Journal of Molecular Graphics and Modelling 36 (2012) 10–19

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

QSPR studies of impact sensitivity of nitro energetic compounds using three-dimensional descriptors

Jie Xu[a,b,*], Ligen Zhu[a], Dong Fang[a], Luoxin Wang[a], Shili Xiao[b], Li Liu[a], Weilin Xu[a]

[a] College of Materials Science & Engineering, Wuhan Textile University, 430073 Wuhan, China
[b] Key Lab of Green Processing & Functional Textiles of New Textile Materials, Ministry of Education, Wuhan Textile University, 430073 Wuhan, China

ARTICLE INFO

Article history:
Received 7 December 2011
Received in revised form 30 January 2012
Accepted 10 March 2012
Available online 20 March 2012

Keywords:
QSPR
Impact sensitivity
Nitro energetic compounds
Three-dimensional descriptors

ABSTRACT

The quantitative structure–property relationship (QSPR) studies were performed between molecular structures and impact sensitivity for a diverse set of nitro energetic compounds based on three-dimensional (3D) descriptors. The entire set of 156 compounds was divided into a training set of 127 compounds and a test set of 29 compounds according to Kennard and Stones algorithm. Multiple linear regression (MLR) analysis was employed to select the best subset of descriptors and to build linear models; while nonlinear models were developed by means of artificial neural network (ANN). The obtained models with ten descriptors involved show good predictive power for the test set: a squared correlation coefficient (R^2) of 0.7222 and a standard error of estimation (s) of 0.177 were achieved by the MLR model; while by the ANN model, R^2 and s were 0.8658 and 0.130, respectively. Therefore, the proposed models can be used to predict the impact sensitivity of new nitro compounds for engineering.

1. Introduction

Nitro energetic compounds are widely used as explosives for industrial, civil and military applications. For energetic compounds, higher performance has always been a prime requirement in the field of research and development of explosives. However, the safety, reliability, and stability of energetic compounds also need to be taken into account because safe handling is one of the most important issues to the scientists and engineers who handle energetic compounds [1].

The impact sensitivity, usually expressed as the impact sensitivity index, is one of the qualities to scale the reliable performance of energetic materials. Experimentally, the impact sensitivity is measured by drop weight impact test, where a height of 50% probability in causing an explosion (H_{50}) was measured when hit by a hammer with a standard weight. However, the impact test is time-consuming and sometimes unreliable. Meanwhile, it is impossible to measure the H_{50} value for yet unsynthesized energetic compounds. Consequently, establishing quantitative relationships between the impact sensitivity and molecular structure of energetic compounds by theoretical approaches has become an urgent and important subject, which has been given considerable attention in the last decades. Kamlet and Adolph [2] employed the oxygen balance (OB_{100}) to estimate the impact sensitivity of some nitro compounds. The oxygen balance was defined as the number of equivalents of oxidant per hundred grams of explosive to burn all hydrogen to water and all carbon to carbon monoxide. Politzer and coauthors [3–6] have developed correlations to predict the impact sensitivity of nitro compounds based on their molecular electrostatic potential using quantum mechanical calculations. Rice and Hare [7] used approximations to the electrostatic potential at bond midpoints, statistical parameters of these surface potentials and the generalized interaction properties function or calculated heats of detonation to predict the impact sensitivity of C—H—N—O explosives. Xiao and coauthors [8–11] studied the relationship between the impact sensitivity and the weakest C—NO_2 bond dissociation energy of nitro compounds. Zhang et al. [12–14] found some relationships between the impact sensitivity and nitro group charges in nitro compounds on the basis of density functional theory. Cao and Gao [15] correlated the impact sensitivity of nitrobenzenes and saturated nitro compounds with both OB_{100} and nitro group charges. Keshavarz [16] and Lai et al. [17] found empirical correlations to predict the impact sensitivity of nitrate, nitroaliphatic and the derivatives, respectively.

Quantitative structure–property relationship (QSPR) provides an alternative method for the prediction of impact sensitivity using descriptors derived solely from the molecular structure to fit experimental data. The QSPR method is based on the assumption that the variation of the behavior of the compounds, as expressed by any measured physicochemical properties, can be correlated with numerical changes in structural features of all compounds, termed

* Corresponding author at: College of Materials Science & Engineering, Wuhan Textile University, 430073 Wuhan, China. Tel.: +86 27 87426559.
E-mail address: xujie0@ustc.edu (J. Xu).

1093-3263/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2012.03.002

"molecular descriptors" [18–24]. The advantage of this method lies in the fact that it requires only the knowledge of the chemical structure and is not dependent on any experimental properties. Once a correlation is established and validated, it can be applicable for the prediction of the property of new compounds that have not been synthesized or found. Thus the QSPR method can expedite the process of development of new molecules and materials with desired properties. The QSPR method has been successfully applied to predict the impact sensitivity of C—H—N—O explosives [25–28]. Nefati et al. [25] developed a thirteen-parameter QSPR model for the prediction of the impact sensitivity of C—H—N—O explosives by means of artificial neural network (ANN) based on topological and quantum-chemical descriptors. However, this model was not evaluated with the external set, thus the true prediction ability of the model for new organic compounds which were not used for model development was not clear. In fact, the external validation is a crucial aspect of any QSPR modeling [29]. Cho et al. [26] obtained a seventeen-parameter ANN model to predict the impact sensitivity of C—H—N—O explosives based on compositional and topological descriptors. Recently, Wang et al. [27] built a sixteen-parameter model to predict the impact sensitivity of nitro compounds by means of ANN based on electrotopological-state (ETSI) indices. There are too many descriptors involved in these two models: better results obtained by adding more descriptors into the correlation should be considered carefully due to the danger of over fitting and chance correlations. Morrill and Byrd [28] used the best multilinear regression (BMLR) approach to find correlations between AM1 quantum chemical descriptors and the impact sensitivity for 227 nitroorganic compounds and obtained an eight-descriptor equation with $R^2 = 0.8141$ after removing seven outliers. However, it is confused that only eight compounds were used as test set even though there were so many compounds available.

Molecular descriptors can be defined as the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number [30]. Each molecular descriptor takes into account a small part of the whole chemical information contained in the real molecule. Molecular descriptors play a fundamental role in developing QSPR models. An efficient descriptor should encode as much structural information as possible. In the literature over 6000 descriptors have been proposed, and the number still grows. Based on the dimensionality of the molecular representation, molecular descriptors can be classified into zero-dimensional (0D), one-dimensional (1D), two-dimensional (2D), three-dimensional (3D) and even higher dimensional groups. The good predictive abilities of 3D descriptors in QSPR studies have been confirmed in comparison with 2D and quantum-chemical descriptors [31], since the higher dimensional structure of a molecule contains more information. In the present work, the QSPR method was employed to predict the impact sensitivity of a diverse set of 156 nitro energetic compounds based on 3D descriptors. Linear and nonlinear models were developed by means of multiple linear regression (MLR) analysis and ANN. In addition, the contributions of the involved descriptors to the models were discussed in detail.

2. Materials and method

2.1. Dataset

Experimental impact sensitivity (measured by the logarithmic 50% impact height, lgH_{50}) of 156 diverse compounds containing C, H, O and N were taken from the literature [27]. The compounds in the dataset include several families: 49 nitroaromatics, 55 nitramine, 40 nitroaliphatic compounds containing other functional groups, 7 nitrate esters, and 5 nitroaliphatic compounds. The experimental lgH_{50} values are given in Table 1.

2.2. Descriptor generation

Three-dimensional descriptors depend explicitly on the conformation of the molecule for which they are being computed. To avoid the introduction of bias in terms of such descriptors, conformational analysis was performed by rotating the rotatable bonds to 180° (in 10° increments). Subsequently, the conformer models were optimized using the MM+ molecular mechanics method (Polak–Ribiere algorithm) and the semi-empirical AM1 method at a restricted Hartree–Fock level with no configuration interaction in the HYPERCHEM program (Version 6.01) [32]. A gradient norm limit of 0.05 kcal $Å^{-1}$ mol^{-1} was chosen as the stopping criterion for the geometry optimization. The resulting lowest energy conformer for each compound was then submitted to the DRAGON software (Version 5.4) [33] to calculate 735 3D descriptors.

In order to reduce redundant and non-useful information, constant or near constant values and descriptors found to be highly correlated pairwise (one of any two descriptors with a correlation greater than 0.99 [34]) were excluded in a pre-reduction step. Consequently, 441 descriptors were remained to undergo subsequent descriptor selection.

2.3. Kennard and Stones algorithm

Kennard and Stones algorithm [35] has been widely used for splitting datasets into two subsets. This algorithm starts by finding two samples, based on the input variables that are the farthest apart from each other. These two samples are removed from the original dataset and put into the training set. Then, the remaining sample farthest away from the selected two samples is again included in training set. This step is repeated until the desired number of samples has been selected in the training set. The advantages of this algorithm are that the training samples always map the measured region of the input variable space completely with respect to the induced metric and that the no test samples fall outside the measured region. Kennard and Stones algorithm has been considered as one of the best ways to build training and test sets [36,37]. Using Kennard and Stones algorithm combined with principal component analysis (PCA), the entire dataset was divided into two subsets: a training set of 127 compounds, and a test set including the remaining 29 compounds.

2.4. Model development and validation

Linear models were developed by applying stepwise MLR analysis with Leave-Many-Out (LMO) cross-validation to the training set. Step-by-step variables are added to the equation, and a new regression is performed. If the new variable contributes significantly to the regression equation, the variable is retained; otherwise, the variable is excluded, hence preventing over-fitting. *F*-to-enter and *F*-to-remove were 4 and 3, respectively; that is, the variable was retained if its *F*-value was higher than 4; while the variable was removed if its *F*-value fell below 3. Meanwhile, for each regression equation five samples of the original training set were removed, and the model was recalculated using the remaining $n - 5$ samples as training set. The response was then predicted for the excluded samples. This process was repeated for all samples of the training set, obtaining a prediction for every one and thus the cross-validated $R^2(R^2_{CV})$. The quality of the models was measured by the R^2, the adjusted R^2, the cross-validated R^2, the *F* ratio values, the standard

Table 1
Experimental and calculated lgH_{50} of nitro energetic compounds.

No.	Compounds	lgH_{50}	lgH_{50}	
			MLR	ANN
1	1-Methoxy-3,5-dichloro-2,4,6-trinitrobenzene	1.88	1.81	1.88
2	1,3,5-Triamino-2,4,6-trinitrobenzene	2.69	2.20	2.34
3	2,4,6-Trinitrophloroglucinol[a]	1.43	1.78	1.69
4	3,3′-Dihydroxy-2,2′,4,4′,6,6′-hexanitrobiphenyl	1.60	1.64	1.65
5	1-Dinitromethyl-3-nitrobenzene	2.02	2.15	1.94
6	N,N′-dinitro-methanediamine[a]	1.11	1.33	1.11
7	N-Methyl-N,N′-dinitro-1,2-ethanediamine	2.06	1.78	1.89
8	N,N′-dinitro-N-[2-(nitroamino)ethyl]-1,2-ethanediamine	1.59	1.64	1.59
9	N,N′-bis-(2,2,2-Trinitroethyl)-N,N′-dinitromethanediamine[a]	0.70	0.83	0.78
10	N,N′-dinitro-N,N′-bis-[2-nitroamino-ethyl]-1,2-ethanediamine	1.72	1.78	1.70
11	1,1,1,3,6,9,11,11,11-Nonanitro-3,6,9-triazaundecane	1.08	1.13	1.05
12	1,1,1,3,6,6,8,10,10,13,15,15,15-Tridecanitro-3,8,13-triazapentadecane	1.36	1.15	1.23
13	N-(t-buty)-trinitroacetamide	2.04	2.10	2.16
14	1,1,1,7,7,7-Hexanitroheptanone-4[a]	1.53	1.34	1.42
15	Trinitroethyl-bis-(trinitroethoxy)-acetate	0.78	0.84	0.75
16	Tetrakis-(2,2,2,-trinitroethyl)-orthocarbonate	0.85	0.81	0.91
17	Methylene-bis-(4,4,4-trinitrobutyramide)	2.05	1.38	1.80
18	bis-(2,2,2-Trinitroethyl)-4,4,6,6,8,8-hexanitro-undecanedioate	1.51	1.63	1.61
19	2,2-bis-(Nitroxymethyl)-1,3-propanediol dinitrate	1.11	0.90	1.07
20	2,4,6-Trinitrobenzyl chloride	1.64	1.79	1.69
21	2,4,6-Trinitrobenzyl alcohol	1.72	1.88	1.74
22	1-Hydroxyethyl-2,4,6-trinitrobenzene	1.83	2.19	2.20
23	2,4,6-Trinitrobenzoic acid	2.04	1.63	1.49
24	2,4,6-Trinitrotoluene	2.20	2.30	2.36
25	1-Ethoxy-2,4,6-trinitrobenzene	2.28	2.44	2.40
26	Hexanitro benzene	1.08	1.19	1.13
27	2′,2′,2′-Trinitroethyl-2,4,6-trinitrobenzoate	1.38	1.82	1.88
28	2,4,6-Trinitro-3-amino-phenol	2.14	1.98	2.08
29	2′,2′,2′-Trinitroethyl-3,5-dinitrosalicylate	1.65	1.88	1.70
30	1-(2,2,2-Trinitroethyl)-2,4,6-trinitrobenzene	1.11	0.93	1.15
31	1-(2,2,2-Trinitroethyl)-2,4-dinitrobenzene[a]	1.49	1.37	1.52
32	2,4,6-Trinitrobenzylalcohol	1.72	1.87	1.73
33	3,5-Dinitro-2,4,6-trinitrophenol	1.89	1.55	1.87
34	N,N′-dinitromethylene-bis-(4,4,4-trinitro)-butyramide	1.11	1.39	1.23
35	bis-(5,5,5-Trinitro-3-nitrazapentanoyl)-methylenedinitramine	1.18	1.35	1.29
36	2,2,2-Trinitroethyl-N-(2,2,2-trinitroethyl)-nitramino acetate	0.95	0.76	1.02
37	2,2,2-Trinitroethyl-4-nitrazavalerate	1.54	1.15	1.32
38	Trinitropropyl-(2,2-dinitropropyl)-nitramine	1.23	1.03	1.14
39	Trinitroethyl-5,5-dinitro-3-nitrazahexanoate	1.40	1.40	1.30
40	2,2,2-Trinitroethyl-2,5,5-trinitro-2-azahexanoate	1.34	1.21	1.19
41	bis-(2,2,2-Trinitroethyl)-3-nitrazaglutarate[a]	1.15	1.29	1.36
42	N,N′-dinitro-N,N′-bis-(3,3,3-trinitropropyl)-oxamide	0.95	0.94	0.90
43	bis-(2,2,2-Trinitroethyl)-4-nitraza-1,7-heptanedioate	1.46	1.38	1.64
44	bis-(2,2,2-Trinitroethyl)-3,6-dinitraza-1,8-octanedioate	1.46	1.57	1.41
45	Trinitroethyl-2-methoxy-ethylnitramine[a]	1.62	1.65	1.78
46	N-methyl-N′-trinitroethyl-N,N′-dinitro-1,2-ethanediamine	1.04	1.24	1.16
47	1,1,1,3,6,9,12,14,14,14-Decanitro-3,6,9,12-tetraza-tetradecane	1.28	1.26	1.36
48	bis-trinitroethyl-5,5-dinitro-2,8-dinitraza-nonanedioate	1.08	0.94	0.95
49	2,2-Dinitro-1,3-propane diol	2.04	2.03	2.03
50	4,4,4-Trinitrobutyramide	1.60	1.57	1.48
51	bis-(Trinitroethoxy)-methane	1.23	1.30	1.20
52	N-(2-propyl)-trinitroacetamide	2.05	2.08	2.18
53	2,2,2-Trinitroethyl-4,4-dinitrovalerate	1.85	1.58	1.71
54	N,N′-bis-(3,3,3-trinitropropyl)-oxamide	1.65	1.37	1.37
55	N,N-bis-(2,2-dinitropropyl)-4,4,4-trinitrobutyramide[a]	1.86	2.08	1.74
56	Trinitroethyl-2,2-dinitropropylcarbonate[a]	1.18	1.39	1.40
57	bis-(2,2,2-Trinitroethyl)-succinate	1.48	1.40	1.59
58	5,5,5-Trinitropentanone-2	2.10	1.72	1.80
59	2,2-Dinitropropyl-4,4,4-trinitrobutyramide	1.86	1.84	1.87
60	3-[N-(2,2,2-trinitroethyl)-nitramino]-propylnitrate[a]	1.08	1.37	1.36
61	1,9-Dinitrato-2,4,6,8-tetranitrazanonane[a]	1.00	1.42	1.25
62	3,5,5-Trinitro-3-azahexyl-nitrate[a]	1.32	1.55	1.24
63	2,2,4,6,6-Pentanitroheptane	1.75	1.71	1.78
64	3,3,4,4-Tetranitro-hexane	1.90	2.03	1.97
65	2,4,6-Trinitro-acetylbenzene[a]	1.90	1.97	1.94
66	Methyl-2,4,6-trinitrobenzoate	1.95	1.89	1.84
67	2,4,6-Trinitroaniline	2.25	2.06	1.92
68	1,3-Diamino-2,4,6-trinitrobenzene	2.51	2.13	2.30
69	1-Methyl-3,5-diamino-2,4,6-trinitrobenzene	2.38	2.43	2.55
70	1,3,5-Trinitrobenzene	2.00	2.04	1.95
71	1,2,4,5-Tetranitrobenzene	1.43	1.53	1.38
72	2,4,6-Trinitroresorcinol	1.63	1.84	1.77
73	2,3,4,6-Tetranitroaniline	1.61	1.80	1.70
74	2,4-Dinitroresorcinol	2.47	2.16	2.31

Table 1 (*Continued*)

No.	Compounds	lgH_{50}	lgH_{50}	
			MLR	ANN
75	1-Hydroxy-3,5-diamino-2,4,6-trinitrobenzene	2.08	2.10	2.23
76	2′,2′,2′-Trinitroethyl-3,5-dinitrobenzoate	1.86	1.71	1.69
77	2,2′,4,4′,6,6′-Hexanitrobiphenyl	1.93	1.94	2.11
78	3-Hydroxy-2,2′,4,4′,6,6′-hexanitrobiphenyl	1.62	1.69	1.69
79	2,2′,4,4′,6,6′-Hexanitrodiphenylamine	1.68	1.80	1.80
80	2,2′,4,4′,6-Pentanitrobenzophenone	1.73	2.02	1.79
81	1-(3,3,3-Trinitropropyl)-2,4-dinitrobenzene	1.49	1.87	1.86
82	1-(3,3,3-Trinitropropyl)-2,4,6-trinitrobenzene[a]	1.32	1.56	1.56
83	3-Methyl-2,2′,4,4′,6′-pentanitrobiphenyl	2.16	2.01	2.00
84	3,3′-Dimethyl-2,2′,4,4′,6,6′-hexanitrobiphenyl	2.13	2.19	2.10
85	N,N′-dinitro-1,2-ethanediamine	1.53	1.42	1.46
86	Cyclotrimethylene-trinitramine	1.41	1.54	1.23
87	bis-(2,2,2-Trinitroethyl)-nitramine	0.70	0.89	0.78
88	N,N′-dimethyl-N,N′-dinitrooxamide	1.90	1.78	1.68
89	1,3,3,5,5-Pentanitropiperidine	1.15	0.99	1.14
90	2,2,2-Trinitroethyl-3′,3′,3′-trinitropropyl-nitramine[a]	0.78	0.97	0.86
91	N,N′-3,3-tetranitro-1,5-pentanediamine[a]	1.54	1.43	1.60
92	N-nitro-N-(3,3,3-trinitropropyl)-2,2,2-trinitroethyl-carbamate	0.95	1.11	1.04
93	2,2,2-Trinitroethyl-3,3-dinitrobutyl-nitramine[a]	1.30	1.13	1.18
94	bis-(Trinitroethyl)-2,4-dinitrazapentanedioate[a]	1.00	1.45	1.35
95	2,2-Dinitropropyl-5,5,5-trinitro-2-nitrazapentanoate	1.20	1.69	1.73
96	N-nitro-N,N′-bis-(trinitropropyl)-urea	1.32	1.22	1.27
97	2,2,2-Trinitroethyl-2,4,6,6-tetranitro-2,4-diazaheptaneoate	1.26	1.08	1.17
98	bis-(Trinitroethyl)-2,4,6-trinitraza-heptanedioate[a]	1.11	0.93	1.09
99	N-(2,2-dinitrobutyl)-N,2,2-trinitro-1-butanamine[a]	1.90	1.99	1.96
100	2,2,4,7,9,9-Hexanitro-4,7-diazadecane[a]	1.86	1.50	1.64
101	1,1,1,4,6,6,8,11,11,11-Decanitro-4,8-diazaundecane	1.04	1.14	1.01
102	1,1,1,3,6,6,9,11,11,11-Decanitro-3,9-diazaundecane	1.00	1.18	1.13
103	bis-(2,2,2-Trinitroethyl)-2,5,8-trinitraza nonanedioate	1.23	1.35	1.07
104	N,N′-dinitro-N,N′-bis-(3,3-dinitrobutyl)-oxamide	1.57	1.59	1.58
105	2,2,4,7,7,10,12,12-Octanitro-4,10-diazatridecane	1.64	1.71	1.60
106	2,2,5,7,7,9,12,12-Octanitro-5,9-diazatridecane	1.57	1.49	1.47
107	2,2-Dinitropropanediol-bis-(5,5-dinitro-2-nitraza-hexanoate)	2.14	1.82	1.81
108	2,2,2-Trinitroethyl-carbamate	1.26	1.20	1.17
109	Methyl-2,2,2-trinitroethyl carbonate	1.45	1.39	1.36
110	bis-(2,2,2-Trinitroethyl)-carbonate[a]	1.20	1.43	1.38
111	N,N′-bis-(2,2,2-trinitroethyl)-urea	1.23	1.37	1.39
112	bis-(Trinitroethyl)-oxalate	1.18	1.14	1.01
113	bis-(Trinitroethyl)-oxamide	1.11	1.07	1.14
114	N-trinitroethyl-4,4,4-trinitrobutyramide	1.26	1.39	1.36
115	tris-(2,2,2-Trinitroethyl)-orthoformate	0.85	0.95	0.96
116	2,2-Dinitropropyl-trinitrobutyrate	2.18	1.67	1.78
117	bis-(2,2-Dinitropropyl)-carbonate	2.48	1.91	2.11
118	bis-(Trinitropropyl)-urea	1.36	1.44	1.36
119	4,4,4-Trinitrobutyricanhydride[a]	1.48	1.64	1.45
120	bis-(2,2-Dinitropropyl)-oxalate	2.36	2.16	2.42
121	2,2,2-Trinitroethyl-4,4-dinitrohexanoate	2.14	1.89	1.95
122	Ethylene-bis-(4,4,4-trinitrobutyrate)[a]	2.08	1.88	1.98
123	2,2-Dinitropropane-1,3-diol-bis-(4,4,4-trinitrobutyrate)	1.70	1.49	1.58
124	1,2,3-Propanetrioltrinitrate	1.30	1.15	1.16
125	N-(2,2,2-trinitroethyl)-nitraminoethyl nitrate	0.85	1.24	1.02
126	1,1,1,3,5,5,5-Heptanitropentane	0.90	0.79	0.86
127	2,2,4,4,6,6-Hexanitroheptane	1.46	1.63	1.64
128	2,4,6-Trinitro-1-chlorobenzene[a]	1.90	1.79	1.89
129	1,3-Dimethoxy-2,4,6-trinitrobenzene	2.40	2.16	2.20
130	2,3,4,5,6-Pentanitroaniline	1.18	1.51	1.31
131	2′,2′-Dinitropropyl-2,4,6-trinitrobenzoate	2.33	2.17	2.11
132	3,3′-Diamino-2,2′,4,4′,6,6′-hexanitrodiphenylamine	2.12	1.87	1.85
133	Picric acid	1.94	1.87	1.81
134	2,4,6-Trinitroanisole	2.28	2.04	2.12
135	2,4,6-Trinitro-m-cresol	2.28	2.24	2.29
136	3-Methyl-2,2′,4,4′,6,6′-hexanitrobiphenyl	1.72	1.48	1.73
137	2,2′,4,4′,6,6′-Hexanitrobibenzyl[a]	2.06	2.06	2.18
138	N-methyl-N-nitro-(trinitroethyl)-carbamate	1.23	1.41	1.36
139	Trinitroethyl-N-ethyl-N-nitro-carbamate	1.28	1.70	1.59
140	2′,2′,2′-Trinitroethyl-2,5-dinitrazahexanoate	1.18	1.31	1.17
141	N-(2,2-dinitropropyl)-N,2,2-trinitro-1-propanamine	1.46	1.61	1.58
142	N,N′-dinitro-N,N′-bis-(3-nitrazabutyl)-oxamide[a]	1.95	1.78	1.85
143	1,1,1,18,18,18-Hexanitro-3,16-dioxa-4,15-dioxo-5,8,11,14-tetranitrazaoctadecane	1.28	1.58	1.26
144	Methyl-2,2,2-trinitroethylnitramine	0.95	1.24	1.03
145	1,7-dimethoxy-2,4,6-trinitrazaheptane	2.22	2.20	2.15
146	Cyclotetramethylene-tetranitramine	1.46	1.45	1.47
147	2,2,6,9,9-Pentanitro-4-oxa-5-oxo-6-azadecane	1.67	1.94	1.84

Table 1 (*Continued*)

No.	Compounds	lgH_{50}	lgH_{50}	
			MLR	ANN
148	2,2,2-Trinitroethyl-4,4,4-trinitrobutyrate[a]	1.26	1.44	1.24
149	bis-(2,2,2-Trinitroethyl)-4,4-dinitroheptanedioate	1.83	1.86	2.02
150	Methylene-bis-N,N′-(2,2,2-trinitroacetamide)[a]	0.95	1.14	1.09
151	Methylene-bis-(trinitroethyl)-carbamate	1.43	1.51	1.53
152	Nitroisobutyl-4,4,4-trinitrobutyrate	2.45	2.35	2.33
153	1,5-bis-(Trinitroethyl)-biuret	1.38	1.17	1.24
154	Ethyl-2,2,2-trinitroethylcarbonate	1.91	1.70	1.77
155	4,4,8,8-Tetranitro-1,11-dinitro-6-nitrazaundecane	1.94	1.40	1.35
156	1,1,1,3-Tetranitrobutane	1.52	1.67	1.82

[a] Compounds in the test set.

error of estimation s and the significance level value p. The adjusted R^2 is calculated using the following formula:

$$R^2_{\text{adj}} = 1 - \left[\left(\frac{N-1}{N-M-1}\right)(1-R^2)\right] \quad (1)$$

where N is the number of members of the training set and M is the number of descriptors involved in the correlation. The adjusted R^2 is a better measure of the proportion of variance in the data explained by the correlation than R^2 (especially for correlations developed using small datasets) because R^2 is somewhat sensitive to changes in N and M. The adjusted R^2 corrects for the artificiality introduced when M approaches N through the use of a penalty function which scales the result. A variance inflation factor (VIF) was calculated to test if multicollinearities existed among the descriptors, which is defined as

$$\text{VIF} = \frac{1}{1-R^2_j} \quad (2)$$

where R^2_j is the squared correlation coefficient between the jth coefficient regressed against all the other descriptors in the model. Models would not be accepted if they contain descriptors with VIF above a value of five [38].

To exclude the possible existence of fortuitous correlations, the proposed MLR model was also checked by randomization tests: new models were recalculated for randomly shuffled lgH_{50} values and original independent-variable matrix. The models obtained from the training set with randomized lgH_{50} values should have significantly lower R^2 and R^2_{CV} values than the proposed one because the relationship between the structure and property has been broken. This is a proof of the proposed model's validity as it can be reasonably excluded that the originally proposed model was obtained due to chance correlation or structural redundancy of the training set.

Nonlinear models were then developed by submitting the selected descriptors from MLR to a three-layer, fully connected, feed-forward ANN. The number of input neurons was equal to that of the descriptors in the linear model. The number of hidden neurons was optimized by trial and error procedure on the training process. One output neuron was used to represent the experimental lgH_{50}. The network was trained using the quasi-Newton BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm [39]. To avoid overtraining, one tenth data from the training set were randomly selected as separate validation set to monitor the training process; that is, during the training of the network the performance was monitored by predicting the values for the systems in the validation set. When the results for the validation set ceased to improve, the training was stopped.

Validation of the linear and nonlinear models was further performed by using the external test set composed of data not used to develop the prediction model. The external $R^2_{\text{CV,ext}}$ for the test sets is determined as:

$$R^2_{\text{CV,ext}} = 1 - \frac{\sum (y_i - \tilde{y}_i)^2}{\sum (y_i - \bar{y}_{\text{tra}})^2} \quad (3)$$

where y_i and $\tilde{y}_i$ are the observed and the calculated values, respectively; and $\bar{y}_{\text{tra}}$ is the averaged value for the response variable of the training set. According to Tropsha and coauthors [29,36], a QSPR model is successful if it satisfies several criteria as follows:

$$R^2_{\text{CV,ext}} > 0.5 \quad (4a)$$

$$r^2 > 0.6 \quad (4b)$$

$$\frac{r^2 - r_0^2}{r^2} < 0.1 \quad \text{or} \quad \frac{r^2 - r'^2_0}{r^2} < 0.1 \quad (4c)$$

$$0.85 \le k \le 1.15 \quad \text{or} \quad 0.85 \le k' \le 1.15 \quad (4d)$$

Here:

$$r = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (5a)$$

$$r_0^2 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \quad (5b)$$

$$r'^2_0 = 1 - \frac{\sum (y_i - y_i^{r_0})^2}{\sum (y_i - \bar{y})^2} \quad (5c)$$

$$k = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2} \quad (5d)$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \quad (5e)$$

where r is the correlation coefficient between the calculated and experimental values in the test set; r_0^2 (calculated versus observed values) and r'^2_0 (observed versus calculated values) are the coefficients of determination; k and k' are slopes of regression lines through the origin of calculated versus observed and observed versus calculated, respectively; $y_i^{r_0}$ and $\tilde{y}_i^{r_0}$ are defined as $y_i^{r_0} = k\tilde{y}_i$ and $\tilde{y}_i^{r_0} = k' y_i$, respectively; and the summations are over all samples in the test set.

2.5. Applicability domain analysis

The applicability domain (AD) of a QSPR model [36,40] must be defined if the model is to be used for screening new compounds. The AD is a theoretical region in the space defined by the descriptors of the model and the modeled response, for which a given QSPR should make reliable predictions. This region is defined by

the nature of the compounds in the training set, and can be characterized in various ways. In this work, the structural AD was verified by the leverage approach. The leverage h_i [40] is defined as follows:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (6)$$

where x_i is the descriptor row-vector the ith compound, x_i^T is the transpose of x_i, X is the descriptor matrix, X^T is the transpose of X. The warning leverage h^* is, generally, fixed at $3(m+1)/n$, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation. In fact, leverage can be used as a quantitative measure of the model AD suitable for evaluating the degree of extrapolation. It represents a sort of compound distance from the model experimental space.

The Williams plot, the plot of leverage values versus standardized residuals, was used to give a graphical detection of both the response outliers (Y outliers) and the structurally influential compounds (X outliers). In this plot, the two horizontal lines indicate the limit of normal values for Y outliers (i.e. samples with standardized residuals greater than 3.0 standard deviation units, $\pm 3.0s$); the vertical straight lines indicate the limits of normal values for X outliers (i.e. samples with leverage values greater than the threshold value, $h > h^*$). For a sample in the external test set whose leverage value is greater than h^*, its prediction is considered unreliable, because the prediction is the result of a substantial extrapolation of the model. Conversely, when the leverage value of a compound is lower than the critical value, the probability of accordance between predicted and experimental values is as high as that for the compounds in the training set. It is noteworthy that the response outliers can be highlighted only for compounds with known responses and the possibility of a compound to be out of the structural AD of a model can be verified for every new compound, the only knowledge needed being the molecular structure information represented by the molecular descriptors selected in the model.

3. Results and discussion

3.1. MLR model

Stepwise MLR analysis with LMO cross-validation was used to develop linear models from the training set and the number of descriptors in the final model was determined on the basis of the dataset size and on the basis of the R^2, the adjusted R^2, the cross-validated R^2, the significance test F and the standard error s. Obviously, lgH_{50} is not linearly correlated with any of the molecular descriptors since univariant correlations between lgH_{50} and the descriptors have poor R^2 values. The R^2 increased gradually with the increased number of descriptors. When adding another descriptor did not significantly improve the statistics of a model, it was determined that the optimum subset size had been achieved. The resulted ten-parameter equation was as the following:

$$\begin{aligned} lgH_{50} &= 1.792 + 0.167[\mathrm{AROM}] - 0.399[\mathrm{Mor29u}] \\ &+ 0.0158[\mathrm{Mor02v}] + 0.896[\mathrm{Ds}] - 1.039[\mathrm{HATS3u}] \\ &- 0.860[\mathrm{HATS7m}] - 1.524[\mathrm{R1m}] + 2.538[\mathrm{R1m+}] \\ &- 2.30[\mathrm{R6m+}] + 2.768[\mathrm{R1v}] \end{aligned} \quad (7)$$

$n = 127$, $R^2 = 0.7684$, $R^2_{CV} = 0.7031$, $R^2_{adj} = 0.7446$, $s = 0.194$, $F = 46.2$, $p < 0.00001$.

Here, AROM is the aromaticity index; Mor29u is the 3D-MoRSE – signal 29/unweighted; Mor02v is the 3D-MoRSE – signal 02/weighted by atomic van der Waals volumes; Ds is the D total accessibility index/weighted by atomic electrotopological states; HATS3u is the leverage-weighted autocorrelation of lag 3/unweighted; HATS7m is the leverage-weighted autocorrelation of lag 3/weighted by atomic masses; R1m is the R autocorrelation of lag 1/weighted by atomic masses; R1m+ is the R maximal autocorrelation of lag 1/weighted by atomic masses; R6m+ is the R maximal autocorrelation of lag 6/weighted by atomic masses; R1v is the R autocorrelation of lag 1/weighted by atomic van der Waals volumes, respectively. More information about these descriptors can be found in Dragon software user's guide [33] and the references therein.

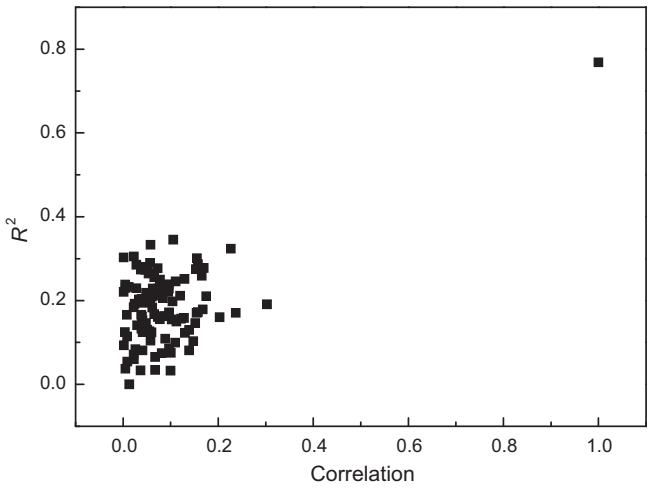

Fig. 1. R^2 vs. correlation coefficient between the original and permuted response data (average value of R^2 is 0.180).

The large F ratio of 46.2 indicates that Eq. (7) does an excellent job of predicting the lgH_{50} values. Eq. (7) has an adjusted R^2 value of 0.7446, which indicates good agreement between the correlation and the variation in the data. The cross-validated correlation coefficient $R^2_{CV} = 0.7031$ illustrates the reliability of the model by focusing on the sensitivity of the model to the elimination of any five data point. The model was further validated by applying the randomization tests and the obtained R^2 vs. the correlation coefficient between the original and permuted response data are plotted in Fig. 1. The lower R^2 values indicate that the good results of the original model are not due to chance correlation or structural dependency of the training set. Some important statistical parameters shown in Table 2 were used to valuate the involved descriptors. The t-value of a descriptor measures the statistical significance of the regression coefficients. The high absolute t-values shown in Table 2 express that the regression coefficients of the descriptors involved in the MLR model are significantly larger than the standard deviation. The t-probability of a descriptor can describe the statistical significance when combined together within an overall collective QSPR model (i.e., descriptors' interactions). Descriptors with t-probability values below 0.05 (95% confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance [41]. The smaller t-probability suggests the more significant descriptor. The t-probability values of the ten descriptors are very small, indicating that all of them are highly significant descriptors. The VIF values (less than five) suggest that these descriptors are weakly correlated with each other. Thus, the model can be regarded as an optimal regression equation.

The calculated lgH_{50} values from the MLR model for the training and test sets are shown in Table 1 and Fig. 2. The errors are distributed on both sides of the zero point, thus one may conclude that there is no systematic error in the model development. The following statistical parameters were obtained for the test set,

Table 2
Characteristics of the descriptors selected in the optimal MLR model.

Descriptor	Descriptor type	*X*	DX	*t*-Value	*t*-Probability	VIF
Constant		1.792	0.228	7.850	0.000	
AROM	Geometrical descriptors	0.167	0.057	2.952	0.004	1.796
Mor29u	3D-MoRSE descriptors	−0.399	0.111	−3.594	0.000	1.697
Mor02v	3D-MoRSE descriptors	0.0158	0.006	2.474	0.015	1.504
Ds	WHIM descriptors	0.896	0.222	4.029	0.000	1.762
HATS3u	GETAWAY descriptors	−1.039	0.215	−4.832	0.000	2.350
HATS7m	GETAWAY descriptors	−0.860	0.137	−6.279	0.000	1.201
R1m	GETAWAY descriptors	−1.524	0.135	−11.272	0.000	2.769
R1m+	GETAWAY descriptors	2.538	0.380	6.676	0.000	1.846
R6m+	GETAWAY descriptors	−2.360	0.407	−5.794	0.000	1.188
R1v	GETAWAY descriptors	2.768	0.336	8.247	0.000	1.642

which obviously satisfy the generally accepted condition and thus demonstrate the predictive power of the present model:

$$R^2_{CV,ext} = 0.7455 > 0.5$$
$$r^2 = 0.7222 > 0.6$$
$$\frac{r^2 - r_0^2}{r^2} = \frac{0.7222 - 0.9724}{0.7222} < 0.1$$
$$\text{or} \quad \frac{r^2 - r'^2_0}{r^2} = \frac{0.7222 - 0.9594}{0.7222} < 0.1$$
$$0.85 \le k = 0.947 \le 1.15 \quad \text{or} \quad 0.85 \le k' = 1.035 \le 1.15$$

3.2. ANN model

The ANN has become an important and widely used nonlinear modeling technique for QSPR studies. The mathematical adaptability of ANN commends it as a powerful tool for pattern classification and building predictive models. A particular advantage of ANN is its inherent ability to incorporate nonlinear dependencies between the dependent and independent variables without using an explicit mathematical function. Among the neural network learning algorithms, the back-propagation (BP) method [42] is one of the most commonly used methods. The drawback of BP is that the training processes slowly, because the gradient-descent algorithm is usually used for minimizing the sum-of-squares error. In this study, the quasi-Newton BFGS algorithm [39] was used to develop nonlinear models. The advantages of the BFGS algorithm are that specifying rate or momentum is not necessary and the training is much more rapid [43]. The ten descriptors from the MLR model were used as inputs to the network. The number of hidden neurons is an important parameter influencing the performances of the ANN. The usual rule of thumb is that the weights and biases should be less than the samples so that the model achieved by the network is stationary [44]. Thus, a 10-7-1 network architecture was obtained after trial and error procedure.

The predictive results from the ANN model for the entire dataset are given in Table 1 and Fig. 3. The R^2, R^2_{CV} and s for the training set are 0.8481, 0.8456 and 0.164, respectively. The proposed ANN model is predictive as it satisfies the conditions for the test set (s = 0.130):

$$R^2_{CV,ext} = 0.8728 > 0.5$$
$$r^2 = 0.8658 > 0.6$$
$$\frac{r^2 - r_0^2}{r^2} = \frac{0.8658 - 0.9840}{0.8658} < 0.1$$
$$\text{or} \quad \frac{r^2 - r'^2_0}{r^2} = \frac{0.8658 - 0.9799}{0.8658} < 0.1$$
$$0.85 \le k = 0.963 \le 1.15 \quad \text{or} \quad 0.85 \le k' = 1.029 \le 1.15$$

By comparison of the models obtained by MLR and ANN, it can be seen that the performance of ANN is significantly better than that of MLR, which confirms the nonlinear relationship between the structural information and the lgH_{50}. To further test the suitability of the proposed models, the obtained results were compared with those calculated by other reported models [25,27,28] (Table 3), especially with those by Wang et al. [27] which employed the same dataset as this study. It can be seen that the performance of the present models is a little better than the models described by Wang et al. Additionally, there are ten descriptors in our models, while there are sixteen descriptors in the models built by Wang et al. Improvement of results by increasing the number of descriptors in the correlation equation should be considered with care, since over fitting and

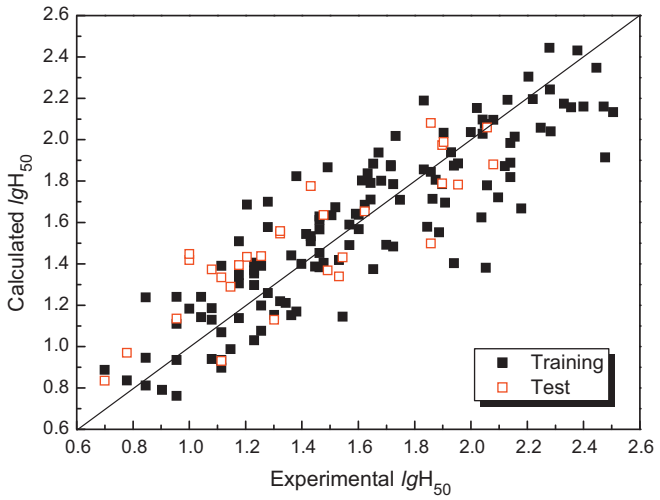

Fig. 2. Calculated vs. experimental lgH_{50} values by the MLR model for the entire dataset.

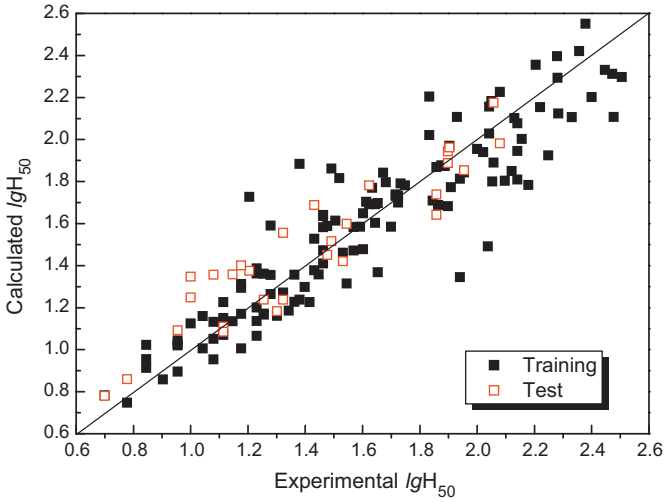

Fig. 3. Calculated vs. experimental lgH_{50} values by the ANN model for the entire dataset.

Table 3
Performance comparison between different QSPR models for lgH_{50}.

Model	Training set			Test set		
	R^2	R^2_{CV}	s	R^2	R^2_{CV}	s
Present MLR	0.7684	0.7031	0.194	0.7222	0.7455	0.177
Present ANN	0.8481	0.8456	0.164	0.8658	0.8728	0.130
Nefati et al. (ANN) [25]	0.795	–	0.203	0.756	0.703	0.257
Wang et al. (MLR) [27]	0.771	0.593	0.212	0.715	0.716	0.251
Wang et al. (PLS) [27]	0.766	0.674	0.214	0.718	0.718	0.250
Wang et al. (BPNN) [27]	0.816	–	0.192	0.740	0.738	0.247
Morrill and Byrd (MLR) [28]	0.8141	0.7951	–	0.6912	–	0.2854

chance correlations may in part be due to such an approach. The better results of the present models are attributed to the use of 3D descriptors which encode more structural information than the ETSI (one type of 2D descriptors).

3.3. Descriptor analysis and interpretation

The descriptors appearing in the present models encode three-dimensional aspects of the molecular structure, and can be classified as follows: (1) a geometrical descriptor: AROM; (2) two 3D-MoRSE descriptors: Mor29u and Mor02v; (3) a WHIM (Weighted Holistic Invariant Molecular) descriptor: Ds; and (4) six GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptors: HATS3u, HATS7m, R1m, R1m+, R6m+ and R1v.

Based on a previously described procedure [45], the relative contributions of the ten descriptors to the MLR and ANN models were determined and are plotted in Fig. 4. The significance of the descriptors involved in the MLR model decreases in the following order: R1m (12.4%) > R1v (10.8%) > R1m+ (10.2%) > HATS7m (10.1%) > R6m+ (9.9%) > HATS3u (9.4%) ≈ Mor02v (9.4%) > AROM (9.3%) ≈ Mor29u (9.3%) > Ds (9.2%). The significance of the descriptors in the ANN model decreases in the order: R1m (11.4%) > HATS7m (10.6%) > R1m+ (10.5%) ≈ R6m+ (10.5%) > R1v (10.3%) > AROM (10.0%) > Ds (9.9%) > Mor29u (9.8%) > Mor02v (9.1%) > HATS3u (7.8%). The most significant descriptor in both the MLR and ANN models is identical, that is, R1m. Although the order of the relative contribution from the other nine descriptors is different from each other in the two models, the individual contribution from all of these descriptors is very close. Thus, the contribution from these descriptors to both models can be regarded as similar. It should also be noted that the difference in descriptor contribution between any two descriptors used in the models is not significant, indicating that all of the descriptors are indispensable in generating the predictive models. Ten descriptors were needed in the QSPR models from a training set containing 127 samples, showing that the analyzed dataset is quite 'noisy', although it is not against the rule of thumb for building a linear model, that is, at least five data point per descriptor must exist in the model.

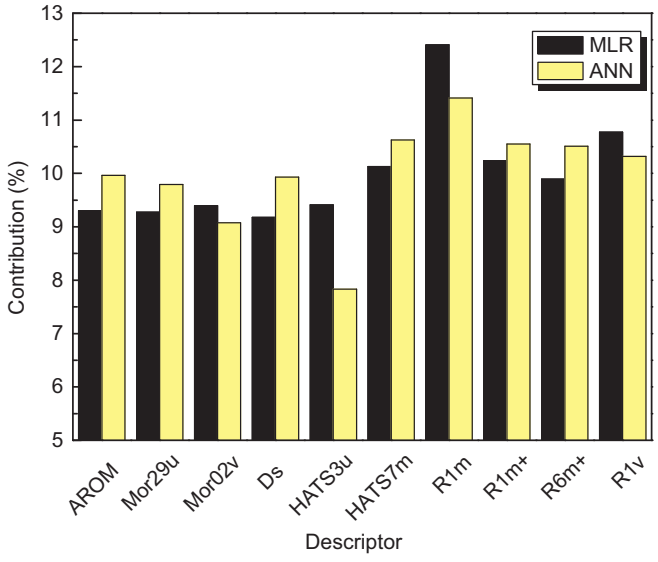

Fig. 4. Relative contributions of the descriptors to the proposed models.

The aromaticity (AROM) is derived from the general aromaticity indices defined by Bird [46] and is calculated as follows:

$$\mathrm{AROM} = 1 - \frac{100/35}{\bar{r}^{\pi}} \cdot \sqrt{\frac{\sum (r_{ij}^{\pi} - \bar{r}^{\pi})^2}{B_{\pi}}} \tag{8}$$

where the sum runs over the bonds belonging to aromatic rings, $\bar{r}^{\pi}$ is the π bond average length and r^{π} are the actual π bond lengths, B_{π} is the number of aromatic bonds. The AROM measures the degree of aromaticity of one compound. The positive sign of this descriptor in Eq. (7) indicates that increase in this descriptor increases the lgH_{50}.

3D-MoRSE descriptors are the 3D molecular representations of structure based on electron diffraction [47,48], which are calculated by summing atomic weights viewed by a different angular scattering function. The values of these descriptor functions are calculated at 32 evenly distributed values of scattering angle(s) in the range of 0–31 $Å^{-1}$ from the three dimensional atomic coordinates of a molecule. The 3D-MoRSE descriptor is calculated using following expression:

$$\mathrm{Morsw} = \sum_{i=i}^{nAT-1} \sum_{j=i+1}^{nAT} w_i \cdot w_j \frac{\sin(s \cdot r_{ij})}{s \cdot r_{ij}} \tag{9}$$

where s is the scattering angle, nAT is the number of atoms, r_{ij} is the interatomic distance between ith and jth atom, w is an atomic property, including atomic number, masses, van der Waals volumes, Sanderson electronegativities, and polarizabilities. The 3D-MoRSE (Mor29u, Mor02v) descriptors appearing in the model are important because they are related to bonds and distances (bond orders, saturation, and ratio of multiple bonds to single bonds).

WHIM descriptors [49,50] are based on principal component analysis of the weighted covariance matrix obtained from the atomic Cartesian coordinates. Ds is calculated by Eq. (10), where λ refers to eigenvalues of the weighted (by the electrotopological state indices) covariance matrix; t refers to atomic coordinates with respect to the principal axes.

$$\mathrm{Ds} = \sum_{k=1}^{3} \frac{\lambda_k^2 \cdot nAT}{\sum t_k^4} \tag{10}$$

Ds indicates a global measure of the degree of participation in intermolecular interactions. The sign of Ds in Eq. (7) is positive, meaning that the lgH_{50} increases when this descriptor increases.

GETAWAY descriptors [51,52] have been proposed as chemical structure descriptors derived from a new representation of

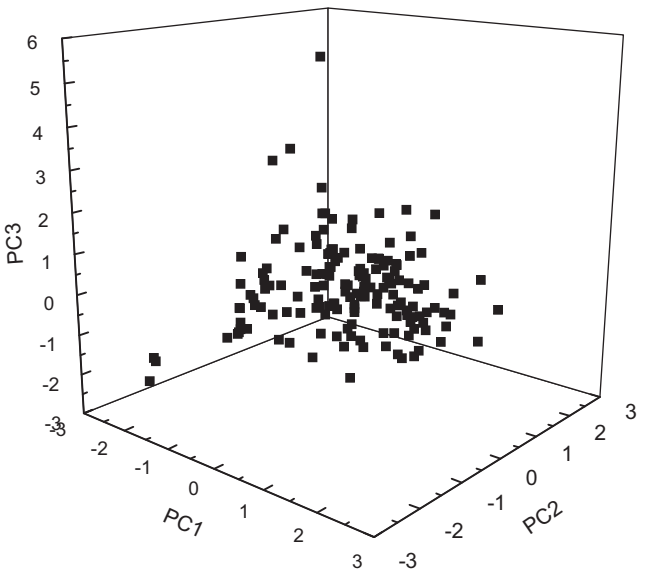

Fig. 5. PCA plots for the dataset.

molecular structure, the molecular influence matrix. These descriptors, as based on spatial autocorrelation, encode information on the effective position of substituents and fragments in the molecular space. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties. The GETAWAY descriptors involved are calculated as follows:

$$\mathrm{Rkw} = \sum_{i=1}^{nAT-1} \sum_{j>1} \frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{ij}} \cdot w_i \cdot w_j \cdot \delta(k, d_{ij}) \tag{11}$$

$$\mathrm{Rkw+} = \max_{ij} \left(\frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{ij}} \cdot w_i \cdot w_j \cdot \delta(k, d_{ij}) \right) \tag{12}$$

$$\mathrm{HATSkw} = \sum_{i=1}^{nAT-1} \sum_{j>1} (w_i \cdot h_{ii}) \cdot (w_j \cdot h_{jj}) \cdot \delta(k, d_{ij}) \tag{13}$$

where h_{ii} and h_{jj} are the leverages of ith and jth atom, d_{ij} is the topological distance, $\delta(k, d_{ij})$ is a Dirac-delta function ($\delta = 1$ if $d_{ij} = k$, zero otherwise).

PCA of the descriptor variables that had been used in construction of the final models were carried out in order to aid in further interpretation of these models. The first three PCs (Fig. 5) explain more than 70% of the total data variance, with a maximum of 44.9% on the first component. PC1 is highly correlated with R1m and Mor29u, PC2 is highly correlated with HATS3u and AROM, and PC3 is highly correlated with R1m+, respectively. Thus, the major factors determining the lgH_{50} are molecular size, shape and aromaticity degree.

3.4. Applicability domain

It needs to be pointed out that no matter how robust, significant and validated a QSPR model may be, it cannot be expected to reliably predict the modeled property for the entire universe of compounds. Therefore, before a QSPR model is put into use for screening compounds, its AD must be defined and predictions for only those compounds that fall in this domain can be considered as reliable. The model AD was analyzed in the Williams plot (shown in Fig. 6). Two compounds (Compound 5: 1-dinitromethyl-3-nitrobenzene and Compound 89: 1,3,3,5,5-pentanitropiperidine)

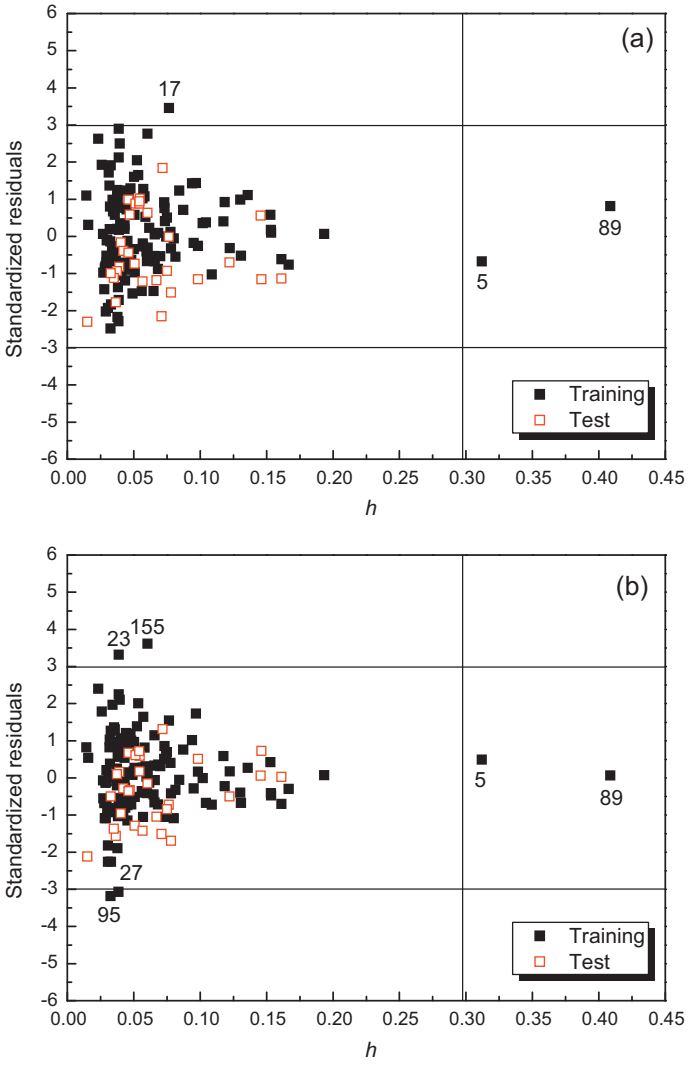

Fig. 6. Williams plots of the developed model: (a) MLR and (b) ANN.

in the training set were found to be an X outlier (with leverage value higher than the warning leverage limit of 0.2598). In Williams plot, X outliers can be explained as compounds with peculiar features poorly represented in the training set, which could affect the variable selection for a better modeling of those compounds. In this study, these two compounds were well predicted by the present models and thus reinforced the models.

Generally, if the residual value is larger than $\pm 3.0s$, the sample can be considered as response outlier (Y outlier), which could be associated with errors in the experimental values. On analyzing the AD of the developed models in the Williams plot, Compound 17: methylene-bis-(4,4,4-trinitrobutyramide)) could be recognized as a response outlier for the MLR model. There are four compounds (Compound 23: 2,4,6-trinitrobenzoic acid, Compound 27: 2′,2′,2′-trinitroethyl-2,4,6-trinitrobenzoate, Compound 95: 2,2-dinitropropyl-5,5,5-trinitro-2-nitrazapentanoate and Compound 155: 4,4,8,8-tetranitro-1,11-dinitro-6-nitrazaundecane) that could be considered as response outliers. However, the majority of the samples within the model AD were calculated accurately, which indicated further the reliability of the predictions.

Due to its high predictive ability, the proposed models could be used to screen existing databases or virtual chemical structures to identify organic compounds with desired impact sensitivity. In this

case, the AD will serve as a valuable tool to filter out "dissimilar" chemical structures.

4. Conclusions

In this paper, the QSPR method was employed to predict the impact sensitivity of nitro energetic compounds with 3D descriptors. A ten-parameter linear model was developed by MLR, with R^2 of 0.7684 and s of 0.194 for the training set. The nonlinear model built by ANN appeared to be more reliable than the linear model, where R^2 was 0.8481 and s was 0.164 for the training set. Satisfactory prediction results for the test set were obtained with both the linear and nonlinear models. The proposed models are predictive and could be used to estimate the impact sensitivity of new nitro energetic compounds because all the descriptors involved can be directly calculated from the molecular structure.

Acknowledgments

This work was supported by the Natural Science Foundation of China (No. 51003082) and the Educational Commission of Hubei Province (Q20101606). The authors gratefully wish to express their thanks to the reviewers for critically reviewing the manuscript and making important suggestions.

References

[1] P.-A. Perssen, R. Holmberg, J. Lee, Rock Blasting and Explosives Engineering, CRC Press, Baca Raton, FL, 1993.
[2] M.J. Kamlet, H.G. Adolph, The relationship of impact sensitivity with structure of organic high explosives, Propellants Explos. Pyrotech. 4 (1979) 30–34.
[3] P. Politzer, L. Abrahmsen, P. Sjoberg, Effects of amino and nitro substituents upon the electrostatic potential of an aromatic ring, J. Am. Chem. Soc. 106 (1984) 855–860.
[4] J.S. Murray, P. Lane, P. Politzer, P.R. Bolduc, A relationship between impact sensitivity and the electrostatic potentials at the midpoints of C—NO_2 bonds in nitroaromatics, Chem. Phys. Lett. 168 (1990) 135–139.
[5] P. Politzer, J.S. Murray, Relationships between dissociation energies and electrostatic potentials of C—NO_2 bonds: applications to impact sensitivities, J. Mol. Struct. 376 (1996) 419–424.
[6] P. Politzer, P. Lane, Comparison of density functional calculations of C-NO_2, N-NO_2 and C-NF_2 dissociation energies, J. Mol. Struct. 388 (1996) 51–55.
[7] B.M. Rice, J.J. Hare, A quantum mechanical investigation of the relation between impact sensitivity and the charge distribution in energetic molecules, J. Phys. Chem. A 106 (2002) 1770–1783.
[8] J. Fan, H.M. Xiao, The theoretical study on sensitivity and stability of polynitro arenas. I. Nitro derivatives of amino-benzenes, Acta Chim. Sin. 43 (1985) 14–18.
[9] J.F. Fan, H.M. Xiao, Theoretical study on pyrolysis and sensitivity of energetic compounds. 2. Nitro derivatives of benzene, J. Mol. Struct. (THEOCHEM) 365 (1996) 225–229.
[10] H.M. Xiao, J.F. Fan, Z.M. Gu, H.S. Dong, Theoretical study on pyrolysis and sensitivity of energetic compounds. 3. Nitro derivatives of amino-benzenes, Chem. Phys. 226 (1998) 15–24.
[11] J.F. Fan, Z.M. Gu, H.M. Xiao, H. Dong, Theoretical study on pyrolysis and sensitivity of energetic compounds. Part 4. Nitro derivatives of phenols, J. Phys. Org. Chem. 11 (1998) 177–184.
[12] C. Zhang, Y. Shu, Y. Huang, X. Wang, Theoretical investigation of the relationship between impact sensitivity and the charges of the nitro group in nitro compounds, J. Energ. Mater. 23 (2005) 107–119.
[13] C. Zhang, Y. Shu, Y. Huang, X. Zhang, H. Dong, Investigation of correlation between impact sensitivities and nitro group charges in nitro compounds, J. Phys. Chem. B 109 (2005) 8978–8982.
[14] C. Zhang, Y. Shu, X. Wang, X. Zhang, A new method to evaluate the stability of the covalent compound: by the charges on the common atomor group, J. Phys. Chem. A 109 (2005) 6592–6596.
[15] C. Cao, S. Gao, Two dominant factors influencing the impact sensitivities of nitrobenzenes and saturated nitro compounds, J. Phys. Chem. B 111 (2007) 12399–12402.
[16] M.H. Keshavarz, Prediction of impact sensitivity of nitroaliphatic, nitroaliphatic containing other functional groups and nitrate explosives, J. Hazard. Mater. 148 (2007) 648–652.
[17] W.-P. Lai, P. Lian, B.-Z. Wang, Z.-X. Ge, New correlations for predicting impact sensitivities of nitro energetic compounds, J. Energ. Mater. 28 (2010) 45–76.
[18] X.J. Yao, Y.W. Wang, X.Y. Zhang, R.S. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, Radial basis function neural network-based QSPR for the prediction of critical temperature, Chemom. Intell. Lab. Syst. 62 (2002) 217–225.
[19] J. Xu, B. Guo, B. Chen, Q. Zhang, A QSPR treatment for the thermal stabilities of second-order NLO chromophore molecules, J. Mol. Model. 12 (2005) 65–75.
[20] J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach, The Netherlands, 1999.
[21] M. Karelson, Molecular Descriptors in QSAR/QSPR, Wiley-Interscience, New York, 2000.
[22] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, 2nd ed., Wiley-VCH, Weinheim, 2009.
[23] J. Xu, H. Zhang, L. Wang, W. Ye, W. Xu, Z. Li, QSPR analysis of infinite dilution activity coefficients of chlorinated organic compounds in water, Fluid Phase Equilib. 291 (2010) 111–116.
[24] J. Xu, L. Wang, L. Wang, X. Shen, W. Xu, QSPR study of Setschenow constants of organic compounds using MLR, ANN, and SVM analyses, J. Comput. Chem. 32 (2011) 3241–3252.
[25] H. Nefati, J.-M. Cense, J.-J. Legendre, Prediction of the impact sensitivity by neural networks, J. Chem. Inf. Comput. Sci. 36 (1996) 804–810.
[26] S.G. Cho, K.T. No, E.M. Goh, J.K. Kim, J.H. Shin, Y.D. Joo, S. Seong, Optimization of neural networks architecture for impact sensitivity of energetic molecules, Bull. Korean Chem. Soc. 26 (2005) 399–408.
[27] R. Wang, J. Jiang, Y. Pan, H. Cao, Y. Cui, Prediction of impact sensitivity of nitro energetic compounds by neural network based on electrotopological-state indices, J. Hazard. Mater. 166 (2009) 155–186.
[28] J.A. Morrill, E.F.C. Byrd, Development of quantitative structure–property relationships for predictive modeling and design of energetic materials, J. Mol. Graph. Model. 27 (2008) 349–355.
[29] A. Golbraikh, A. Tropsha, Beware of q^2!, J. Mol. Graph. Model. 20 (2002) 269–276.
[30] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
[31] E. Estrada, E. Molina, 3D connectivity indices in QSPR/QSAR studies, J. Chem. Inf. Comput. Sci. 41 (2001) 791–797.
[32] HYPERCHEM, Version 6.01, Hypercube, Inc., Gainesville, 2000.
[33] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON for Windows (Software for Molecular Descriptor Calculations), Version 5.4, TALETE srl, Milan, 2006.
[34] H. Liu, P. Gramatica, QSAR study of selective ligands for the thyroid hormone receptor β, Bioorg. Med. Chem. 15 (2007) 5251–5261.
[35] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[36] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77.
[37] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, K.A. Prebble, Artificial neural networks in classification of NIR spectral data: design of the training set, Chemom. Intell. Lab. Syst. 33 (1996) 35–46.
[38] A.J. Holder, D.M. Yourtee, D.A. White, A.G. Glaros, R. Smith, Chain melting temperature estimation for phosphatidyl cholines by quantum mechanically derived quantitative structure property relationships, J. Comput. Aided Mol. Des. 17 (2003) 223–230.
[39] M.D. Wessel, P.C. Jurs, Prediction of reduced ion mobility constants from structural information using multiple linear regression analysis and computational neural networks, Anal. Chem. 66 (1994) 2480–2487.
[40] A. Atkinson, Plots, Transformations, and Regression, Clarendon Press, Oxford, UK, 1985.
[41] L.F. Ramsey, W.D. Schafer, The Statistical Sleuth, Wadsworth Publishing Company, USA, 1997.
[42] P.A. Jansson, Neural networks: an overview, Anal. Chem. 63 (1991) 357A–362A.
[43] L. Xu, J.W. Ball, S.L. Dixon, P.C. Jurs, Quantitative structure–activity relationships for toxicity of phenols using regression analysis and computational neural networks, Environ. Sci. Chem. 13 (1994) 841–851.
[44] Y.-H. Qi, Q.-Y. Zhang, L. Xu, Correlation analysis of the structures and stability constants of gadolinium(III) complexes, J. Chem. Inf. Comput. Sci. 42 (2002) 1471–1475.
[45] F. Zheng, E. Bayram, S.P. Sumithran, J.T. Ayers, C.-G. Zhan, J.D. Schmitt, L.P. Dwoskin, P.A. Crooks, QSAR modeling of mono- and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release, Bioorg. Med. Chem. 14 (2006) 3017–3037.
[46] C.W. Bird, The application of a new aromaticity index to six-membered ring heterocycles, Tetrahedron 48 (1986) 89–92.
[47] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, Chemical information in 3D space, J. Chem. Inf. Comput. Sci. 36 (1996) 1030–1037.
[48] J. Schuur, P. Selzer, J. Gasteiger, The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure–spectra correlations and studies of biological activity, J. Chem. Inf. Comput. Sci. 36 (1996) 334–344.
[49] R. Todeschini, M. Lasagni, E. Marengo, New molecular descriptors for 2D and 3D structures. Theory, J. Chemometr. 8 (1994) 263–272.
[50] R. Todeschini, P. Gramatica, R. Provenzani, E. Marengo, Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons, Chemom. Intell. Lab. Syst. 27 (1995) 221–229.
[51] V. Consonni, R. Todeschini, M. Pavan, Structure/response sorrelations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors, J. Chem. Inf. Comput. Sci. 42 (2002) 682–692.
[52] V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies, J. Chem. Inf. Comput. Sci. 42 (2002) 693–705.