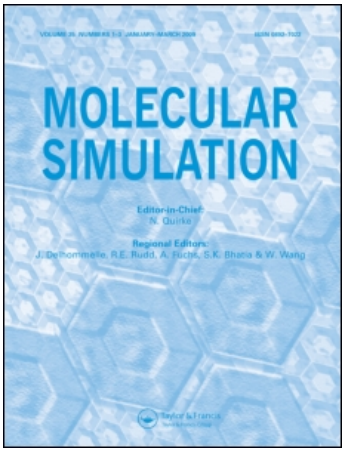


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

PLS and N-PLS-based MIA-QSTR modelling of the acute toxicities of phenylsulphonyl carboxylates to *Vibrio fischeri*

Mohammad Goodarzi[a]; Matheus P. Freitas[b]
[a] Department of Chemistry, Faculty of Sciences, and Young Researchers Club - Islamic Azad University, Arak, Markazi, Iran [b] Department of Chemistry, Federal University of Lavras, CP 3037, Lavras, MG, Brazil

Online publication date: 02 November 2010

To cite this Article Goodarzi, Mohammad and Freitas, Matheus P.(2010) 'PLS and N-PLS-based MIA-QSTR modelling of the acute toxicities of phenylsulphonyl carboxylates to *Vibrio fischeri*', Molecular Simulation, 36: 12, 953 — 959
To link to this Article: DOI: 10.1080/08927022.2010.492836
URL: http://dx.doi.org/10.1080/08927022.2010.492836

Molecular Simulation
Vol. 36, No. 12, October 2010, 953–959

Taylor & Francis
Taylor & Francis Group

PLS and N-PLS-based MIA-QSTR modelling of the acute toxicities of phenylsulphonyl carboxylates to *Vibrio fischeri*

Mohammad Goodarzi[a] and Matheus P. Freitas[b]*

[a]*Department of Chemistry, Faculty of Sciences, and Young Researchers Club – Islamic Azad University, Arak Branch, Arak, Markazi, Iran;* [b]*Department of Chemistry, Federal University of Lavras, CP 3037, 37200-000 Lavras, MG, Brazil*

(*Received 18 March 2010; final version received 3 May 2010*)

Descriptors based on multivariate image analysis have been used to derive predictive QSTR models of the acute toxicities of phenylsulphonyl carboxylates to *Vibrio fischeri*. Classical and multilinear partial least squares, PLS and N-PLS, respectively, were applied as regression methods, demonstrating similar predictive capability to each other. Model performance was improved in c. 10% after removing an outlier, and the results were in general agreement with the ones previously obtained through CoMFA and extended topochemical atom indices analysis. Overall, this study showed that a simple procedure is able to give highly predictive models, useful in ecotoxicology, independent of the regression method used for this class of compounds.

Keywords: MIA-QSTR; phenylsulphonyl carboxylates; *Vibrio fischeri*; PLS; N-PLS

1. Introduction

Vibrio fischeri is a gram-negative rod-shaped bacterium found globally in the marine environments. The bacterium is a key research organism for the examination of microbial bioluminescence, which is caused by transcription induced by population-dependent quorum sensing [1]. Because of this emission of the visible light, assays may be carried out for wastewater and environmental monitoring [2–5]. These tests rely on the measurement of the per cent inhibition in the light produced by the bacteria caused by the addition of the sample to the bacteria after an incubation period [6]. Thus, *V. fischeri* is a convenient probe to monitor toxicity of pollutants, such as phenylsulphonyl carboxylates. The potential hazard of phenylsulphonyl carboxylates to living species has been demonstrated, for example, to *Daphnia magna*, probably because of the endoergic effect of cavity formation and the exoergic hydrogen bonding between the solute as the acceptor and water as the donor [7].

The toxicities of a series of 56 phenylsulphonyl carboxylates have been modelled using extended topochemical atom (ETA) indices and genetic function approximation as a statistical tool [8], demonstrating a comparable predictive ability to the previous 3D-QSTR analysis (CoMFA) [9]. MIA-QSTR is another technique applied to predict toxicities, but for a series of organic compounds to *Pimephales promelas* [10]. In addition to QSTR studies, multivariate image analysis (MIA) descriptors have also been used in several QSAR studies [11–15], providing models with satisfactory predictive capability. Its main advantage over most of the structure-based methodologies lies on the no need for conformational screening and 3D alignment; it just requires a 2D alignment, which refers to simply superimpose 2D images–2D chemical structures drawn with the aid of an appropriate drawing program.

Moreover, in most of the 3D-QSAR studies, partial least squares (PLS) are the main regression method applied to correlate descriptors with the corresponding dependent variables. However, multilinear PLS (N-PLS) [16] is supposed to be superior to the unfolding PLS due to its simplicity (the number of variables can be effectively reduced) and predictive ability [17], though no important advantage of this method on traditional PLS has been found in an MIA-QSAR study of antifungal agents [18]. Thus, this work is devoted to building a predictive MIA-QSTR model from toxicity data of the literature [8], and comparing the abilities of prediction from PLS and N-PLS regressions.

2. QSAR modelling

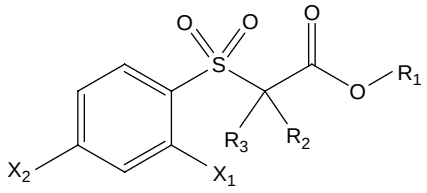
The data-set of 56 phenylsulphonyl carboxylates and the corresponding acute toxicity values (pC, C in mol l^{-1}) were obtained from the literature [8] (Table 1). The 2D chemical structures were built using the ChemSketch program [19], then aligned by a common pixel among them (Figure 1) in a defined workspace of dimension 400×250 pixels, and finally saved as bitmaps. The 56 images were read as double arrays in Matlab [20] and

*Corresponding author. Email: matheus@dqi.ufla.br

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927022.2010.492836
http://www.informaworld.com

Table 1. The series of 56 phenylsulphonyl carboxylates used in the MIA-QSTR analysis



.

Compound	R_1	R_2	R_3	X_1	X_2
1	CH_3	$-[CH_2]_2-$		H	H
2	CH_3	$-[CH_2]_3-$		H	H
3	CH_3	$-[CH_2]_4-$		H	H
4	CH_3	$-[CH_2]_5-$		H	H
5	CH_3	$-[CH_2]_2-$		H	NO_2
6	$CH(CH_3)_2$	$-[CH_2]_2-$		H	NO_2
7	$CH(CH_3)_2$	$-[CH_2]_3-$		H	NO_2
8	$CH(CH_3)_2$	$-[CH_2]_5-$		H	NO_2
9	$CH(CH_3)_2$	$-[CH_2]_6-$		H	NO_2
10	CH_3	$-[CH_2]_2-$		H	Br
11	CH_3	$-[CH_2]_3-$		H	Br
12	CH_3	$-[CH_2]_4-$		H	Br
13	CH_3	$-[CH_2]_5-$		H	Br
14	CH_3	$-[CH_2]_2-$		H	Cl
15	CH_3	$-[CH_2]_3-$		H	Cl
16	$CH(CH_3)_2$	$-[CH_2]_2-$		H	Cl
17	$CH_2(CH_2)_2CH_3$	$-[CH_2]_2-$		H	Cl
18	$CH(CH_3)_2$	$-[CH_2]_4-$		H	Cl
19	$CH(CH_3)_2$	$-[CH_2]_5-$		H	Cl
20	$CH(CH_3)_2$	$-[CH_2]_6-$		H	Cl
21	$CH(CH_3)_2$	$-[CH_2]_2-$		H	CH_3
22	$CH(CH_3)_2$	$-[CH_2]_3-$		H	CH_3
23	CH_3	$-[CH_2]_2-$		H	CH_3
24	CH_2CH_3	$-[CH_2]_2-$		H	CH_3
25	CH_2CH_3	$-[CH_2]_3-$		H	CH_3
26	$CH(CH_3)_2$	$-[CH_2]_4-$		H	CH_3
27	$CH(CH_3)_2$	$-[CH_2]_5-$		H	CH_3
28	CH_3	$-[CH_2]_5-$		H	CH_3
29	CH_3	H	H	H	NO_2
30	$CH(CH_3)_2$	H	H	H	NO_2
31	CH_3	H	H	Cl	NO_2
32	$CH(CH_3)_2$	H	H	Cl	NO_2
33	CH_3	H	H	NO_2	H
34	$CH(CH_3)_2$	H	H	NO_2	H
35	CH_3	H	H	NO_2	Cl
36	$CH(CH_3)_2$	H	H	NO_2	Cl
37	CH_3	H	CH_3	H	NO_2
38	CH_3	CH_3	CH_3	H	NO_2
39	CH_3	CH_2CH_3	CH_2CH_3	H	NO_2
40	CH_3	$CH_2(CH_2)_2CH_3$	$CH_2(CH_2)_2$	CH_3	H
41	CH_3	CH_2Ph	CH_2Ph	H	NO_2
42	CH_2CH_3	$CH_2(CH_2)_2CH_3$	$CH_2(CH_2)_2CH_3$	H	NO_2
43	CH_2CH_3	CH_3	CH_2Ph	H	NO_2
44	CH_2CH_3	CH_3	$CH_2CH{=}CH_2$	H	NO_2
45	CH_2CH_3	CH_3	CH_2-1-Naph	H	NO_2
46	$CH(CH_3)_2$	$CH_2(CH_2)_2CH_3$	$CH_2(CH_2)_2CH_3$	H	NO_2
47	Cyclohexyl	H	CH_3	H	NO_2
48	CH_3	H	$CH_2CO_2CH_2CH_3$	H	NO_2
49	$CH(CH_3)_2$	H	$CH_2CO_2CH(CH_3)_2$	H	NO_2
50	$CH(CH_3)_2$	$CH_2CO_2CH_2CH_3$	$CH_2CO_2CH_2CH_3$	H	NO_2
51	CH_3	$=$CHPh		H	NO_2
52	CH_2CH_3	$=$CHPh		H	NO_2

Table 1 – *continued*

Compound	R_1	R_2	R_3	X_1	X_2
53	$CH(CH_3)_2$		=CHPh	H	NO_2
54	$CH_2CH(CH_3)_2$		=CHPh	H	NO_2
55	$CH(CH_3)_2$		=CHPh	H	CH_3
56	$CH(CH_3)_2$		=CHPh	H	Cl

superimposed to give a $56 \times 400 \times 250$ three-way array. This three-mode array, in which descriptors state for pixels (binaries), was regressed against the **y** block (the toxicities column vector) through N-PLS. Subsequently, the three-way array was unfolded to give a $56 \times 100{,}000$ **X**-matrix, which was regressed against the acute toxicity values through PLS. The superposition of congruent structural scaffolds, the generation of the three-way array and the unfolding step are illustrated in Figure 1. The statistical parameters used to evaluate the model performances were the root mean square errors of calibration, leave-one-out cross-validation (LOOCV) and leave-20%-out cross-validation (RMSEC, RMSECV and $RMSECV_{20\%}$, respectively), and the squared correlation coefficients of the regression lines of experimental vs. fitted (r^2) and predicted (q^2 or $q^2_{20\%}$) acute toxicity values.

3. Results and discussion

The hydrophobicity of a compound can give scientists an indication of how easily a compound might be taken up in groundwater to pollute waterways, and its toxicity to animals and aquatic life [21]. This parameter is usually represented by the logarithm of the octanol–water partition coefficient (log *P*), which may be easily estimated through calculations. However, we have found that a simple correlation between calculated log *P* (obtained through the Molinspiration software [22]) and acute toxicities for the 56 phenylsulphonyl carboxylates results was very poor (Figure 2(a)), and thus hydrophobicity alone is not a suitable parameter to predict the toxicities of the title-based compounds. Therefore, MIA-QSTR arises as an alternative method to derive useful models without having to proceed with conformational screening and 3D alignment.

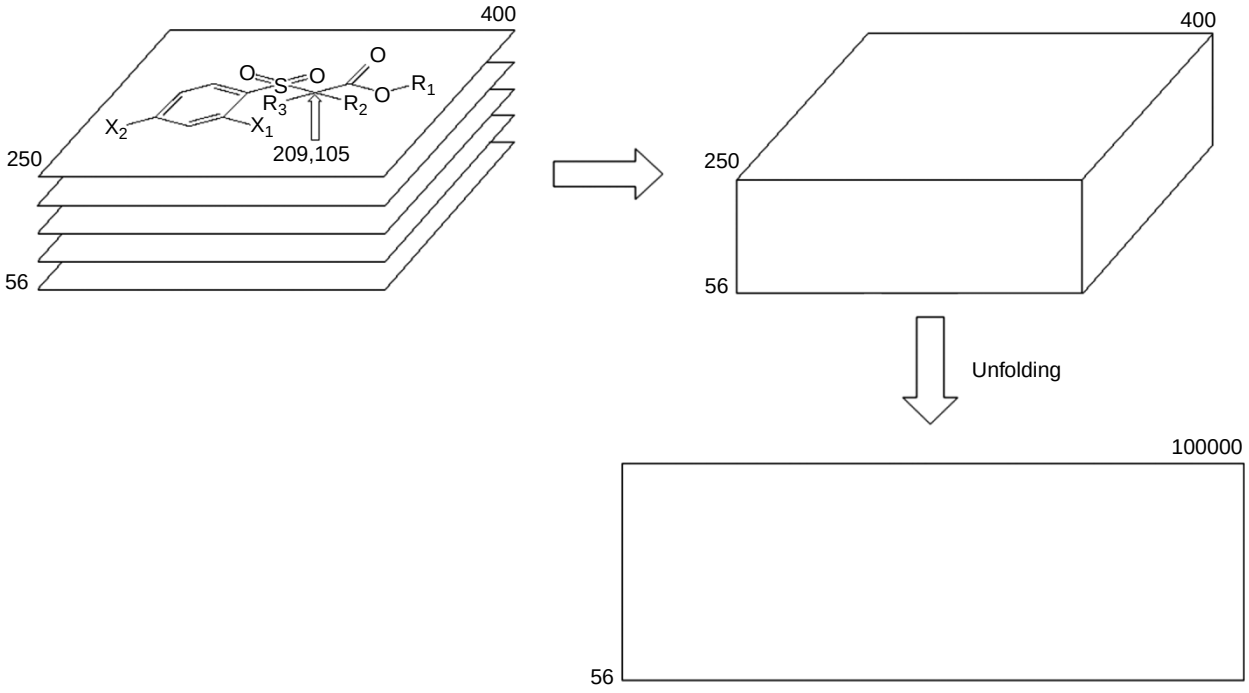


Figure 1. Image superposition, building of the three-way array (suitable to N-PLS regression), and unfolding to a two-way array (**X**-matrix, suitable to PLS regression). The arrow in the molecular structure indicates a pixel in common among the whole set of images (2D chemical structures) fitted at the 209,105 coordinated used for the 2D alignment step.

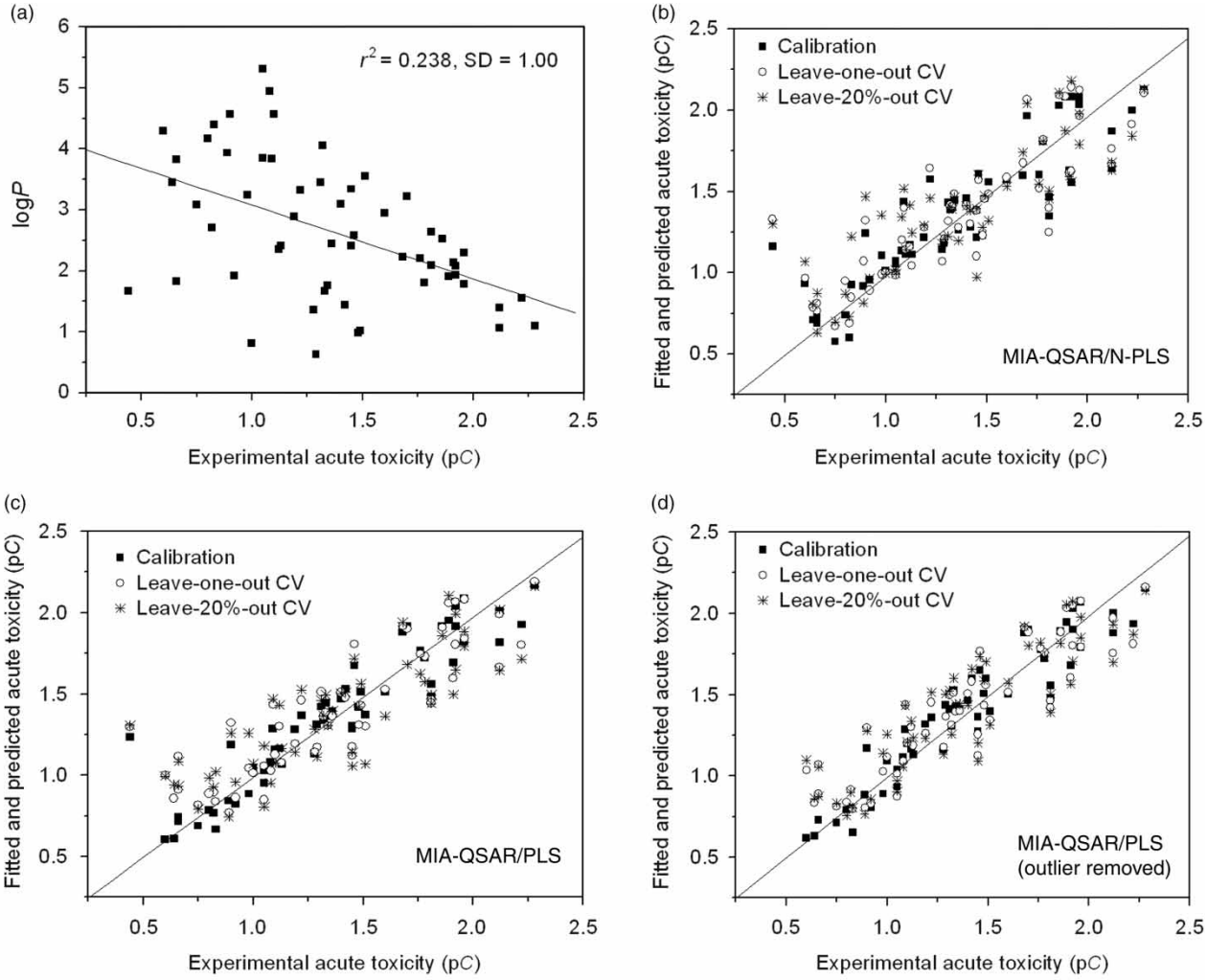


Figure 2. QSAR models: (a) correlation of acute toxicities (p*C*) vs. calculated log *P* and (b)–(d) plots of experimental vs. calibrated–predicted p*C* for the MIA-QSTR models built.

In a first approach, an MIA-QSTR model was built using N-PLS to correlate the three-way array (the descriptors block) with the acute toxicity values. The best number of latent variables was searched by analysing the minimum RMSECV, and three N-PLS components were found to be optimum. A reasonable r^2 of 0.787 (RMSEC = 0.21) was achieved, which is comparable to that found elsewhere [8], where r^2 of 0.820–0.865 was obtained using ETA indices. The N-PLS-based model was validated through LOOCV, in which 56 models were developed with one different prediction sample at a time; a q^2 of 0.715 (RMSECV = 0.24) was obtained. LOOCV has often been considered to be an incomplete validation method; external validation has been strongly recommended instead [23]. Randomly selected samples, 20% from the total series of 56 compounds, were also used as the external test set. Randomisation was performed 10 times, and an average $q^2_{20\%}$ was considered, i.e. 0.669 ($\text{RMSECV}_{20\%} = 0.26$). The estimation and prediction using MIA-QSTR/N-PLS are shown in Table 2 and illustrated in Figure 2(b)–(d).

The three-way array used for the N-PLS treatment was unfolded to a two-way array, an **X**-matrix of dimension $56 \times 100{,}000$ suitable to be regressed against the **y** block through classical (bilinear) PLS. The calibration using four PLS components gave an r^2 of 0.914 (RMSEC = 0.17), which is slightly superior to the correlation found in the literature [8] (Table 2 and Figure 2(b)–(d)). The calibration model was validated by LOOCV and leave-20%-out cross-validations, giving q^2 of 0.736 (RMSECV = 0.24) and $q^2_{20\%}$ of 0.671 ($\text{RMSECV}_{20\%} = 0.26$). However, compound **31** appeared as an outlier, since the residual in calibration for this sample was 4.6 times higher than RMSEC. The statistics of this model was improved after removing such outlier, to r^2 of 0.906 (RMSEC = 0.13), q^2 of 0.801

Table 2. Experimental, calibrated and predicted toxicity (p*C*) of phenylsulphonyl carboxylates to *V. fischeri*.

		MIA-QSTR/N-PLS			MIA-QSTR/PLS			MIA-QSTR/PLS (outlier removed)		
Compound	Exp.	Cal.	LOOCV	L-20%-OCV	Cal.	LOOCV	L-20%-OCV	Cal.	LOOCV	L-20%-OCV
1	2.28	2.13	2.11	2.13	2.17	2.19	2.16	2.15	2.16	2.14
2	2.12	1.64	1.67	1.63	2.01	1.99	2.02	2.00	1.97	1.93
3	1.91	1.63	1.61	1.59	1.69	1.60	1.50	1.68	1.60	1.56
4	1.81	1.47	1.40	1.45	1.56	1.47	1.49	1.56	1.46	1.50
5	2.12	1.87	1.76	1.68	1.82	1.66	1.64	1.88	1.75	1.70
6	1.78	1.81	1.82	1.81	1.72	1.73	1.57	1.72	1.76	1.73
7	1.81	1.35	1.25	1.50	1.49	1.44	1.44	1.48	1.42	1.39
8	1.45	1.22	1.10	0.97	1.28	1.12	1.06	1.27	1.12	1.09
9	1.05	1.05	1.00	0.99	1.03	0.85	0.81	1.04	0.87	0.90
10	1.89	2.08	2.08	1.87	1.95	2.06	2.11	1.94	2.03	2.05
11	1.76	1.60	1.52	1.55	1.77	1.74	1.62	1.78	1.78	1.82
12	1.60	1.57	1.59	1.53	1.51	1.53	1.36	1.50	1.52	1.57
13	1.31	1.43	1.32	1.22	1.42	1.51	1.46	1.41	1.50	1.52
14	1.96	2.04	2.12	1.98	2.09	2.08	1.88	2.07	2.08	1.98
15	1.92	1.56	1.62	1.56	1.91	1.80	1.65	1.90	1.80	1.70
16	1.86	2.03	2.09	2.11	1.92	1.91	1.86	1.89	1.89	1.81
17	1.70	1.96	2.07	2.04	1.92	1.90	1.68	1.90	1.88	1.80
18	1.51	1.56	1.49	1.32	1.37	1.30	1.07	1.40	1.34	1.31
19	1.32	1.39	1.41	1.40	1.34	1.36	1.37	1.31	1.31	1.25
20	0.90	1.24	1.32	1.47	1.19	1.32	1.26	1.17	1.30	1.28
21	1.96	2.08	1.97	1.79	1.82	1.84	1.79	1.79	1.79	1.85
22	1.46	1.61	1.57	1.60	1.67	1.81	1.72	1.65	1.77	1.73
23	2.22	2.00	1.91	1.84	1.93	1.80	1.71	1.93	1.81	1.87
24	1.92	2.08	2.14	2.18	2.04	2.07	1.99	2.03	2.05	2.07
25	1.68	1.60	1.68	1.74	1.88	1.92	1.94	1.88	1.91	1.92
26	1.22	1.58	1.64	1.46	1.37	1.46	1.52	1.36	1.45	1.51
27	1.09	1.44	1.40	1.52	1.29	1.43	1.47	1.29	1.44	1.43
28	1.40	1.46	1.41	1.41	1.47	1.51	1.50	1.46	1.51	1.43
29	1.29	1.18	1.21	1.21	1.31	1.17	1.11	1.43	1.36	1.50
30	1.28	1.15	1.07	1.18	1.13	1.14	1.28	1.16	1.18	1.13
31	0.44	1.16	1.33	1.30	1.23	1.30	1.31			
32	1.13	1.11	1.04	1.25	1.07	1.08	1.16	1.13	1.19	1.24
33	1.49	1.46	1.46	1.47	1.51	1.43	1.56	1.60	1.56	1.70
34	1.34	1.45	1.49	1.47	1.36	1.31	1.30	1.43	1.40	1.44
35	1.33	1.40	1.41	1.39	1.44	1.39	1.49	1.53	1.51	1.60
36	1.45	1.39	1.38	1.39	1.30	1.18	1.14	1.36	1.26	1.20
37	1.48	1.23	1.23	1.28	1.42	1.31	1.42	1.51	1.43	1.58
38	1.42	1.28	1.30	1.38	1.53	1.48	1.51	1.60	1.58	1.66
39	1.36	1.26	1.28	1.20	1.39	1.36	1.40	1.41	1.40	1.44
40	1.10	1.11	1.14	1.14	1.16	1.13	1.07	1.20	1.20	1.21
41	0.60	0.93	0.97	1.07	0.61	1.00	0.99	0.62	1.04	1.10
42	1.08	1.13	1.20	1.34	1.08	1.03	0.95	1.11	1.09	1.05
43	0.98	1.11	0.99	1.35	0.89	1.04	1.26	0.89	1.03	1.14
44	1.12	1.17	1.16	1.42	1.16	1.30	1.43	1.17	1.30	1.34
45	0.83	0.93	0.85	1.22	0.67	0.84	1.02	0.65	0.79	0.80
46	1.05	1.07	0.99	1.02	0.95	1.06	1.18	0.93	1.01	0.97
47	1.19	1.22	1.28	1.29	1.28	1.19	1.14	1.32	1.26	1.23
48	1.00	1.01	1.00	0.99	1.04	1.01	1.07	1.09	1.12	1.26
49	0.92	0.96	0.89	0.96	0.82	0.86	0.96	0.80	0.83	0.86
50	0.66	0.73	0.76	0.87	0.74	1.12	1.08	0.73	1.07	1.05
51	0.82	0.60	0.69	0.73	0.77	0.90	0.93	0.80	0.92	0.90
52	0.75	0.58	0.67	0.70	0.69	0.82	0.79	0.71	0.81	0.83
53	0.64	0.71	0.79	0.80	0.61	0.86	0.94	0.63	0.83	0.86
54	0.66	0.69	0.81	0.63	0.71	0.91	0.93	0.73	0.89	0.87
55	0.89	0.92	1.07	0.81	0.84	0.77	0.74	0.88	0.80	0.76
56	0.80	0.74	0.95	0.87	0.79	0.89	0.98	0.79	0.84	0.76

Table 3. Experimental, calibrated, cross-validated and predicted toxicity (p*C*) of phenylsulphonyl carboxylates to *V. fischeri* for the data-set split into training- and test-set compounds (outlier **31** removed)[a].

		MIA-QSTR/N-PLS				MIA-QSTR/PLS			
Compound	Exp.	Cal.	LOOCV	L-20%-OCV	Predicted	Cal.	LOOCV	L-20%-OCV	Predicted
1	2.28	2.09	2.08	2.07		2.07	2.00	2.03	
2[a]	2.12				1.69				1.93
3	1.91	1.57	1.51	1.45		1.64	1.53	1.59	
4	1.81	1.44	1.46	1.48		1.55	1.41	1.41	
5	2.12	1.89	1.97	1.78		1.90	1.78	1.78	
6	1.78	1.88	1.91	1.94		1.66	1.66	1.68	
7	1.81	1.43	1.38	1.34		1.52	1.36	1.39	
8[a]	1.45				1.20				1.19
9	1.05	1.06	1.06	1.06		1.13	1.31	1.38	
10	1.89	2.06	2.11	2.11		2.00	1.93	1.96	
11	1.76	1.63	1.69	1.65		1.76	1.74	1.73	
12[a]	1.60				1.49				1.60
13	1.31	1.37	1.49	1.37		1.48	1.47	1.49	
14[a]	1.96				2.04				1.92
15	1.92	1.63	1.67	1.71		1.77	1.73	1.73	
16	1.86	2.05	1.98	1.93		1.81	1.81	1.82	
17	1.70	1.99	1.98	1.98		1.83	1.84	1.84	
18	1.51	1.49	1.24	1.37		1.50	1.32	1.44	
19	1.32	1.39	1.39	1.29		1.35	1.31	1.33	
20[a]	0.90				1.27				1.29
21	1.96	2.06	1.87	1.82		1.82	1.71	1.71	
22	1.46	1.67	1.60	1.62		1.70	1.70	1.69	
23	2.22	2.08	2.06	1.99		1.86	1.80	1.80	
24[a]	1.92				2.06				1.98
25	1.68	1.66	1.57	1.68		1.84	1.81	1.84	
26	1.22	1.54	1.49	1.57		1.43	1.46	1.33	
27	1.09	1.41	1.42	1.37		1.32	1.49	1.46	
28	1.40	1.44	1.57	1.52		1.51	1.49	1.47	
29[a]	1.29				1.25				1.44
30	1.28	1.22	1.27	1.37		1.25	1.21	1.23	
32	1.13	1.15	1.17	1.33		1.22	1.24	1.20	
33	1.49	1.43	1.48	1.48		1.57	1.53	1.63	
34	1.34	1.44	1.42	1.42		1.41	1.37	1.48	
35	1.33	1.38	1.52	1.55		1.52	1.52	1.56	
36[a]	1.45				1.38				1.37
37	1.48	1.29	1.42	1.37		1.52	1.47	1.59	
38	1.42	1.35	1.52	1.49		1.62	1.61	1.63	
39	1.36	1.32	1.43	1.41		1.35	1.36	1.49	
40	1.10	1.13	1.18	1.22		1.27	1.32	1.32	
41	0.60	0.72	0.80	0.87		0.65	1.13	1.17	
42[a]	1.08				1.16				1.17
43	0.98	1.00	1.04	1.04		0.99	0.98	0.98	
44	1.12	1.21	1.55	1.53		1.27	1.32	1.30	
45	0.83	0.78	0.80	0.53		0.78	0.75	0.72	
46	1.05	1.10	1.00	0.97		1.06	1.01	1.01	
47	1.19	1.29	1.30	1.34		1.34	1.26	1.42	
48[a]	1.00				0.97				1.18
49	0.92	0.93	0.94	0.93		0.92	0.95	0.94	
50	0.66	0.73	0.73	0.87		0.86	1.08	1.09	
51	0.82	0.70	0.94	0.82		0.61	0.83	0.89	
52	0.75	0.69	0.65	0.65		0.53	0.73	0.68	
53[a]	0.64				0.81				0.77
54	0.66	0.74	0.74	0.75		0.50	0.83	0.82	
55	0.89	0.99	0.82	0.83		0.81	0.85	1.02	
56	0.80	0.81	0.74	0.75		0.74	0.89	0.96	

[a] Compounds pertain to the test set.

(RMSECV = 0.21) and $q^2_{20\%}$ of 0.764 ($RMSECV_{20\%} = 0.22$). These results are in agreement with the best results found through ETA [8] and CoMFA [9] analyses.

Also, the whole data-set (excepting outlier **31**) was in fact randomly split into training (80% of the whole set of compounds) and test sets, as depicted in Table 3, in order to give insight about a real external validation; three N-PLS components were found to be better, and r^2 of 0.863 and r^2_{test} of 0.795 were achieved. The corresponding test-set result for the PLS model was improved in comparison to the N-PLS one: the r^2 and r^2_{test} obtained were 0.871 and 0.868, respectively (using three PLS components). Additional statistic has been proposed by Roy and Roy [24] in order to test the external predictability, namely r^2_m, which is defined as $r^2[1-(r^2-r^2_0)^{1/2}]$, where r^2 and r^2_0 are the squared correlation coefficient values between observed and predicted values of the test-set compounds with and without the intercept, respectively. For a model with good external predictability, the r^2_m value should be greater than 0.5. The r^2_m values for the N-PLS and PLS models were similar, both for the training and test sets: 0.742 and 0.758 for the training set (N-PLS and PLS, respectively), and 0.584 and 0.581 for the test set (N-PLS and PLS, respectively). Therefore, these models are equally predictive according to this validation method.

Overall, we found that N-PLS and bilinear PLS are similarly predictive when applied to the MIA-QSTR analysis for a series of 56 phenylsulphonyl carboxylates. Also, this work was an attempt to show that a general statement that N-PLS is better than PLS in all QSARs is inadequate. In fact, the MIA-QSTR/N-PLS model was slightly more parsimonious than the PLS-based model (three N-PLS components used in the modelling using the whole data-set against four PLS components), but the predictive ability of both models were comparable to the available data from the literature, only requiring a modest computational investment and neither conformational screening nor 3D alignment rules to achieve reliable models to predict toxicities of compounds harmful to the environment.

Acknowledgements

The authors thank FAPEMIG for the financial support of this work. CNPq is also gratefully acknowledged for the fellowship (to M.P.F.).

References

[1] M. Madigan and J. Martinko, *Prokaryotic diversity: The bacteria,* in *Brock Biology of Microorganisms (11th ed.)*, M. Madigan and J. Martinko, eds., Prentice Hall, Upper Saddle River, 2005, pp. 329–418.

[2] M. Castillo, M.C. Alonso, J. Riu, M. Reinke, G. Klöter, H. Dizer, B. Fischer, B. Hansen, and D. Barceló, *Identification of cytotoxic compounds in European wastewaters during a field experiment*, Anal. Chim. Acta 426 (2001), pp. 265–277.

[3] D.J.B. Dalzell, S. Alte, E. Aspichueta, A. de la Sota, J. Etxebarria, M. Gutierrez, C.C. Hoffmann, D. Sales, U. Obst, and N. Christofi, *A comparison of five rapid direct toxicity assessment methods to determine toxicity of pollutants to activated sludge*, Chemosphere 47 (2002), pp. 535–545.

[4] M. Farré, G. Klöter, M. Petrovic, M.C. Alonso, M.J.L. de Alda, and D. Barceló, *Identification of toxic compounds in wastewater treatment plants during a field experiment*, Anal. Chim. Acta 456 (2002), pp. 19–30.

[5] V.L.K. Jennings, M.H. Rayner-Brandes, and D.J. Bird, *Assessing chemical toxicity with the bioluminescent photobacterium (*Vibrio fischeri*): A comparison of three commercial systems*, Water Res. 35 (2001), pp. 3448–3456.

[6] E.C. Faria, B.J.T. Brown, and R.D. Snook, *Water toxicity monitoring using* Vibrio fischeri*: A method free of interferences from colour and turbidity*, J. Environ. Monit. 6 (2004), pp. 97–102.

[7] Y.-B. He, L.-S. Wang, and Z.-T. Liu, *Acute toxicity of alkyl(1-phenylsulfonyl)cycloalkane–carboxylates to* Daphnia magna *and quantitative structure–activity relationships*, Chemosphere 31 (1995), pp. 2739–2746.

[8] K. Roy and G. Ghosh, *QSTR with extended topochemical atom indices. Part 5: Modeling of the acute toxicity of phenylsulfonyl carboxylates to* Vibrio fischeri *using genetic function approximation*, Bioorg. Med. Chem. 13 (2005), pp. 1185–1194.

[9] X. Liu, Z. Yang, and L. Wang, *CoMFA of the acute toxicity of phenylsulfonyl carboxylates to* Vibrio fischeri, SAR QSAR Environ. Res. 14 (2003), pp. 183–190.

[10] M.P. Freitas, *MIA-QSTR study of different organic compounds to* Phimephales promelas, Med. Chem. Res. 18 (2009), pp. 648–655.

[11] M.P. Freitas, S.D. Brown, and J.A. Martins, *MIA-QSAR: A simple 2D image-based approach for quantitative structure–activity relationship analysis*, J. Mol. Struct. 738 (2005), pp. 149–154.

[12] M.P. Freitas, *MIA-QSAR modelling of anti-HIV-1 activities of some 2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners*, Org. Biomol. Chem. 4 (2006), pp. 1154–1159.

[13] M.P. Freitas, *Multivariate QSAR: From classical descriptors to new perspectives*, Curr. Comput. Aided Drug. Des. 3 (2007), pp. 235–239.

[14] J.R. Pinheiro, M. Bitencourt, M.P. Freitas, E.F.F. da Cunha, and T.C. Ramalho, *Novel anti-HIV cyclotriazadisulfonamide derivatives as modeled by ligand- and receptor-based approaches*, Bioorg. Med. Chem. 16 (2008), pp. 1683–1690.

[15] J.E. Antunes, M.P. Freitas, E.F.F. da Cunha, T.C. Ramalho, and R. Rittner, In silico *prediction of novel phosphodiesterase type-5 inhibitors derived from sildenafil, vardenafil and tadalafil*, Bioorg. Med. Chem. 16 (2008), pp. 7599–7606.

[16] R. Bro, *Multiway calibration. Multilinear PLS*, J. Chemom. 10 (1996), pp. 47–61.

[17] M.M.C. Ferreira, *Multivariate QSAR*, J. Braz. Chem. Soc. 13 (2002), pp. 742–753.

[18] M. Bitencourt and M.P. Freitas, *Bi- and multilinear PLS coupled to MIA-QSAR in the prediction of antifungal activities of some benzothiazole derivatives*, Med. Chem. 5 (2009), pp. 5, 79–86.

[19] *ACD/ChemSketch Version 10.02*, Advanced Chemistry Development, Inc., Toronto, 2006.

[20] *Matlab Version 7.5*, MathWorks Inc., Natick, 2007.

[21] D. Cronin and T. Mark, *The role of hydrophobicity in toxicity prediction*, Curr. Comp. Aided Drug Des. 2 (2006), pp. 405–413.

[22] *Molinspiration cheminformatics*, Bratislava, Slovak Republic (http://www.molinspiration.com).

[23] A. Golbraikh and A. Tropsha, *Beware of q^2!*, J. Mol. Graph. Model. 20 (2002), pp. 269–276.

[24] P.P. Roy and K. Roy, *On some aspects of variable selection for partial least squares regression models*, QSAR Comb. Sci. 27 (2008), pp. 302–313.