Struct Chem (2014) 25:29–35
DOI 10.1007/s11224-013-0229-x

ORIGINAL RESEARCH

Spectral density distribution moments as novel descriptors for QSAR/QSPR

D. Bielińska-Wąż · P. Wąż · K. Jagiełło · T. Puzyn

Received: 10 November 2012 / Accepted: 1 February 2013 / Published online: 28 February 2013

Abstract We propose spectral density distribution moments as molecular descriptors. We apply the new descriptors for developing a QSPR model that predicts the logarithmic values of subcooled liquid vapor pressure. We consider the infrared spectra of chloronaphthalenes.

keywords Descriptors · Molecular similarity · Statistical spectroscopy

Introduction

Nowadays, the majority of environmental studies are focused on group of chemicals called persistent organic pollutants (POPs), which pose a vast range of threats to human health and natural ecosystems. Due to their high lipophilicity and resistance to naturally occurring degradation processes, they are prone to bioaccumulate in human and animal tissues and to biomagnify in food chains [1]. Moreover, after entering the organism, they can induce a variety of toxic effects, including cancer, allergies, and hypersensitivity, damage to the central and peripheral nervous systems, reproductive disorders, and disruption of the immune system [2, 3]. Therefore, according to the Stockholm Convention, the emission of POPs to the environment needs to be eliminated or reduced [1].

Typical representatives of POPs are chloronaphthalenes (CNs). This group includes of 75 congeners—chemicals based on the same skeleton (naphthalene), but differ by a number of chlorine atoms or by the substituted pattern [4]. Despite of the fact that the synthesis of CNs is formally abandoned, there are still some commercial products (i.e., insulating materials, rubber belts) containing CNs available [5]. Moreover, chloronahthalenes are released to the environment during thermal-related synthesis (i.e., industrial waste incineration as well as domestic heating) [6], which, in fact, is assumed to be currently the main source of CNs in the environment [7]. Since the emission of CNs to the atmosphere, estimated only for Europeans countries, is still high, equal to 1.03 tons per year [8], there is an urgent need to perform comprehensive risk studies of these pollutants.

Among main factors influencing the environmental behavior of CNs are: their overall environmental persistence, mobility, and (eco) toxicity. The first two characteristics cannot be measured directly. They are usually determined with employing multimedia mass-balance (MM) models. And every MM model requires a set of phys/chem parameters (e.g., partition coefficients, half-live times, enthalpies of phase transfer, vapor pressure, etc.) as input data. These parameters can be obtained empirically. However, high costs of experiments and time required for performing them for large arrays of chemicals motivate the scientific community to search for alternative, non-experimental, and ways of receiving the lacking parameters.

Nowadays, the significance of a computational technique known as quantitative structure–property relationships (QSPR) modeling and their various applications in chemical risk assessment have being highlighted by many international organizations and regulations (e.g., REACH in Europe) [9].

D. Bielińska-Wąż (✉)
Instytut Fizyki, Uniwersytet Mikołaja Kopernika,
Grudziądzka 5, 87-100 Toruń, Poland
e-mail: dsnake@fizyka.umk.pl

P. Wąż
Centrum Astronomii, Uniwersytet Mikołaja Kopernika,
Gagarina 11, 87-100 Toruń, Poland

K. Jagiełło · T. Puzyn
Pracownia Chemometrii Środowiska, Wydział Chemii,
Uniwersytet Gdański, Sobieskiego 18/19,
80-952 Gdańsk, Poland

This approach is based on the assumption that the phys/chem properties of chemical compounds are the functions of so-called molecular descriptors, representing structural features of the molecules. Thus, based on the experimental data available even for few compounds, it is possible to develop mathematical equation describing the correlation(s) between their molecular structures and properties and, on this basis, to predict the lacking information for other, structurally similar molecules [10].

There are many examples of successful applications of the QSPR approach for predicting environmentally relevant properties of CNs [7, 11, 12]. However, still there is a need of searching for novel structural descriptors that more appropriately would express molecular variance in particular groups of structurally similar congeners of POPs.

Intensity distribution moments, recently proposed by us as new molecular descriptors [13–15], proved to be an efficient tool in the identification of specific groups of molecules. For example, using these descriptors, one could distinguish nitriles from amides [16]. The general methodology used in this study, the statistical spectroscopy, is known in many different areas of science. The basic quantities, the distribution moments, may be derived from atomic or molecular spectra. Similar methods of statistical spectroscopy we have already applied in studies on stellar spectra [17, 18], in analyzing properties of chaotic dynamical systems [19], and in bioinformatics [20]. Now, we continue the investigation on the usefulness of different kinds of moments as molecular descriptors. This time we check the spectral density distribution moments. In the present study, the moments are obtained from the frequencies (rather than from the intensities, as it was done before [16]) of the infrared (IR) spectra of CNs. However, the statistical distributions may also be created from any function describing the system under consideration. The new descriptors are applied for developing a QSPR model that predicts the logarithmic values of subcooled liquid vapor pressure at 25 °C.

Theory

Let us consider a discrete frequency spectrum $\nu_1, \nu_2, \ldots \nu_D$ treated as a statistical ensemble. The density of the frequencies (spectral density distribution) is defined as:

$$\rho(\nu) = \frac{1}{D}\sum_{i=1}^{D}\delta(\nu - \nu_i). \tag{1}$$

Convenient characteristics of distributions are their moments.

The q-th moment of $\rho(\nu)$ is defined as

$$M_{\rho,q} = \int_{-\infty}^{\infty} \nu^q \rho(\nu) d\nu. \tag{2}$$

Using Eq. 1 we get

$$M_{\rho,q} = \frac{1}{D}\sum_{i=1}^{D} \nu_i^q. \tag{3}$$

The corresponding q-th spectral density scaled moments are

$$M'_{\rho,q} = \frac{1}{D}\sum_{i=1}^{D}(\nu_i - M_{\rho,1})^q, \tag{4}$$

$$M''_{\rho,q} = \frac{1}{D}\sum_{i=1}^{D}\left[\frac{(\nu_i - M_{\rho,1})}{\sqrt{M_{\rho,2} - (M_{\rho,1})^2}}\right]^q. \tag{5}$$

In the present study, we construct spectral density distributions $\rho(\nu)$ from the frequencies ν_i of the IR spectra. The aim of this study is to introduce spectral density distribution moments $M_{\rho,q}$, $M'_{\rho,q}$, $M''_{\rho,q}$ as molecular descriptors. Usually, four lowest moments are used in statistical investigations ($M_{\rho,1}$ is the mean frequency, $M'_{\rho,2}$ describes the width, $M''_{\rho,3}$ - the asymmetry, and $M''_{\rho,4}$ - the excess of the distribution). Higher-order moments do not have any direct geometrical equivalents and usually are neglected. In most cases they supply no new information about the system [15]. However, in some cases, they may also be useful [16].

Results and discussion

We perform the calculations for 76 compounds: CNs containing from zero through eight chlorine atoms. They are listed in Table 1, where $r = 0, 1, \ldots 75$ are the labels of the compounds.

We study spectral density distributions of the frequencies of the IR spectra of the CNs.

The vibrational spectra we obtained from density functional theory (DFT) calculations. A hybrid B3LYP functional and 6-311++G** basis were used as implemented in the Gaussian 03 code [21].

Figure 1 shows the first moments $M_{\rho,1}$ for all 76 compounds. In this figure, one can recognize particular descriptors corresponding to particular compounds numbered by r (see Table 1) in the horizontal axis. The descriptors corresponding to the molecules with different number of chlorine atoms are represented by different symbols in the figure. The same symbols are used in Fig. 2, where $M'_{\rho,2}$ and $M''_{\rho,q}$ for $q = 3, 4, \ldots 10$ are shown. The patterns of $M''_{\rho,5}, M''_{\rho,6}, \ldots M''_{\rho,10}$ are similar to each other (Fig. 2). That suggests strong correlations between these moments. Therefore, the four lowest moments $M_{\rho,1}, M'_{\rho,2}, M''_{\rho,3}$, and $M''_{\rho,4}$ carry sufficient characteristics of the compounds.

Table 1 Experimental and predicted values of *log* P_L, leverage values and $M_{\rho,1}$ used as molecular descriptors (T-training set, V-validation set)

r	Substitution pattern	Splitting	Exp *log* P_L	Pred *log* P_L	Leverage	$M_{\rho,1}$
0	Naphthalene			1.05	0.41	1343.08
1	1-CN	T	0.57	0.39	0.28	1256.80
2	2-CN	T	0.58	0.38	0.28	1254.90
3	1,2-diCN	T	−0.48	−0.27	0.18	1169.15
4	1,3-diCN			−0.28	0.18	1168.64
5	1,4-diCN	V	−0.38	−0.27	0.18	1170.25
6	1,5-diCN			−0.27	0.18	1169.95
7	1,6-diCN			−0.27	0.18	1169.11
8	1,7-diCN			−0.28	0.18	1168.55
9	1,8-diCN			−0.29	0.18	1167.50
10	2,3-diCN			−0.29	0.18	1167.19
11	2,6-diCN			−0.29	0.18	1167.06
12	2,7-diCN			−0.29	0.18	1167.58
13	1,2,3-triCN	T	−1.19	−0.95	0.12	1080.51
14	1,2,4-triCN			−0.94	0.12	1082.01
15	1,2,5-triCN			−0.94	0.12	1081.52
16	1,2,6-triCN			−0.95	0.12	1080.53
17	1,2,7-triCN			−0.95	0.12	1080.24
18	1,2,8-triCN			−0.97	0.11	1077.50
19	1,3,5-triCN			−0.95	0.12	1080.88
20	1,3,6-triCN			−0.95	0.12	1080.13
21	1,3,7-triCN			−0.96	0.11	1079.46
22	1,3,8-triCN			−0.96	0.11	1078.51
23	1,4,5-triCN			−0.96	0.11	1079.25
24	1,4,6-triCN			−0.94	0.12	1081.52
25	1,6,7-triCN			−0.95	0.12	1080.60
26	2,3,6-triCN			−0.96	0.11	1079.29
27	1,2,3,4-tetraCN	T	−1.76	−1.62	0.09	992.27
28	1,2,3,5-tetraCN	T	−1.75	−1.62	0.09	992.02
29	1,2,3,6-tetraCN			−1.62	0.09	991.80
30	1,2,3,7-tetraCN			−1.63	0.09	991.45
31	1,2,3,8-tetraCN			−1.65	0.09	988.29
32	1,2,4,5-tetraCN			−1.63	0.09	990.65
33	1,2,4,6-tetraCN			−1.62	0.09	992.02
34	1,2,4,7-tetraCN	V	−1.61	−1.61	0.09	992.84
35	1,2,4,8-tetraCN			−1.65	0.09	988.73
36	1,2,5,6-tetraCN			−1.61	0.09	993.33
37	1,2,5,7-tetraCN			−1.62	0.09	992.36
38	1,2,5,8-tetraCN			−1.65	0.09	988.83
39	1,2,6,7-tetraCN			−1.62	0.09	991.71
40	1,2,6,8-tetraCN			−1.65	0.09	988.67
41	1,2,7,8-tetraCN			−1.65	0.09	988.27
42	1,3,5,7-tetraCN			−1.63	0.09	991.08
43	1,3,5,8-tetraCN			−1.63	0.09	990.18
44	1,3,6,7-tetraCN			−1.63	0.09	991.14
45	1,3,6,8-tetraCN			−1.64	0.09	989.60
46	1,4,5,8-tetraCN			−1.67	0.08	985.78
47	1,4,6,7-tetraCN			−1.61	0.09	992.90

Table 1 continued

r	Substitution pattern	Splitting	Exp $log\ P_L$	Pred $log\ P_L$	Leverage	$M_{\rho,1}$
48	2,3,6,7-tetraCN			−1.62	0.09	992.15
49	1,2,3,4,5-pentaCN			−2.32	0.09	899.81
50	1,2,3,4,6-pentaCN	T	−2.32	−2.30	0.09	902.47
51	1,2,3,5,6-pentaCN			−2.29	0.09	903.64
52	1,2,3,5,7-pentaCN	V	−2.23	−2.30	0.09	902.68
53	1,2,3,5,8-pentaCN	T	−2.48	−2.32	0.09	899.54
54	1,2,3,6,7-pentaCN			−2.29	0.09	903.39
55	1,2,3,6,8-pentaCN			−2.33	0.09	899.01
56	1,2,3,7,8-pentaCN			−2.33	0.09	898.78
57	1,2,4,5,6-pentaCN			−2.32	0.09	899.77
58	1,2,4,5,7-pentaCN			−2.31	0.09	901.00
59	1,2,4,5,8-pentaCN			−2.34	0.09	897.05
60	1,2,4,6,7-pentaCN			−2.29	0.09	903.57
61	1,2,4,6,8-pentaCN			−2.33	0.09	899.00
62	1,2,4,7,8-pentaCN			−2.32	0.09	899.89
63	1,2,3,4,5,6-hexaCN			−3.00	0.13	810.68
64	1,2,3,4,5,7-hexaCN			−3.00	0.13	810.68
65	1,2,3,4,5,8-hexaCN			−3.02	0.13	808.09
66	1,2,3,4,6,7-hexaCN	V	−2.92	−2.97	0.13	813.97
67	1,2,3,5,6,7-hexaCN	T	−2.92	−2.97	0.13	814.39
68	1,2,3,5,6,8-hexaCN			−3.00	0.13	810.68
69	1,2,3,5,7,8-hexaCN	T	−3.00	−3.00	0.13	810.26
70	1,2,3,6,7,8-hexaCN			−3.01	0.13	809.67
71	1,2,4,5,6,8-hexaCN	V	−3.02	−3.02	0.13	807.75
72	1,2,4,5,7,8-hexaCN			−3.02	0.13	808.03
73	1,2,3,4,5,6,7-heptaCN	T	−3.65	−3.68	0.21	721.19
74	1,2,3,4,5,6,8-heptaCN			−3.69	0.21	719.27
75	OctaCN	T	−4.25	−4.37	0.32	630.09

In the present study, the spectral density distribution moments are applied as molecular descriptors for developing a QSPR model of the logarithmic values of subcooled liquid vapor pressure ($log\ P_L$) at 25 °C. Experimental data, available for 17 CNs (22 % of the investigated group), have been taken from [22]. The compounds, for which the experimentally derived $log\ P_L$ values have been available, were divided into two smaller sets: a training set (12 compounds) and a validation set (5 compounds). The splitting algorithm was as follows. The 17 compounds have been sorted along with the decreasing $log\ P_L$ value and then every third compound was selected to the validation set, whereas the remaining ones formed the training set. This method produces two representative sets of compounds, since the compounds are evenly distributed along with the range of $log\ P_L$. The training set was then utilized for the model development and calibration, whereas the validation set, according to the golden standards [23] and the OECD recommendations for QSAR [24], was employed for performing external validation of the model.

Multiple linear regression (MLR)—a standard statistical method—was selected for modeling. But, regarding the limited number of training compounds, complexity of the model was restricted to using maximum one descriptor. Among all the descriptors taken into account (namely: $M_{\rho,1}, M'_{\rho,2}, M''_{\rho,3}, \ldots M''_{\rho,10}$) only one ($M_{\rho,1}$-spectral density distribution moment of the first order) has enabled the construction of a statistically significant ($p < 0.05$) QSPR model (Eq. 6):

$$logP_L = -9.1601 + 0.0076M_{\rho,1}. \quad (6)$$

The model was characterized not only by high goodness-of-fit (measured by the high correlation between the observed/experimental and predicted values of $log\ P_L$, $R^2 = 0.991$), but also by high robustness (expressed by the cross-validated correlation coefficient, $Q^2_{CV} = 0.985$) and—which is the most important—by high predictive ability (based on its external validation coefficient, $Q^2_{Ext} = 0.991$). The values of root mean square error (RMSE) of calibration

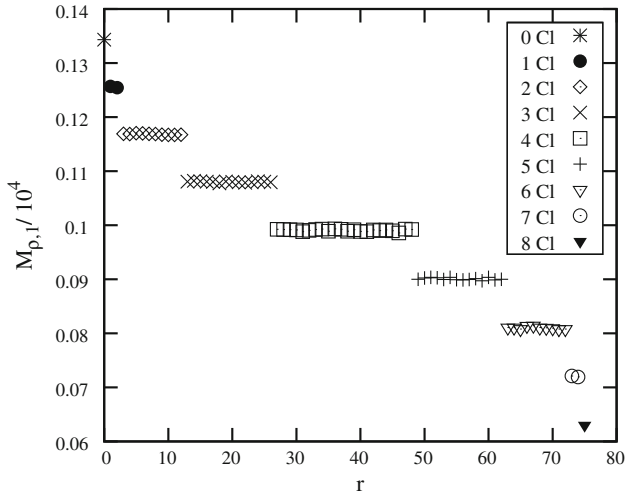

Fig. 1 $M_{\rho,1}$ for 76 compounds

(C), cross-validation (CV), and external prediction for the validation compounds (P) were as follows: $RMSE_{C=0.14}$, $RMSE_{CV} = 0.18$, $RMSE_P = 0.06$. Visual analysis of the correlation plot (Fig. 3) did indicate any outlying results. Similarly, the analysis of the model's applicability domain with Williams plot (Fig. 4) indicated ability of the model for performing reliable predictions for CNs having leverage values h_i below $h^* = 0.5$. This means that the model predicts correctly also for those CNs that have not been previously used for fitting (training). For more details related to developing and validating QSAR/QSPR models one should refer to [23] and [25]. The values of *log* P_L predicted for all 76 CNs together with the calculated leverage values were listed in Table 1.

It is worth noting that the subcooled liquid vapor pressure (*log* P_L) at 25 °C has been already successfully predicted for CNs with QSPR models [26]. The models utilized other

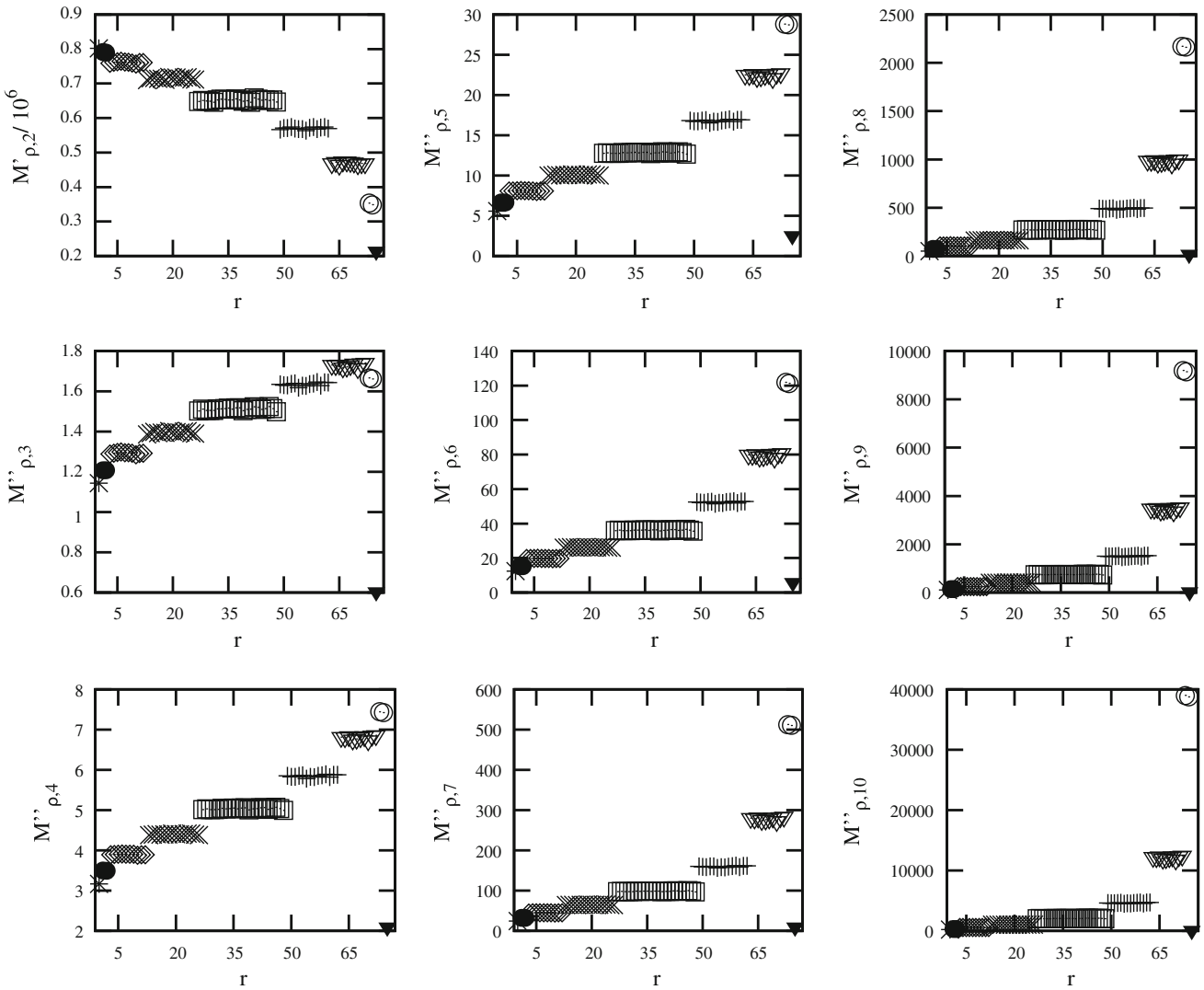

Fig. 2 Spectral density distribution moments

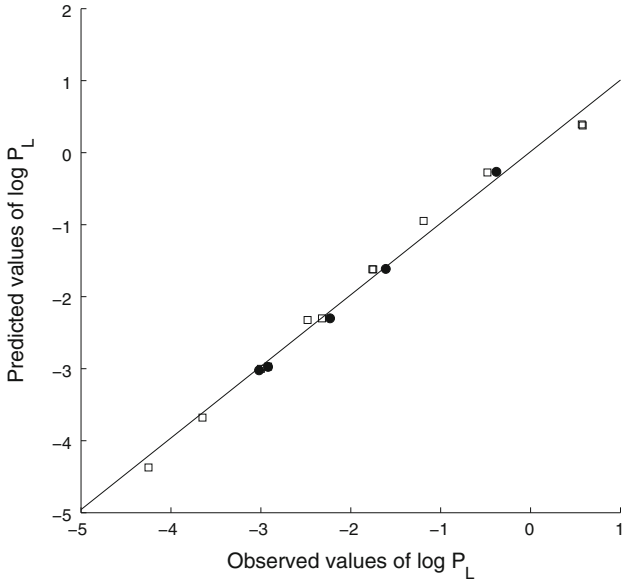

Fig. 3 Plot of the observed (experimental) versus predicted values of *log* P_L. Training compounds are indicated by the *squares*, whereas validation compounds are indicated with the *circles*

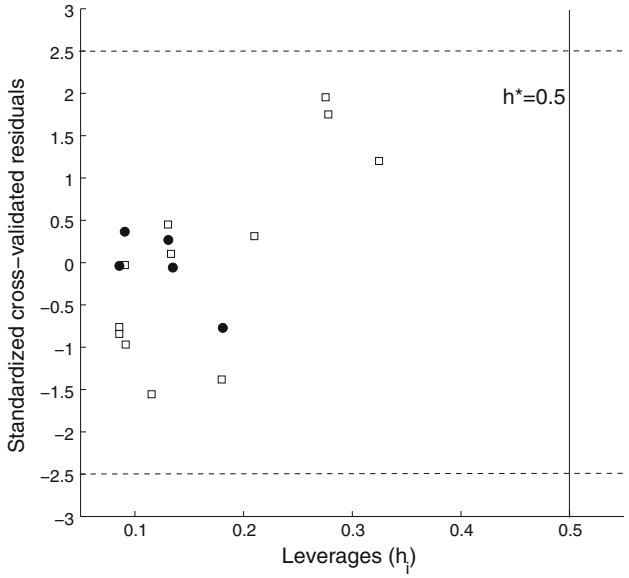

Fig. 4 Applicability domain of the model explored with the Williams plot. Training compounds are indicated by the *squares*, whereas validation compounds are indicated with the *circles*. Standardized cross-validated residuals do not exceed the absolute value of 2.5. All the compounds from the training and validation set are characterized by leverage values lower than the critical leverage $h^* = 0.05$. Predictions for compounds, for which $h_i < h^*$ should be considered as reliable, since they are the results of interpolation by the model

popular quantum-mechanical descriptors (averaged polarizability, dipole moment etc.) calculated at the same level of theory (B3LYP/6-311++G**) with statistical modeling methods of different complexity, including MLR, principal component regression (PCR), partial least square (PLS) regression, and its two modifications: PLS regression with uninformative variable elimination (UVE-PLS) and partial least square regression with variable selection by genetic algorithm (GA-PLS). However, by employing the spectral density distribution moments as novel molecular descriptors in the current study it was possible to develop a model characterized by both lower complexity and better predictive ability than the best model obtained in the previous study. The best original model was developed with GA-PLS, utilized eight descriptors and the prediction error $RMSE_P = 0.108$.

Summarizing, the model presented in the current study has been developed with much simpler algorithm MLR, utilizing only one descriptor with $RMSE_P$ equal to 0.06. This finally confirms the usefulness of the proposed spectral density distribution moments in QSPR.

The proposed descriptors characterize statistical properties of the distributions of the frequencies (not of the intensities) used for their computation. Therefore, the descriptors defined in this study are useful for the description of the properties which are mainly determined by the frequency distributions, such as *log* P_L of chloronaphthalens. The frequency distributions are different if the molecules contain different numbers of chlorine atoms but all isomers of CNs with a fixed number of substituents have nearly the same frequency distributions. Therefore, the moments shown in Figs. 1 and 2 are nearly constant for the compounds with the same number of substituents. In order to distinguish different isomers one has to use the intensity distribution moments [16]. For CNs such descriptors are different for each compound and they will be considered in a subsequent paper.

Acknowledgements The contribution of TP was supported by the Polish Ministry of Science and Higher Education (Grant No. DS/8430-4-0171-1).

References

1. UNEP (2001) Stockholm Convention on persistent organic pollutants, Stockholm, 2001
2. Blankenship AL, Kannan K, Villalobos SA, Villeneuve DL, Falandysz J, Imagawa T, Jakbsson E, Giesy JP (2000) Environ Sci Technol 34:3153
3. Villeneuve DL, Kannan K, Khim JS, Falandysz J, Nikiforov VA, Blankenship AL, Giesy JP (2000) Arch Environ Contam Toxicol 39:273
4. Haranczyk M, Puzyn T, Sadowski P (2008) QSAR Comb Sci 27:826
5. Yamashita N, Taniyasu S, Hanari N, Horii Y, Falandysz J (2003) J Environ Sci Health Part A Toxic/Hazard Subst Environ Eng 38:1745

6. Lee SC, Harner T, Pozo K, Shoeib M, Wania F, Muir DCG, Barrie LA, Jones KC (2007) Environ Sci Technol 41:2680
7. Puzyn T, Mostra A, Suzuki N, Falandysz J (2008) Atmos Environ 42:6627
8. Weem AP (2007) Exploration of management options for polychlorinated naphthalenes (PCN). In: Paper for the Sixth Meeting of the UNECE CLRTAP Task Force on persistent organic pollutants, Vienna, 4–6 June 2007
9. REACH, Regulation (EC) No 1907/2006 of the European Parliament and of the Coincil of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. (2006)
10. Schultz TW, Cronin MTD, Walker JD, Aptula AO (2003) J Mol Struct 622:1
11. Puzyn T, Suzuki N, Haranczyk M (2008) Environ Sci Technol 42:5189
12. Gajewicz A, Haranczyk MTP (2010) Atmos Environ 44:1428
13. Bielińska-Wąż D, Wąż P, Basak SC (2006) Eur Phys J B 50:333
14. Bielińska-Wąż D, Wąż P, Basak SC (2007) J Math Chem 42: 1003
15. Bielińska-Wąż D, Wąż P (2008) J Math Chem 43:1287
16. Bielińska-Wąż D, Nowak W, Pepłowski Ł, Wąż P, Basak SC, Natarajan R (2008) J Math Chem 43:1560
17. Wąż P, Bielińska-Wąż D, Pleskacz A, Strobel A (2008) Acta Phys Pol B 39:1993
18. Wąż P, Bielińska-Wąż D, Strobel A, Pleskacz A (2010) Acta Astron 60:283
19. Wąż P, Bielińska-Wąż D (2009) Acta Phys Pol A 116:987
20. Bielińska-Wąż D, Nowak W, Wąż P, Nandy A, Clark T (2007) Chem Phys Lett 443:408
21. Frisch JM et.al. (2004) Gaussian. Inc.: Wallingford CT
22. Lei YD, Wania F, Shiu WY (1999) J Chem Eng Data 44:577
23. Gramatica P (2007) QSAR Comb Science 26:694
24. OECD, (2007) Guidance document on the validation of (Quantitative) structure–activity relationships, (QSAR) Models, Organization for Economic Co-Operation and Development, Paris
25. Puzyn T, Leszczynski J, Cronin MTD (eds) (2010) Recent advances in QSAR studies: methods and applications. In: Challenges and advances in computational chemistry and physics, ISBN: 978-1-4020-9782-9; Springer, Dordrecht
26. Puzyn T, Falandysz J (2007) SAR QSAR Environ Res 18:299