasteiger, *Neural Networks for Chemists*, VCH Publishers, Weinheim, Germany, 1993.
[19] ADAPT, P. C. Jurs, 152 Davey Lab, Chemistry Dpt, Penn State University, University Park, PA 16802 USA, E–mail: pcj@psu.edu www http://zeus.chem.psu.edu/ADAPT.html.
[20] O. G. Mekenyan, S. H. Karabunarliev, J. M. Ivanov, and D. N. Dimitrov, A New Development of the OASIS Computer System for Modeling Molecular Properties. *Comput. Chem.* **1994**, *18*, 173–187.
[21] L. Tarko and O. Ivanciuc, QSAR Modeling of the Anticonvulsant Activity of Phenylacetanilides with PRECLAV (PRoperty Evaluation by CLAss Variables), *MATCH* (*Commun. Math. Comput. Chem.*) **2001**, *44*, 201–214.
[22] SciQSAR, SciVision, Inc., 200 Wheeler Road, Burlington, MA 01803, U.S.A., Phone: 1–781–272–4949, Fax: 1–781–272–6868, E–mail: scivision@delphi.com, www http://www.scivision.com.
[23] CODESSA 2.13, Semichem, 7204 Mullen, Shawnee, KS 66216, U.S.A., E–mail andy@semichem.com, http://www.semichem.com.
[24] CERIUS2, Accelrys Inc. 9685 Scranton Road San Diego, CA 92121–3752 USA, http://www.accelrys.com.
[25] DRAGON 1.11/2001, Milano Chemometrics and QSAR Research Group, University of Milano–Bicocca P.za d. Scienza, 1 – 20126 Milano, Italy, http://www.disat.unimib.it.
[26] A. T. Balaban (Ed.), *From Chemical Topology to Three–Dimensional Geometry*, Plenum, New York, 1997.
[27] A. R. Katritzky, *Understanding How Chemical Structure Determines Physical Properties*, http://ark2.chem.ufl.edu/research/qspr_2000/QSPR_files
[28] 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.
[29] 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.
[30] 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.
[31] 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.* **1994**, *34*, 1118–1121.
[32] 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.
[33] 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.
[34] 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.
[35] 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.
[36] T. Suzuki, R–U. Ebert, and G. Schüürmann, Development of Both Linear and Nonlinear Method to Predict the Liquid Viscosity at 20 ^{o}C of Organic Compounds, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1122–1128.
[37] O. Ivanciuc, The neural network MolNet prediction of alkane enthalpies, *Anal. Chim. Acta* **1999**, *384*, 271–284.
[38] 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.
[39] 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.
[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] 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.
[45] G. Espinoza, D. Yaffe, A. Arenas, Y. Cohen, and F. Giralt, A Fuzzy ARTMAP–Based Quantitative Structure–Property Relationship (QSPR) for Predicting Physical Properties of Organic Compounds, *Ind. Eng. Chem. Res.* **2001** *40*, 2757–2766.
[46] A. J. Chalk, B. Beck, and T. Clark, A Quantum Mechanical/Neural Net Model for Boiling Points with Error Estimation, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 457–462.
[47] T. Suzuki, R–U. Ebert, and G. Schüürmann, Application of Neural Networks to Modeling and Estimating Temperature–Dependent Liquid Viscosity of Organic Compounds, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 776–790.
[48] J. Kozioł, Neural Network Modeling of Physical Properties of Chemical Compounds, *Int. J. Quantum Chem.* **2001**, *84*, 117–126.
[49] 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.
[50] M. Vračko and J. Gasteiger, A QSAR Study on a Set of 105 Flavonoid Derivatives Using Descriptors Derived From 3D Structures, *Internet Electron. J. Mol. Des.* **2002**, *1*, 527–544, http://www.biochempress.com.
[51] A. I. Vogel, W. T. Cresswell, G.H. Jeffery, and J. Leicaster, Calculation of the Refractive Indices of Liquid Organic Compounds: Bond Molecular Refraction Coefficients, *Chem. Ind.* **1951**, *5*, 376–376.
[52] M. Huggins, Densities and Optical Properties of Organic Compounds in the Liquid State. VI. The Refractive Indices of Parafin Hydrocarbons and Some of Their Derivatives, *Bull. Chem. Soc. Jpn.* **1956**, *29*, 336–339.
[53] T. D. Le and J. G. Weers, Group Contribution Additivity and Quantum Mechanical Models for predicting the Molar Refractions, Indices of Refraction, and Boiling Points of Fluorochemicals, *J. Phys. Chem.* **1995**, *99*, 13909–13916.
[54] L. B. Kier and L. H. Hall, Molecular Connectivity in Structure–Activity Analysis; *Research Studies Press, Ltd.*, Letchworth, Hertfordshire, England, 1986, p 26.
[55] S. Liu, S. Cai, Ch. Cao, and Z. Li, Molecular Electronegative Distance Vector (MEDV) Related to 15 Properties

of Alkanes, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1337–1348.
[56] A. R. Katritzky, S. Sild, and M. Karelson, General Quantitative Structure–Property Relationship Treatment of the Refractive Index of Organic Compounds, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 840–844.
[57] G. P. Romanelli, L. R. F. Cafferata, and E. A. Castro, Ameliorate QSPR Study of Alkyl Hydroperoxides, *Rus. J. Gen. Chem.* **2001**, *71*, 257–260.
[58] R. Zhang, S. Liu, M. Liu, and Z. Hu, Neural Network–Molecular Descriptor Approach to the Prediction of Properties of Alkanes, *Comput. Chem.* **1997**, *21*, 335–341.
[59] S. Liu, R. Zhang, M. Liu, and Z. Hu, Neural Network–Topological Indices Approach to the Prediction of Properties of Alkanes, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1146–1151.
[60] O. Ivanciuc, Molecular Structure Encoding Into Artificial Neural Networks Topology. *Roum. Chem. Quart. Rev.* **2000**, *8*, 197–220.
[61] Beilstein Handbuch der Organishen Chemie, Vierter Auflange, Springer–Verlag, Berlin, 1958.
[62] 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.