CODEN IEJMAT *Internet Electronic Journal of Molecular Design* **2006**, *5*, 116–134 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

February 2006, Volume 5, Number 2, Pages 116–134

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Danail Bonchev on the occasion of the 65th birthday

QSTR Study on Aquatic Toxicity Against *Poecilia reticulata* and *Tetrahymena pyriformis* Using Topological Indices

Adina Costescu and Mircea V. Diudea

Faculty of Chemistry and Chemical Engineering, "Babes–Bolyai" University, 400028 Cluj, Romania

Received: November 3, 2005; Revised: February 1, 2006; Accepted: February 6, 2006; Published: February 28, 2006

Citation of the article:
A. Costescu and M. V. Diudea, QSTR Study on Aquatic Toxicity Against *Poecilia reticulata* and *Tetrahymena pyriformis* Using Topological Indices, *Internet Electron. J. Mol. Des.* **2006**, *5*, 116–134, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

QSTR Study on Aquatic Toxicity Against *Poecilia reticulata* and *Tetrahymena pyriformis* Using Topological Indices

Adina Costescu and Mircea V. Diudea*

Faculty of Chemistry and Chemical Engineering, "Babes–Bolyai" University, 400028 Cluj, Romania

Received: November 3, 2005; Revised: February 1, 2006; Accepted: February 6, 2006; Published: February 28, 2006

Internet Electron. J. Mol. Des. **2006**, *5* (2), 116–134

Abstract

Motivation. Considering the importance of quantitative structure–toxicity relationship (QSTR) studies in the field of aquatic toxicology from the viewpoint of ecological safety assessment, fish and algae toxicity of various aromatic compounds has been modeled by the multiple regression technique. Topological indices are used to relate the chemical structures to their toxic activity against aquatic organisms. In our experiment we do not look for the best model for a property on a given set of molecules, but for the best set of descriptors modeling a given functional group.

Method. Correlation analysis and Multiple Linear Regression (MLR) have been carried out to derive the best QSAR models, giving important information on functional groups by topological descriptors. The models belonging to benzene derivative toxicity against *Poecilia reticulata* were validated with an external validation set, and the indices in the best models were used to predict the toxicity for other two sets of aromatic compounds against *Tetrahymena pyriformis*. The robustness and prediction power of each model was evaluated by a leave–half–out (LHO) procedure.

Results. The classification performance of topological descriptor models generated with multiple regressions show that the statistical results depend strongly on the functional groups description. The basic set, toxicity data of 92 diverse aromatic compounds against *Poecilia reticulata*, provided very good results witch were validated more then 93%. For the both prediction sets against *Tetrahymena pyriformis* the best models are: $R = 0.98$; $R^2 = 0.96$; $Q^2 = 0.94$; $F = 188.36$; $s = 0.13$ (the set with 37 nitrobenzene derivatives) and $R = 0.92$; $R^2 = 0.84$; $Q^2 = 0.82$; $F = 198.64$; $s = 0.2$ (for 167 aromatic compounds).

Conclusions. Our statistical study demonstrated that topological indices based on Cluj matrices show a good predictive ability of the aquatic toxicity against *Poecilia reticulata* and *Tetrahymena pyriformis*. These indices account for molecular bulk, halogen, nitro and amino substitutions in the benzene ring. The four descriptors that describe the equation of nitrobenzene derivative are very useful to predict the activity of other compounds containing the nitro functional group. It appears that topological descriptors have significant potential in QSTR studies, which warrants extensive evaluation. This procedure can be used to approach the aquatic toxicity and to select the appropriate model for new chemical compounds.

Keywords. QSAR; quantitative structure–activity relationships; QSTR; quantitative structure–toxicity relationships; aquatic toxicity; topological indices; aromatic compounds; benzene derivatives; *Poecilia reticulata*; *Tetrahymena pyriformis*.

Dedicated on the occasion of the 65th birthday to Danail Bonchev.

* Correspondence author; E–mail: diudea@chem.ubbcluj.ro.

1 INTRODUCTION

Toxicity of organic compounds is one of the particularly interesting biological activities in the scientific community due to its impact on environment and human health. The term "toxicity" defines just a biological end–point, including several different mechanisms at the molecular level. Data sets usually consist of compounds acting by different mechanisms making the use of linear models questionable. The use of short–term toxicity assays and computational models over their traditional counterparts is preferred for many reasons including the ease of use, speed and relatively low cost. In addition, they pose fewer problems with regard to the stability of chemicals and potential losses during the testing due to volatility. The acute toxicity, assessed in short and low–cost unicellular tests, is also considered to be a surrogate for the prediction of toxicity to higher aquatic organisms [1].

Chemical byproducts from industrial systems that are allowed to escape into the environment can have toxic effects. Each of these chemicals can be harmful, and it is crucial that each compound be assessed for its toxicity level. However, this can be costly, time–consuming, and could potentially produce toxic side products from the experimental methods used today [2].

Recently, computational methods have been used to solve complex problems in many aspects of science. One particularly useful method, the development of quantitative structure activity relationships (QSAR) [3], has found diverse applications in chemistry. These applications include biological activity (QSAR) prediction [4–6], physical property (QSPR), prediction [6–8], and toxicity (QSTR) prediction [9–12]. QSAR has great advantages over both experimental techniques and other computational methods. First, QSAR is a purely computational method that does not require the use of expensive equipment or hazardous chemicals. Second, QSAR has the advantage of being computationally inexpensive, as compared with molecular dynamics, Monte Carlo, and *ab initio* quantum mechanical methods.

QSAR approach is based on the assumption that the structure of a molecule must contain the features responsible for its physical, chemical, and biological properties and on the possibility of representing a molecule by numerical descriptors. The difficulty of predicting toxicity is due to the lack of knowledge of the toxic mechanisms for reactive chemicals and the complexity and heterogeneity of the data available. More powerful computational approaches have now opened new avenues to QSAR studies and several research papers have been published [1–9]. While toxicity databases on a number of fish species, including the guppy (*Poecilia reticulata*) exist, the largest and most chemically diverse of the fish data sets is that for 96–h flow through fathead minnow (*Pimephales promelas*) 50% mortality [13].

The aim of this study, therefore, is to evaluate a series of aromatic compounds, such as phenols, anilines, nitrobenzenes and other poly–substituted benzenes *vs*. their toxicity on the above fish populations. Substituted benzenes, including anilines and phenols, represent a significant portion of

the chemicals that are of environmental concern. Their toxic potency may vary, however, in a wide range. According to Fish Acute Toxicity Syndromes [14] for fathead minnows, substituted benzenes can act as nonpolar narcotics, polar narcotics, uncouplers of oxidative phosphorylation, or even invoke reactive mode of toxic action.

Nitroaromatic compounds form an important class of industrial chemicals with substantial marketing volumes and a diverse range of use patterns [15]. Besides applications as solvents, they are used for the synthesis of dyestuffs, urethane polymers, and other plastics as well as of anilines, and among derivative products are also insecticides, herbicides, and pharmaceuticals. The latter is of ongoing interest also in an effort to better understand the mutagenicity of ambient air, pointing to an important aspect of the toxicological profile of this class of compound [16–17]. Moreover, there is evidence that nitrophenols are formed photochemically from aromatic precursor compounds in rain, which was discussed in the context of xenobiotics contributing to forest decline [18]. From the chemical viewpoint, the nitro group is a strong *ð*–electron acceptor, lowering the electron density of the aromatic ring. Inside the nitro group, excess electronic charge is mainly localized at the oxygen atoms, while the nitrogen atom is typically electron–deficient.

As a consequence, nitroaromatic compounds show enhanced reactivity for the attack of nucleophiles at aromatic ring carbons as well as for reactions with reducing agents. In phenol derivatives the nitro group leads to a pronounced enhancement of the acidity of the OH group. The presence of nitroaromatics in aquatic systems has led to various studies on the associated hazard potential. Many studies have concentrated on investigating the acute toxicity toward fish [19–20] and other aquatic species [21], including the QSARs for elucidating underlying modes of action and their link to characteristics of the molecular structure of these compounds. Excess toxicity of bioreactive compounds can be identified by its upward deviation from hydrophobicity–based QSARs, and there are various molecular descriptors available for modeling the reactivity profile of the chemicals, which allows a mechanistic interpretation of specific toxicity effects as related to metabolic pathways and chemical interactions with endogenous macromolecules [20,21].

Different QSAR models, descriptors and statistical methods have been applied to model toxicity data of diverse chemicals by different group of workers. Recently, Rose and Hall [22] have published a paper on E–state modeling of fish toxicity independent of 3D structure information. QSARs were developed for the prediction of aqueous toxicities for *Poecilia reticulata* using CODESSA treatment by Katritzky *et al*. [23]. Optimization of correlation weights of local graph invariants was used by Toropov *et al*. [24] to predict the aquatic toxicity. Seward *et al*. [25] have reported toxicity modeling of aliphatic carboxylic acids and salts to *Tetrahymena pyriformis* using physicochemical and quantum mechanical parameters. Roy and Ghosh [26] have modeled toxicity of substituted phenols against *Tetrahymena pyriformis* and fish toxicity data of substituted benzenes against *Poecilia reticulata* [27] to explore the suitability of the newly developed extended

topochemical atom (ETA) indices in modeling studies.

Table 1. Observed [22] fish toxicity values of substituted benzenes

No	Compounds	$pLC_{50\ obs}$	No	Compounds	$pLC_{50\ obs}$
1	phenol	3.45	**47**	nitrobenzene	2.97
2	2–methylphenol	3.77	**48**	2–nitrotoluene	3.59
3	3–methylphenol	3.48	**49**	3–nitrotoluene	3.65
4	4–methylphenol	3.74	**50**	4–nitrotoluene	3.67
5	2,4–dimethylphenol	3.86	**51**	2,3–dimethylnitrobenzene	4.39
6	2,6–dimethylphenol	3.75	**52**	3,4–dimethylnitrobenzene	4.21
7	3,4–dimethylphenol	3.92	**53**	2–chloronitrobenzene	3.72
8	2,3,6–trimethylphenol	4.21	**54**	3–chloronitrobenzene	4.01
9	4–Ethylphenol	4.07	**55**	4–chloronitrobenzene	4.42
10	4–propylphenol	4.09	**56**	2,3–dichloronitrobenzene	4.66
11	4–butylphenol	4.47	**57**	2,4–dichloronitrobenzene	4.46
12	4–*tert*–butylphenol	4.46	**58**	2,5–dichloronitrobenzene	4.59
13	2–*tert*–butyl–4–methylphenol	4.90	**59**	3,5–dichloronitrobenzene	4.58
14	4–pentylphenol	5.12	**60**	2–chloro–6–nitrotiluene	4.52
15	4–*tert*–pentylphenol	4.81	**61**	4–chloro–2–nitrotoluene	4.44
16	2–allylphenol	3.96	**62**	aniline	2.91
17	2–phenylphenol	4.76	**63**	2–methylaniline	3.12
18	1–naphthol	4.50	**64**	3–methylaniline	3.47
19	4–chlorophenol	4.18	**65**	4–methylaniline	3.72
20	4–chloro–3–methylphenol	4.33	**66**	*N*,*N*–dimethylaniline	3.33
22	3–methoxyphenol	3.22	**68**	3–ethylaniline	3.65
23	4–methoxyphenol	3.05	**69**	4–ethylaniline	3.52
24	4–phenoxyphenol	4.58	**70**	4–butylaniline	4.16
25	quinoline	3.63	**71**	2,6–diisopropylaniline	4.06
26	chlorobenzene	3.77	**72**	2–chloroaniline	4.31
27	1,2–dichlorobenzene	4.40	**73**	3–chloroaniline	3.98
28	1,3–dichlorobenzene	4.28	**74**	4–chloroaniline	3.67
29	1,4–dichlorobenzene	4.56	**75**	2,4–dichloroaniline	4.41
30	1,2,3–trichlorobenzene	4.89	**76**	2,5–dichloroaniline	4.99
31	1,2,4–trichlorobenzene	4.83	**77**	3,4–dichloroaniline	4.39
32	1,3,5–trichlorobenzene	4.47	**78**	3,5–dichloroaniline	4.62
33	1,2,3,4–tetrachlorobenzene	5.35	**79**	2,3,4–trichloroaniline	5.15
34	1,2,3,5–tetrachlorobenzene	5.43	**80**	2,3,6–trichloroaniline	4.73
35	1,2,4,5–tetrachlorobenzene	5.85	**81**	2,4,5–trichloroaniline	4.92
36	3–chlorotoluene	3.84	**82**	ααα–4–tetrafluoro–3–methylaniline	3.77
37	4–chlorotoluene	4.33	**83**	ααα –4–tetrafluoro–2–methylaniline	3.78
38	2,4–dichlorotoluene	4.54	**84**	pentafluoroaniline	3.69
39	2,4,5–trichlorotoluene	5.06	**85**	2–nitroaniline	4.15
40	3,4,5–trichlorotoluene	4.60	**86**	3–nitroaniline	3.24
41	pentachlorotoluene	6.15	**87**	4–nitroaniline	3.23
42	benzene	3.09	**88**	2–chloro–4–nitrianiline	3.93
43	toluene	3.13	**89**	4–bromoaniline	3.56
44	2–xylene	3.48	**90**	3–benzyloxyaniline	4.34
45	3–xylene	3.45	**91**	4–hexyloxyaniline	4.78
46	4–xylene	3.48	**92**	4–ethoxy–2–nitroaniline	3.85

In the present work we modeled the toxicity of benzene derivatives against *Poecilia reticulata* (taken from Rose *et al*. [22]) using topological descriptors by multiple regression technique. The best found relations were used to predict the toxicity of nitrobenzene derivatives against *Tetrahymena pyriformis* [28] and the toxicity of aromatic compounds against *Tetrahymena pyriformis* [29].

2 MATERIALS AND METHODS

2.1 Chemical Data

In our present QSAR study, fish toxicity data of 92 diverse aromatic compounds against *Poecilia reticulate* [22] have been modeled by using topological parameters and the multiple regression technique. The data set is chemically heterogeneous and includes phenols, anilines, nitrobenzenes, as well as compounds with more than one functional group on the benzene ring (Table 1). The data set does not include chemicals that are anticipated to elicit their toxicity via target–specific mechanisms such as enzyme inhibition.

Table 2. Observed [28] nitrobenzene toxicity values to *Tetrahymena pyriformis*

No	Compound	pLC_{50}	No	Compound	pLC_{50}
1	2,6–Dimethylnitrobenzene	0.30	**20**	6–Bromo–1,3–nitrobenzene	2.31
2	2,3–Dimethylnitrobenzene	0.56	**21**	3–Bromonitrobenzene	1.03
3	2–Methyl–3–chloronitrobenzene	0.68	**22**	2,4,6–Trimethylnitrobenzene	0.86
4	2–Methylnitrobenzene	0.05	**23**	5–Methyl–1,2–dinitrobenzene	1.52
5	2–Chloronitrobenzene	0.68	**24**	2,4–Dichloronitrobenzene	0.99
6	2–Methyl–5–chloronitrobenzene	0.82	**25**	3,5–Dichloronitrobenzene	1.13
7	2,4,5–Trichloronitrobenzene	1.53	**26**	6–Iodo–1,3–dinitrobenzene	2.12
8	2,5–Dichloronitrobenzene	1.13	**27**	2,3,4,5–tetrachloronitrobenzene	1.78
9	6–Chloro–1,3–dinitrobenzene	1.98	**28**	2,3–Dichloronitrobenzene	1.07
10	Nitrobenzene	0.14	**29**	2,5–Dibromonitrobenzene	1.37
11	3–Methylnitrobenzene	0.05	**30**	1,2–Dichloro–4,5–dinitrobenzene	2.21
12	3,4–Dichloronitrobenzene	1.16	**31**	3–Methyl–4–bromonitrobenzene	1.16
13	4–Methylnitrobenzene	0.17	**32**	2,3,4–Trichloronitrobenzene	1.51
14	1,4–Dinitrobenzene	1.30	**33**	2,4,6–Trichloronitrobenzene	1.43
15	4–Chloronitrobenzene	0.43	**34**	4,6–Dichloro–1,2–dinitrobenzene	2.42
16	2,3,5,6–Tetrachloronitrobenzene	1.82	**35**	2,4,6–Trichloro–1,4–dinitrobenzene	2.19
17	3–Chloronitrobenzene	0.73	**36**	2,3,5,6–Tetrachloro–1,4–dinitrobenzene	2.74
18	1,2–Dinitrobenzene	1.25	**37**	2,4,6–Trichloro–1,3–dinitrobenzene	2.59
19	2–Bromonitrobenzene	0.75			

A problem in computational chemistry pertains to the dependent variable in the QSAR. To make any sense at all, the left–hand term in the QSAR (dependent variable) must be uniform. Often researchers take great satisfaction in demonstrating how many molecules can be covered by a correlation equation. As some workers demonstrated [26] with a large set of phenols, it makes no sense to lump compounds together unless it can be shown that they are all acting by the same mechanism.

Data set representing several mechanisms of toxic action was divided by functional group in four sets: phenol substitutes, benzene substitutes, nitrobenzene substitutes and aniline substitutes. In an effort to further demonstrate the utility of the successful indices, models for two other toxicity data sets [28–29] were calculated, by using the same set of descriptors.

Table 3. Observed [29] aromatic compounds toxicity values to *Tetrahymena pyriformis*

No	Compound	$pIGC_{50}$	No	Compound	$pIGC_{50}$
1	4–cyanopyridine	–0.82	**82**	1–cyanonaphthalene	0.69
2	2–cyanopyridine	–0.79	**83**	ethyl–4–nitrobenzoate	0.71
3	3–cyanopyridine	–0.74	**84**	4–methyl–3–nitrophenol	0.74
4	benzonitrile	–0.52	**85**	2–chloro–4–nitroaniline	0.75
5	2–hydroxy–4–methyl–3–nitropyridine	–0.50	**86**	4,5–difluoro–2–nitroaniline	0.75
6	3–cyanoaniline	–0.47	**87**	2–chloromethyl–4–nitrophenol	0.75
7	4–cyanobenzamide	–0.38	**88**	4–ethoxy–2–nitroaniline	0.76
8	4–acetylbenzonitrile	–0.37	**89**	3–chloro–4–fluoronitrobenzene	0.80
9	1,2–dicyanobenzene	–0.34	**90**	2–chloro–5–nitropyridine	0.80
10	2–cyanobenzamide	–0.32	**91**	5–chloro–2–methylnitrobenzene	0.82
11	4–fluorobenzonitrile	–0.26	**92**	4–nitrophenetole	0.83
12	3–tolunitrile	–0.25	**93**	3–chloronitrobenzene	0.84
13	2–tolunitrile	–0.24	**94**	2,6–dinitroaniline	0.84
14	4–tolunitrile	–0.10	**95**	2–bromonitrobenzene	0.86
15	3–chlorobenzonitrile	–0.06	**96**	2,4,6–trimethylnitrobenzene	0.86
16	methyl–4–cyanobenzoate	–0.06	**97**	6–methyl–1,3–dinitrobenzene	0.87
17	3–cyanophenol	–0.06	**98**	3–hydroxy–2–nitropyridine	0.87
18	3–cyanobenzaldehyde	–0.02	**99**	2–chloro–3–nitropyridine	0.87
19	2–amino–3–nitropyridine	–0.01	**100**	1,3–dinitrobenzene	0.89
20	2–methoxy–2–nitropyridine	–0.01	**101**	3–fluoro–4–nitrophenol	0.93
21	4–chlorobenzonitrile	0.00	**102**	3,5–dinitroaniline	0.94
22	3–nitroaniline	0.03	**103**	2,5–dinitrophenol	0.95
23	2–cyanophenol	0.04	**104**	4–amino–2–nitrophenol	0.98
24	4–cyanobenzaldehyde	0.04	**105**	2,4–dichloronitrobenzene	0.99
25	3–methoxybenzonitrile	0.05	**106**	1–nitronaphthalene	1.00
26	2–methylnitrobenzene	0.05	**107**	2–methyl–1–nitronaphthalene	1.04
27	3–methylnitrobenzene	0.05	**108**	2,3–dichloronitrobenzene	1.07
28	4–methoxybenzonitrile	0.10	**109**	2–bromo–5–nitropyridine	1.07
29	4–nitrobenzyl alcohol	0.12	**110**	1–fluoro–3–iodo–5–nitrobenzene	1.09
30	4–nitrophenylacetonitrile	0.13	**111**	3,4–dinitrobenzyl alcohol	1.09
31	3–nitrobenzaldehyde	0.14	**112**	2,4–dinitrophenol	1.10
32	nitrobenzene	0.14	**113**	4–nitro–1–naphthylamine	1.12
33	2–nitrobenzaldehyde	0.17	**114**	5–fluoro–2–nitrophenol	1.12
34	4–methylnitrobenzene	0.17	**115**	2,5–dichloronitrobenzene	1.13
35	4–nitrobenzamide	0.18	**116**	3,5–dichloronitrobenzene	1.13
36	4–nitrobenzaldehyde	0.20	**117**	3,4–dichloronitrobenzene	1.16
37	1–fluoro–3–nitrobenzene	0.20	**118**	2–bromo–5–nitrotoluene	1.16
38	2–amino–5–nitropyridine	0.22	**119**	2–amino–4–chloro–5–nitrophenol	1.18
39	1–fluoro–2–nitrobenzene	0.23	**120**	3,5–dinitrobenzonitrile	1.22
40	4–cyanoaniline	0.24	**121**	2–chloro–4,6–dinitroaniline	1.22
41	4–fluoronitrobenzene	0.25	**122**	3–bromonitrobenzene	1.22
42	4–fluoro–2–nitrotoluene	0.25	**123**	2,6–dinitromethylphenol	1.23
43	3–hydroxy–4–nitrobenzaldehyde	0.27	**124**	2–bromo–4,6–dinitroaniline	1.24
44	2–chlorobenzonitrile	0.28	**125**	4–biphenylcarbonitrile	1.24
45	4–bromobenzonitrile	0.29	**126**	1,2–dinitrobenzene	1.25
46	2,6–dimethylnitrobenzene	0.30	**127**	2,4–dichloro–6–nitroaniline	1.26
47	3–nitroacetophenone	0.32	**128**	4–chloro–3–nitrophenol	1.27
48	5–hydroxy–2–nitrobenzaldehyde	0.33	**129**	2–phenylnitrobenzene	1.30
49	2–fluoro–4–nitrotoluene	0.33	**130**	2–chloro–6–methoxy–3–nitropyridine	1.36
50	4–methyl–2–nitroaniline	0.37	**131**	2,6–dibromo–4–nitrophenol	1.36
51	ethyl–4–cyanobenzoate	0.37	**132**	2–nitro–1–naphthol	1.36
52	2–amino–4–methyl–5–nitropyridine	0.37	**133**	2,5–dibromonitrobenzene	1.37
53	3,4–dinitrophenol	0.37	**134**	3,5–dinitrophenol	1.39
54	4–nitroanisole	0.38	**135**	4–butoxynitrobenzene	1.42
55	3–hydroxy–6–methyl–2–nitropyridine	0.39	**136**	2,4,6–trichloronitrobenzene	1.43
56	methyl–4–nitrobenzoate	0.40	**137**	2,3,4–trichloronitrobenzene	1.51

Table 3. (Continued)

No	Compound	$pIGC_{50}$	No	Compound	$pIGC_{50}$
57	4–nitropyridine	0.41	**138**	3,4–dinitrotoluene	1.52
58	2–chloro–4–methyl–5–nitropyridine	0.42	**139**	2,4,5–trichloronitrobenzene	1.53
59	4–ethylnitrobenzene	0.43	**140**	3–phenylnitrobenzene	1.57
60	4–chloronitrobenzene	0.43	**141**	2,4–dibromo–6–nitroaniline	1.62
61	2–amino–5–chlorobenzonitrile	0.44	**142**	3–trifluoromethyl–4–nitrophenol	1.65
62	3–nitrobenzonitrile	0.45	**143**	4,5–dichloro–2–nitroaniline	1.66
63	4,5–dimethyl–2–nitroaniline	0.45	**144**	2,4–dinitro–5–fluoroaniline	1.69
64	2,5–difluoronitrobenzene	0.45	**145**	2,4–dinitrofluorobenzene	1.71
65	2–amino–4–nitrophenol	0.47	**146**	2–methyl–4,6–dinitrophenol	1.73
66	2–methyl–4–nitroaniline	0.49	**147**	2,4–dichloro–6–nitrophenol	1.75
67	3–nitrophenol	0.51	**148**	2,3,4,5–tetrachloronitrobenzene	1.78
68	4–nitrophenylene–1,2–diamine	0.52	**149**	4–tertbutyl–2,6–dinitrophenol	1.8
69	2,3–dimethylnitrobenzene	0.56	**150**	2,6–diiodo–4–nitrophenol	1.81
70	4–methyl–2–nitrophenol	0.57	**151**	2,3,4,6–tetrafluoronitrobenzene	1.87
71	1,2–dimethyl–4–nitrobenzene	0.59	**152**	1,2,3–trifluoro–4–nitrobenzene	1.89
72	2–chloro–5–nitrobenzaldehyde	0.60	**153**	4–nitrodiphenylamine	1.89
73	4–hydroxy–3–nitrobenzaldehyde	0.61	**154**	2,4–dinitronaphth–1–ol	1.89
74	2–nitroresorcinol	0.66	**155**	1,5–difluoro–2,4–dinitrobenzene	2.08
75	2–methyl–5–nitrophenol	0.66	**156**	4–iodo–1,3–dinitrobenzene	2.12
76	2–nitrophenol	0.67	**157**	2,4,6–trichloro–1,3–dinitrobenzene	2.19
77	3–methoxynitrobenzene	0.67	**158**	1,2–dichloro–4,5–dinitrobenzene	2.21
78	4–nitrobenzaldoxime	0.68	**159**	3,5–dichloro–1,2–dinitrobenzene	2.42
79	2–chloronitrobenzene	0.68	**160**	pentafluoronitrobenzene	2.43
80	2–nitroaniline	0.68	**161**	1,3–dinitro–2,4,5–trichlorobenzene	2.60
81	3–chloro–2–methylnitrobenzene	0.68	**162**	2,3,5,6–tetrachloro–1,4–dinitrobenzene	2.74

For the first set, containing 37 nitrobenzene compounds, acting against *Tetrahymena pyriformis* [28] (Table 2) the toxicity is 50% inhibitory growth impairment concentration ($\log 1/LC_{50} = pLC_{50}$). For the second one, consisting of 167 aromatic compounds acting against *Tetrahymena pyriformis* [29] (Table 3), the toxicity data as 50% growth inhibitory concentration ($\log 1/IGC_{50} = pIGC_{50}$) was considered.

2.2 Calculation of Molecular Descriptors

Topological indices were calculated with TOPOCLUJ, release 3.0, molecular modeling software package [30], developed in our laboratory. A single number, representing a chemical structure, in graph–theoretical terms, is called a topological descriptor. Being a structural invariant, it does not depend on the labeling or the pictorial representation of the graph. Despite the considerable loss of information by the projection in a single number of a structure, such descriptors found broad applications in the correlation and prediction of several molecular properties [31–33] and also in tests of similarity and isomorphism [33,34].

When a topological descriptor correlates with a molecular property, it can be denominated as molecular index or topological index (TI). Only an index having a direct and clear structural interpretation can help to the interpretation of a complex molecular property. If the index correlates with a single molecular property it could indicate the structural composition of that property [35]. If the index can be generalized to higher analogues or it can be built up on various bases (*e.g.*, on

various matrices [34, 35]) it could offer a larger pool of descriptors for the regression analysis.

Almost all descriptors used in this study are based on the Cluj matrices (CJ and CF). Definitions of some basic parameters used in building the Cluj indices are given below. *Cluj matrix*, **CJ**, proposed by Diudea [36–38], is defined by using either the *distance* or the *detour* concept. The non–diagonal entries, [**UM**]*ij* , **M = CJD** (Cluj–Distance) or **CJΔ** (Cluj–Detour), are defined as:

$$[\mathbf{UM}]_{ij} = \max_{k=1,2,...} |V_{i,j,pk}| \tag{1}$$

$$V_{i,j,p_k} = \left\{ v \middle| v \in V(G);\ \ d_{iv} < d_{jv};\ \ (i,v)_h \cap p_k = \{i\};\ \ p_k \in D(G)\ \ or\ \ \Delta(G) \right\}\ \ h,k = 1,2,... \tag{2}$$

where $|V_{i,j,pk}|$ is the cardinality of the set $V_{i,j,pk}$, which is taken as the maximum over all paths $pk = (i,j)_k$. D(G) *and* Δ(G) are the sets of distances (*i.e.*, geodesics) and detours (*i.e.*, elongations), respectively.

If $V_{i,j,pk}$ real (connected) chemical fragments are wanted, the *Cluj fragmental matrices* [38], **CF** are defined. In this version, the sets $V_{i,j,pk}$ are:

$$V_{i,j,pk} = \{v / v \in V(G_p);\ \ G_p = G - p_k;\ \ d_{iv}(G_p) < d_{jv}(G_p);\ \ p_k \in D(G)\ \ or\ \ \Delta(G)\} \tag{3}$$

where $d_{iv}(G_p)$ and $d_{jv}(G_p)$ are the topological distances between a vertex v and vertices i and j, respectively, in the spanning subgraph G_p resulted by cutting the path $p_k = (i,j)_k$ (except its endpoints) from G.

The Cluj indices are calculated as half–sum of the entries in a Cluj symmetric matrix, **M**, (**M = CJD**, **CJD**, **CFD**, **CFD**) [36–38].

$$IE(M) = (1/2)\sum_i \sum_j [M]_{ij}[A]_{ij} = (1/2)\sum_i \sum_j [UM]_{ij}[UM]_{ji}[A]_{ij} \tag{4}$$

$$IP(M) = (1/2)\sum_i \sum_j [M]_{ij} = (1/2)\sum_i \sum_j [UM]_{ij}[UM]_{ji} \tag{5}$$

The number defined on edge, *IE*, is an *index* while the number defined on path, *IP* is a *hyper–index*.

Another descriptor used for the model is partial charges as electronic descriptor. Within TOPOCLUJ program the partial charges Ch_i are calculated as follows [30]:

$$Ch_{i,j} = \log(S_j / S_i)^{1/(d_{i,j})^2} \tag{5}$$

$$Ch_i = \sum_j ch_{i,j} \tag{7}$$

In the above relations, S_i, S_j represent the Sanderson group electronegativities calculated for the hydride groups (*i.e.*, the heavy atoms with their surrounding hydrogen atoms) in the molecule and d_{ij} is the Euclidean distance separating atoms i and j in a minimal energy optimized chemical

structure (HyperChem [45]). $Ch_{i,j}$ is the perturbation of the electronegativity of atom *i* by any *j* atom in molecule while Ch_i is the resultant of these perturbations on the atom *i*.

2.3 Data Analysis

The toxicity data analyzed by means of quantitative structure activity relationships resulted in the finding that the best descriptors set in modeling the toxicity on a given aquatic species can be an excellent predictor of the toxicity of chemicals to other species. Prediction of toxicity can be done from the chemical structure alone and the methods are easily automated [39]. The quality of the QSAR models should meet a number of criteria, which are currently subject of intensive discussion [40]. QSAR models also are required to undergo some form of validation [41]. Benzene derivatives comprise a significant component of the pollutant burden on the environment. The toxicity of these compounds can arise from a multitude of mechanisms of toxic action including a range of narcoses as well as reactive mechanisms in which the compounds are able to form covalent bonds with biological macromolecules.

A total of 92 benzene derivatives representing several mechanisms of toxic action were considered in this study. These are listed in Table 1. The data set is chemically heterogeneous and includes phenols, anilines, nitrobenzenes, other poly–substituted benzenes as well as compounds with more than one functional group on the benzene ring. The data set was distributed in 4 subsets by functional groups, namely: subset 1 (phenol group, compounds **1**–**25**) subset 2 (benzene group, compounds **26**–**46**) subset 3 (nitrobenzene group, compounds **47**–**61**) subset 4 (aniline group, compounds **62**–**92**).

A QSAR model requires to be validated in prediction. To do so, every subset was randomly divided into two groups: a training set and a validation set. A total of four different random tset/vset pairs were generated. The toxicity of the compounds in the validation sets were treated as unknowns and were calculated using the best equation obtained in the corresponding training set.

In order to test the reliability and the prediction potential of entire model, validation procedure was performed using the best models in each subset. These 21 validation compounds are shown in Table 5. The four models were tested for statistical outliers. A compound was considered as an outlier if the residual is more than twice the standard error of estimate for a particular equation. Compounds **6** and **23** (subset 1) and compound **76** (subset 3) appear to be outliers. These compounds were not included in the further analyses. With these outliers removed, we observed somewhat improved correlation with the same descriptors. With TOPOCLUJ software program package [30] we computed 878 molecular descriptors.

2.3.1 Descriptor analysis

Once the desired set of descriptors had been calculated and stored, the process of descriptor analysis is started. It is important to examine the pool of descriptors in an objective manner and to remove from further consideration those descriptors which are redundant or do not contain enough discriminatory information to be of any significant value. All descriptors containing identical values for 90% or more of the compounds in a given data set, including both zero and non–zero values, were removed. All possible combinations of remaining descriptor pairs were examined to identify those pairs that are highly correlated. As a rule of thumb, a critical value of 0.950 for the correlation coefficient (*r*) was used. If two descriptors were correlated at or above the critical value, one descriptor was discarded. The decision of which one to retain was based on the possible physical interpretation of the descriptor, ease of calculation, or usefulness in the past studies. The result of this analysis is a reduced pool of information–rich descriptors that can then be screened by using multiple linear regression analysis. After all these procedures we reduced the searching space from 828 to 498 descriptors.

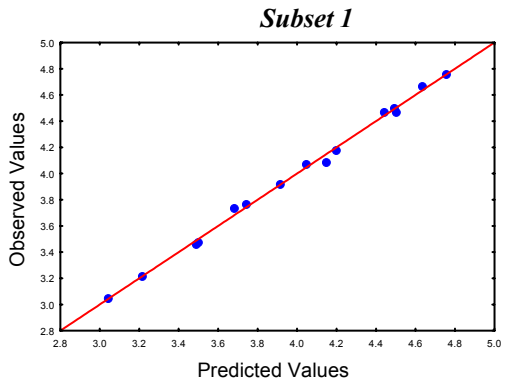
Table 4. Statistics for derived QSTR models for benzene derivative compounds in Table 1

Subset 1				
	B	Std.Err.	t	p–level
Intercept	–0.3839	0.66522	–0.57716	0.579705
C[LM[vdWRadius]]	–3.1417	1.38478	–2.26872	0.052996
C[Sh[CjMax[Covalent radius]]]	3.4842	1.30389	2.67215	0.028268
CS[Sh[CfMin[Charge]]]	0.3807	0.06509	5.84876	0.000383
IP[CfMax[Charge]]	66.7206	7.12156	3.89687	0.004565
IP[CjMax[Charge]]	–84.7958	2.18025	–4.20192	0.002989
IP[CjMin[Charge]]	7.8711	1.81513	4.33635	0.002491
Subset 2				
	B	Std.Err.	t	p–level
Intercept	0.61566	0.296057	2.07952	0.064250
PDS1[LM[Mass]]	0.02283	0.002325	9.81923	0.000002
PDS2[Sh[CjMin[Charge]]]	0.75033	0.098268	7.63561	0.000018
Charges	–4.24963	0.616803	–6.88978	0.000042
PDS2[Sh[CjMin]]	–0.01296	0.002378	–5.45066	0.000281
X[Sh[Distance]]	–1.28584	0.665023	–1.93353	0.081954
Subset 3				
	B	Std.Err.	t	p–level
Intercept	0.63535	1.007826	0.63041	0.551663
PDS2[LM[Electronegativity]]	0.00569	0.013955	0.40792	0.697487
C[Sh[CjMax[Atomic radius]]]	–1.94433	0.579010	–3.35802	0.015267
PDS1[LM[Mass]]	0.00835	0.002771	3.01437	0.023567
C[Sh[Adjacency]]	0.40779	0.230279	1.77084	0.126977
Subset 4				
	B	Std.Err.	t	p–level
Intercept	0.850818	0.886617	0.95962	0.349959
IE[CfMin[Mass]]	–0.000129	0.000059	–2.18713	0.042176
SCH[AdjacencyAdjacencyCharges]	–0.109875	0.028315	–3.88050	0.001096
W2[Covalent radius_Adjacency]	1.109039	0.189763	5.84432	0.000016
PDS3[LM[Density]]	–0.025828	0.005147	–5.01788	0.000089
W2[Electronegativity_Adjacency]	–0.121353	0.046058	–2.63479	0.016821
PDS1[Sh[CfMin[Mass]]]	0.002553	0.001568	1.62852	0.120792
PDS2[Sh[CfMax[Charge]]]	0.120145	0.119824	1.00268	0.329305

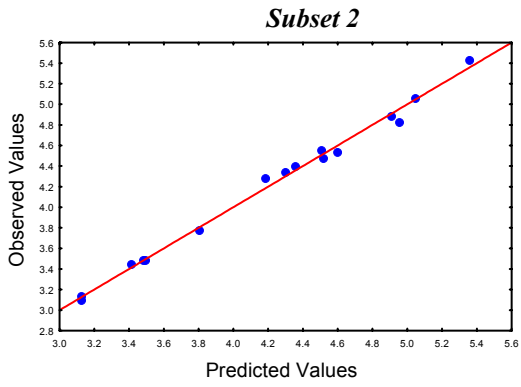
2.3.2 Regression Analysis

In most cases, linear regression models were developed by the method of multiple regression with progressive deletions. The process builds up a model through stepwise addition of terms (descriptors), where the inclusion of a given term is based on the F statistic values. A deletion process is then employed where each independent variable is held out in turn, and a model is developed by using the remaining pool of descriptors. Then all pairs and triplets are held out, and the process is repeated.

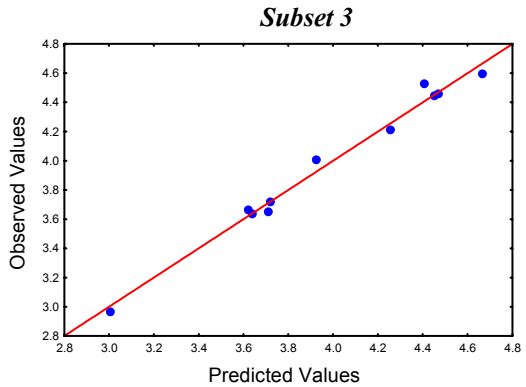
This series of steps has the effect of finding the best equations. The best descriptors and the models were also examined for robustness and predictive ability through both internal and external validation methods. These evaluations are included in the discussion below. As we discussed above, every subset is divided in a training set and a validation set. Subset 1 (tset, $N = 15$; vset, $N = 7$), subset 2 (tset, $N = 16$; vset, $N = 5$), subset 3 (tset, $N = 11$; vset, $N = 4$), subset 4 (tset, $N = 25$; vset, $N = 5$).



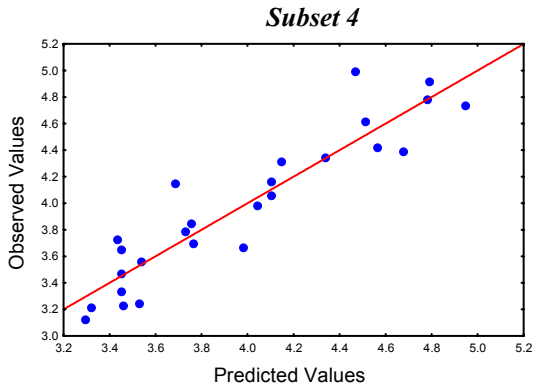
(*a*) $N = 15$; $R = 0.998$; $R^2 = 0.996$; $R^2_{CV} = 0.918$; F = 220.08; p < 0.00001; s = 0.047



(*b*) $N = 16$; $R = 0.997$; $R^2 = 0.994$; $R^2_{CV} = 0.933$; F = 322.57; p < 0.000001; s = 0.068



(*c*) $N = 11$; $R = 0.993$; $R^2 = 0.986$; $R^2_{CV} = 0.958$; F = 104.40; p<0.00001; s = 0.077



(*d*) $N = 25$; $R = 0.926$; $R^2 = 0.857$; $R^2_{CV} = 0.83$; F = 15.438; p < 0.000001; s = 0.25

Figure 1. Observed *vs*. predicted values for the training subsets.

2.3.3 Regression Analysis

To find the best correlations, the four subsets were submitted to STATISTICA software [42] and the results are present in Table 4. Statistical quality of the multivariate relations was judged by parameters such as regression coefficients of the variables (B), standard error (Std. Err.), test for independent samples (t), probability of error (p–level). The corresponding regression equations are shown in Figure 1.

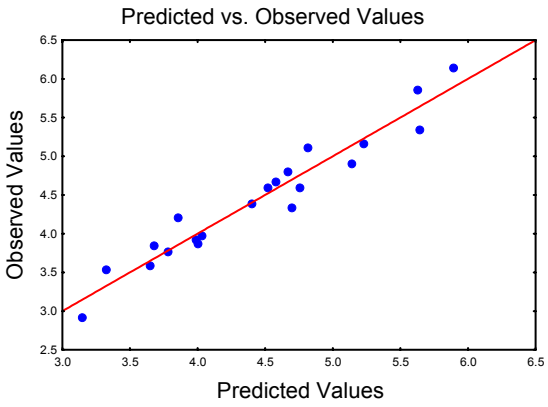
Plots of observed versus predicted values also give evidence of the validity of the models (Figure 1 *a–d*: (*a*) calculated values *cf*. Table 4, subset 1; (*b*) calculated values *cf*. Table 4, subset 2, (*c*) calculated values *cf*. Table 4, subset 3; (*d*) calculated values *cf*. Table 4, subset 4). The quality of the models was estimated by: correlation coefficient (R), squared correlation coefficient (R^2), standard error of estimate (s), Fischer test (F) cross–validated correlation coefficient (R^2_{CV}), and probability of error (p). The best equations for the training sets are from Table 4 subset 2 and subset 3, as shown in Figure 1 with the high correlation coefficient R, low standard deviations, and least variables.

Table 5. Observed, predicted, residual values and the coefficient of variance (CV%) for the validation data

Compd.	$pLC_{50\ obs}$	$pLC_{50\ pred}$	$pLC_{50\ residual}$	CV %
62	2.91	3.24	–0.331	11.383
69	3.52	3.41	0.105	2.987
48	3.59	3.71	–0.119	3.311
82	3.77	3.83	–0.059	1.562
36	3.84	3.74	0.104	2.715
5	3.86	4.04	–0.177	4.581
88	3.93	4.01	–0.084	2.146
16	3.96	4.06	–0.097	2.454
8	4.21	3.89	0.316	7.508
20	4.33	4.67	–0.343	7.919
77	4.39	4.40	–0.008	0.174
59	4.58	4.50	0.078	1.697
40	4.6	4.72	–0.123	2.682
56	4.66	4.56	0.100	2.153
15	4.81	4.64	0.167	3.477
13	4.9	5.07	–0.171	3.498
14	5.12	4.78	0.345	6.732
79	5.15	5.16	–0.013	0.252
33	5.35	5.53	–0.181	3.382
35	5.85	5.53	0.322	5.509
41	6.15	5.77	0.378	6.144

2.3.4 Validation set

After all the toxicity values for the training subsets were predicted and cross–validated correlation coefficient calculated we can find the residual toxicity values for the validation data. The 21 benzene derivative compounds randomly selected for every subset are presented in Table 5, while the plot of predicted and residual values in the validation set in Figure 2.



$pLC_{50\ obs} = -0.03879 + 1.0896\ pLC_{50\ pred}$

Figure 2. Observed and predicted values in the validation set.

The prediction ability of the derived QSAR model (Figure 2) has the following statistics: $N = 21$; $R = 0.967$; $R^2 = 0.935$; $F = 277.07$; $p = 8.71E{-}13$; $s = 0.2$. The residuals (*i.e.*, differences between predicted and experimental data) for the acute toxicity over *Poecilia reticulata*, listed in Table 5, are very low. The results confirm the excellent prediction and the robustness of the QSAR models derived by topological descriptors analysis. The prediction ability of the model (93.5%) shows that topological descriptors used for the models are useful to detect some functional groups of molecules, particularly for NO_2 group, observation that will be demonstrated in section **2.3.5**. The goal of the experiment was to select those descriptors able to describe better a given functional group but not to describe well a mixing set of molecules.

Recall that, benzene derivatives sets were modeled by Roy and Ghosh [26] and the model with their non–ETA indices showed the following quality: $N = 92$, $Q^2 = 0.718$, $R^2 = 0.738$, $R = 0.859$, $F = 82.8$, $s = 0.340$. With extended topochemical atom (ETA) indices the quality of the models was: $N = 92$, $Q^2 = 0.865$, $R^2 = 0.885$, $R = 0.941$, $F = 92.6$, $s = 0.230$. However, our experiment is different in that we do not look for the best model for a property on a given set of molecules, but for the best set of descriptors modeling a given functional group. In conclusion, based on the results from Table 5 and Figure 2, we can say that toxicity increases as the number of electronegative substituents on the benzene nucleus, such as chlorine atoms, increase. The presence of NO_2 (or NH_2) groups on benzene decreases the toxicity.

2.3.5 Prediction sets

A more convincing test is to use the QSAR models in predicting the values of a biological activity for an entirely new set of compounds, and to examine how well these descriptors will predict the experimental values. The indices from the best models were used to predict the toxicity of nitrobenzene derivatives and aromatic compounds to *Tetrahymena pyriformis*. To investigate the robustness of the model, a leave–half–out (LHO) method, one of many cross–validation (CV) techniques has been applied. LHO method simulates the predictive quality of the regression

equation in a satisfactory manner. The first set comprising the toxicity data of 37 nitrobenzene derivatives against to *Tetrahymena pyriformis*, was taken from Ref. [28] and are shown in Table 2. The statistics are presented in Table 6, together with the LHO cross–validated value (Q^2).

Table 6. Statistics model for the toxicity of nitrobenzenes in Table 2

Indices from subset 1 **(Table 4)**										
	B	Std.Err.	t(30)	p–level	R	R^2	Q^2	F	p	s
Intercept	–4.05E–02	1.19E+00	–3.40E–02	9.73E–01	0.917	0.841	0.78	26.45	1.03E–10	0.32
C[LM[vdWRadius]]	2.25E+01	4.15E+00	5.42E+00	7.15E–06						
C[Sh[CjMax[Covalent radius]]]	–2.18E+01	4.42E+00	–4.93E+00	2.83E–05						
CS[Sh[CfMin[Charge]]]	–6.39E–02	2.50E–02	–2.56E+00	1.58E–02						
IP[CfMax[Charge]]	1.40E+00	3.71E+00	3.78E–01	7.08E–01						
IP[CjMax[Charge]]	–1.93E+00	3.70E+00	–5.21E–01	6.06E–01						
IP[CjMin[Charge]]	–1.25E–01	2.72E–01	–4.59E–01	6.49E–01						
Indices from subset 2 **(Table 4)**										
	B	Std. Err.	t(31)	p–level	R	R^2	Q^2	F	p	s
Intercept	–3.67E+00	8.95E–01	–4.10E+00	2.80E–04	0.977	0.955	0.92	136.17	3.88E–20	0.16
Charges	1.28E+00	2.82E–01	4.55E+00	7.79E–05						
PDS1[LM[Mass]]	1.60E–03	3.18E–03	5.02E–01	6.19E–01						
PDS2[Sh[CjMin[Charge]]]	1.12E–05	1.05E–05	1.07E+00	2.95E–01						
PDS2[Sh[CjMin]]	–1.23E–02	6.96E–03	–1.77E+00	8.65E–02						
X[Sh[Distance]]	1.11E–01	5.22E–02	2.13E+00	4.17E–02						
Indices from subset 3 **(Table 4)**										
	B	Std. Err.	t(32)	p–level	R	R^2	Q^2	F	p	s
Intercept	–4.74E+00	5.09E–01	–9.31E+00	1.26E–10	0.98	0.96	0.94	188.36	9.42E–22	0.13
C[Sh[Adjacency]]	–2.16E–05	7.06E–06	–3.06E+00	4.40E–03						
C[Sh[CfMax[Atomic radius]]]	1.81E–01	4.79E–02	3.77E+00	6.64E–04						
PDS1[LM[Mass]]	1.63E–02	3.12E–03	5.24E+00	9.84E–06						
PRDS[LM[Electronegativity]]	–8.08E–02	2.55E–02	–3.17E+00	3.33E–03						
Indices from subset 4 **(Table 4)**										
	B	Std. Err.	t(31)	p–level	R	R^2	Q^2	F	p	s
Intercept	–4.60E+00	1.51E+00	–3.06E+00	4.74E–03	0.975	0.95	0.91	79.699	3.27E–17	0.18
PDS1[Sh[CfMin[Mass]]]	6.47E–04	1.26E–04	5.15E+00	1.69E–05						
W2[Electronegativity_Adjacency]	–4.85E–01	5.24E–01	–9.24E–01	3.63E–01						
SCH[AdjacencyAdjacencyCharges]	–2.37E–02	9.66E–03	–2.46E+00	2.03E–02						
W2[Covalent radius_Adjacency]	1.22E+00	5.48E–01	2.22E+00	3.45E–02						
PDS3[LM[Density]]	–1.05E–02	2.60E–03	–4.03E+00	3.69E–04						
IE[CfMin[Mass]]	–7.71E–02	4.22E–01	–1.83E–01	8.56E–01						
PDS2[Sh[CfMax[Charge]]]	–9.11E–02	2.99E–02	–3.05E+00	4.84E–03						

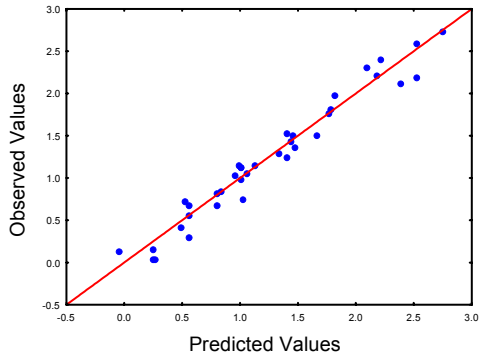
The models in Table 6 were developed by the best indices in Table 4, namely the best models are obtained with indices for the subset 3 which are proper for describing the nitrobenzene derivatives subset.

$$pLC_{50} = -4.74 - 0.00002\ C[Sh[Adjacency]] + 0.18\ C[Sh[CfMax[Atomic\ radius]]] + 0.016\ PDS1[LM[Mass]] - 0.08\ PRDS[LM[Electronegativity]] \quad (8)$$

So, if the activity of nitrobenzene derivatives is assumed unknown, from the predicted toxicity by Eq. (8), we figure up that it speaks on nitrobenzene compounds. Figure 3 shows the plot of observed vs. predicted values by subset 3 (Table 6).

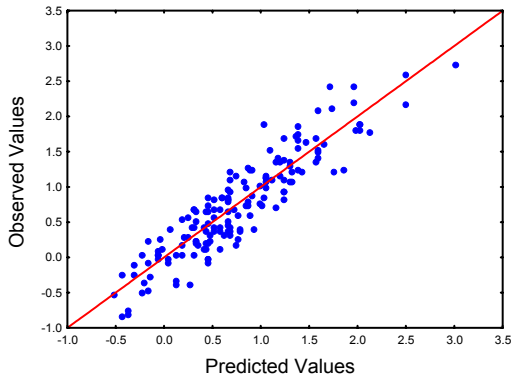
The prediction of toxicity for nitrobenzene derivatives against *Tetrahymena pyriformis* shows very good quality: four topological descriptors, Eq. (8), explain 96% and predict 94% of the variance of toxicity of nitrobenzene derivative with a low standard error (s = 13%). Toxicity of

nitrobenzenes to *Tetrahymena pyriformis* was modeled by Estrada *et al*. [28] using fragmental contributions adopting TOPSMODE approach, and R^2 statistic for the best developed model was 0.910 (Q^2 = 0.901, F = 93.9 [df 4,37], s = 0.22). Nitrobenzenes toxicity data against *Tetrahymena pyriformis* was also modeled by Cronin *et al*. [41] using physicochemical descriptors (1–octanol–water partition coefficient and molecular orbital parameters) and R^2 value of the best model was 0.881 (Q^2 = 0.866, F = 154 [df 2,39], s = 0.246).



$pLC_{50\ obs}$ = 6.779E–8+ 0.997 $pLC_{50\ pred}$

Figure 3. Plot of observed *vs*. predicted values for the best model with indices for subset 3 (Table 6) for 37 nitrobenzene derivatives (Table 2) against *Tetrahymena pyriformis*.



$pIGC_{50\ obs}$ = 1.531E–8+ 0.999 $pIGC_{50\ pred}$

Figure 4. Plot of observed *vs*. predicted values for the best model with indices for subset 3 (Table 7) for 162 aromatic compounds (Table 3) against *Tetrahymena pyriformis*.

The positive coefficient of PDS1[LM[Mass]] in Eq. (8) indicates that the nitrobenzene toxicity increases with rise in molecular bulk while the negative coefficient of PRDS[LM[Electronegativity]] implies the decrease of toxicity with decreasing atom electronegativity. Molecules with two nitro groups attached to aromatic ring together with three, four chlorine atoms have the highest toxicity. The presence of $–CH_3$ in meta or para position *vs*. nitrobenzene, results in some decrease of toxicity. Some data on frequent probability (*e.g*., HazardExpert) [3] show that the probability that a substance containing both aromatic nitro groups and chlorine atoms to be carcinogenic is 68%.

The second prediction set contains 167 aromatic compounds tested against *Tetrahymena pyriformis* is taken from Ref. [29] and are shown in Table 3. From this set, 135 are nitro derivative compounds. The statistics of the models for the toxicity of these aromatic compounds are presented in Table 7.

Table 7. Statictics for the toxicity of aromatic compounds in Table 3

Indices from subset 1 (**Table 4**)										
	B	Std.Err.	t(158)	p–level	R	R^2	Q^2	F	p	s
Intercept	–3.38E+00	6.92E–01	–4.88E+00	2.59E–06	0.83	0.69	0.61	58.02	3.7E–40	0.4
C[LM[vdWRadius]]	8.07E+00	1.82E+00	4.44E+00	1.73E–05						
C[Sh[CjMax[Covalent radius]]]	–6.37E+00	1.69E+00	–3.77E+00	2.33E–04						
CS[Sh[CfMin[Charge]]]	6.07E–02	1.53E–02	3.96E+00	1.14E–04						
IP[CfMax[Charge]]	6.26E–02	2.00E+00	3.12E–02	9.75E–01						
IP[CjMax[Charge]]	9.01E–02	2.25E+00	4.00E–02	9.68E–01						
IP[CjMin[Charge]]	–6.41E–02	2.70E–01	–2.38E–01	8.13E–01						
Indices from subset 2 (**Table 4**)										
	B	Std.Err.	t(156)	p–level	R	R^2	Q^2	F	p	s
Intercept	–2.02E+00	1.57E–01	–1.29E+01	1.87E–26	0.896	0.80	0.77	128.36	0.0E–01	0.32
Charges	7.05E–01	2.19E–01	3.21E+00	1.59E–03						
PDS1[LM[Mass]]	7.83E–03	7.64E–04	1.02E+01	3.76E–19						
PDS2[Sh[CjMin[Charge]]]	1.52E–02	5.71E–02	2.65E–01	7.91E–01						
PDS2[Sh[CjMin]]	2.96E–04	5.37E–04	5.50E–01	5.83E–01						
X[Sh[Distance]]	–7.73E–01	9.61E–01	–8.05E–01	4.22E–01						
Indices from subset 3 (**Table 4**)										
	B	Std.Err.	t(157)	p–level	R	R^2	Q^2	F	p	s
Intercept	–2.80E+00	2.51E–01	–1.12E+01	1.13E–21	0.917	0.84	0.82	198.64	0.0E–01	0.2
PDS1[LM[Mass]]	6.56E–03	8.95E–04	7.33E+00	1.16E–11						
C[Sh[Adjacency]]	1.48E–01	4.64E–02	3.19E+00	1.71E–03						
C[Sh[CjMax[Atomic radius]]]	–8.73E–03	1.18E–03	–7.37E+00	9.09E–12						
PRDS[LM[Electronegativity]]	–1.08E–03	2.93E–03	–3.68E–01	7.13E–01						
Indices from subset 4 (**Table 4**)										
	B	Std.Err.	t(154)	p–level	R	R^2	Q^2	F	p	s
Intercept	–2.03E+00	6.98E–01	–2.91E+00	4.14E–03	0.90	0.82	0.79	101.04	0.0E–01	0.3
IE[CfMin[Mass]]	–1.32E–05	9.73E–06	–1.36E+00	1.76E–01						
PDS1[Sh[CfMin[Mass]]]	1.06E–03	3.84E–04	2.75E+00	6.63E–03						
PDS2[Sh[CfMax[Charge]]]	1.46E–01	4.42E–02	3.29E+00	1.23E–03						
PDS3[LM[Density]]	7.97E–01	3.37E–01	2.37E+00	1.93E–02						
SCH[Adjacency AdjacencyCharges]	–4.51E–02	1.39E–02	–3.23E+00	1.50E–03						
W2[ElectronegativityAdjacency]	3.23E–02	2.91E–02	1.11E+00	2.70E–01						
W2[Covalent radius_Adjacency]	1.96E–01	1.66E–01	1.18E+00	2.39E–01						

The best equation was obtained with the same descriptors as for the nitrobenzene derivatives, with indices for subset 3 (Table 7). The four topological descriptors, Eq. (9), explain 84% and predict 82% of the variance of toxicity for 162 aromatic compounds against *Tetrahymena pyriformis*.

$$\begin{aligned} pIGC_{50} = &- 2.08 + 0.0065\ PDS1[LM[Mass]] + 0.148\ C[Sh[Adjacency]] \\ &- 0.0087\ C[Sh[CjMax[Atomic\ radius]]] - 0.001\ PRDS[LM[Electronegativity]] \end{aligned} \quad (9)$$

Plot of observed *vs*. predicted values for the best model (subset 3, Table 7) is presented in Figure 4. The slope near unity and the intercept near to zero, demonstrates again a good fit between observed and predicted values. Eq. (9) has the following quality (Table 7): $N = 162$, $R = 0.917$, $R^2 = 0.84$, $Q^2 = 0.82$, $F = 198.64$, $s = 0.2$ and the results are comparable with those reported in literature

for this type of compounds. Cronin *et al*. [42] studied aromatic compounds against *Tetrahymena pyriformis* and they obtained the following model with electrophylic parameter: $N = 203$, $R^2 = 0.70$, $s = 0.42$, $F = 237$. In another study Cronin *et al*. [43] developed a model with 268 aromatic compounds and with *LUMO* and *LogP* as descriptors, they obtained a better model: $N = 239$, $R^2 = 0.800$, $R_{CV}^2 = 0.796$, $s = 0.335$, $F = 476$. Substances that contain two nitro groups and more than three chlorine atoms bring to the benzene nucleus increasing toxicity. Further, the toxicity decreases with the presence of nitrile group and methyl on the benzene nucleus, as well as with the presence of cyano substituted pyridine nucleus.

3 CONCLUSIONS

In this paper, a QSAR method based on topological descriptors was employed to predict acute toxicity of benzene derivatives against *Poecilia reticulata*. The predicted values are very close to the experimental ones. The goal of the experiment was to select those descriptors able to describe better a given functional group but not to describe well a mixing set of molecules. To evaluate the importance of the descriptors from the best models we predicted the acute toxicity for two other sets of benzene derivatives against *Tetrahymena pyriformis*. The results showed this QSAR approach a highly predictive one for aquatic toxicity of pollutants.

The prediction ability of the toxicity model of benzene derivatives against *Poecilia reticulata* describes 93.5% of the variance with a low standard error, better than those presented in previous studies. The prediction of toxicity for nitrobenzene derivatives against *Tetrahymena pyriformis* shows very good quality: it explains 96% and predicts 94% of the variance with very good statistical parameters. These results have more statistical significance that those reported in the literature. The same four descriptors that describe NO_2 group explain 84% and predict 82% of the variance of toxicity for 162 aromatic compounds against *Tetrahymena pyriformis*.

For all the three sets of aromatic compounds toxicity increases with the presence of electronegative atoms, particularly the chlorine atoms bound to the benzene nucleus, in various positions. Further, nitrobenzene derivatives, with two nitro groups attached to aromatic ring show an increased toxicity. The presence of nitrile group on pyridine rings or amino group as the same as methyl radical attached to aromatic ring decrease toxicity.

In the present study, it was shown that topological indices could explore the important chemical information contributing to the aquatic toxicity of substituted benzenes and the quantitative relations obtained could predict the activity of unknowns with good accuracy. It appears that topological descriptors have significant potential in QSAR/QSPR/QSTR studies, which warrants extensive evaluation.

5 REFERENCES

[1] T. W. Schultz, TETRATOX: The *Tetrahymena pyriformis* population growth impairment endpoint. A surrogate for fish lethality. *Toxicol. Methods* **1997**, *7*, 289–309.
[2] D. L. Hill, The Biochemistry and Physiology of Tetrahymena, **1972**, 1st ed., pp 230, Academic Press, New York.
[3] Eldred, C. L. Weikel, P. C Jurs, and K. L. E. Kaiser, Prediction of fathead minnow acute toxicity of organic compounds from molecular structure. *Chem. Res. Toxicol.* **1999**, *12*, 670–678.
[4] G. W. Kauffman and P. C. Jurs, Prediction of the sodium ion–proton antiporter by benzoylguanidine derivatives from molecular structure, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 753–761.
[5] S. J. Patankar and P. C. Jurs, Prediction of IC50 values for ACAT inhibitors from molecular structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 706–723.
[6] M. D. Wessel, P. C. Jurs, J. W. Tolan, and S. M. Muskal, Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726–735.
[7] S. R. Johnson and P. C. Jurs, Prediction of the clearing temperatures of a series of liquid crystals from molecular structure. *Chem. Mater.* **1999**, *11*, 1007–1023.
[8] H. E. McClelland and P. C. Jurs, Quantitative structure–property relationships for the prediction of vapor pressures of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 967–975.
[9] P. J. Cumpson, Estimation of inelastic mean free paths for polymers and other organic materials: use of quantitative structure–property relationships. *Surf. Interface Anal.* **2001**, *31*, 23– 34.
[10] A. D. DeWeese and T. W. Schultz, Structure–activity relationships for aquatic toxicity to *Tetrahymena*: Halogensubstituted aliphatic esters. *Environ. Toxicol.* **2001**, *16*, 54–60.
[11] V. A. Filov, A. A. Golubev, E. I. Liublinaand, and N. A. Tokontsev, *Quantitative Toxicology*, John Wiley & Sons, New York, **1979**, 1st ed., pp 462
[12] J. D. LeBlond, B. M. Applegate, F. M. Menn, T. W. Schultz, and G. S. Sayler, Structure–toxicity assessment of metabolites of the aerobic bacterial transformation of substituted naphthalenes. *Environ. Toxicol. Chem.* **2000**, *19*, 1235–1246.
[13] C. L. Russom, S. P. Bradbury, S. J. Broderius, D. E. Hammermeister, and R. A. Drummond, Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ. Toxicol. Chem.* **1997**, *16*, 948–967.
[14] S. P. Bradbury, T. R. Henry, and R. W. Carlson, Fish Acute Toxicity Syndromes in the Development of Mechanism–Specific QSARs. In *Practical Applications of QSARs in Environmental Chemistry and Toxicology*; W. Karcher, J. Devillers, Eds.; Kluwer: Dordrecht, **1990**, pp 295–316.
[15] D. R. Hartter, The use and importance of nitroaromatic chemicals in the chemical industry. In *Toxicity of nitroaromatic compounds* (Rickert, D. E., Ed.), Hemisphere, New York, **1985**, pp 1–13.
[16] J. Arey, Atmospheric reactions of PAHs including formation of nitroarenes. In *The Handbook of Environmental Chemistry*, *Volume 3*, *Part I. PAHs and Related Compounds* (Neilson, A. H., Ed.) Springer–Verlag, Berlin, **1998**, pp 347–385.
[17] W. F. Jr. Bushby, H. Smitz, C. L. Crespi, B. W. Penman, and A. L Lafleur, Mutagenicity of the atmospheric transformation products 2–nitrofluoranthene and 2–nitrodibenzopyranone in *Salmonella* and human cell forward mutation assays. *Mutat. Res.* **1997**, *389*, 261–270.
[18] G. Rippen, E. Zietz, R. Frank, T. Knacker, and W. Klopffer, Do airborne nitrophenols contribute to forest decline, *Environ. Technol. Lett.* **1987**, *8*, 475–482.
[19] H. C. Bailey and R. J. Spanggord, The relationship between the toxicity and structure of nitroaromatic chemicals. In *Aquatic Toxicology and Hazard Assessment*, *Sixth Symposium*, *ASTM STP 802* (Bishop, W. E., Cardwell, R. D., and Heidolph, B. B., Eds.), American Society for Testing and Materials, Philadelphia. **1983**, pp 98–107.
[20] J. W. Deneer, T. L. Sinnige, W. Seinen, and J. L. M. Hermens, Quantitative structure–activity relationships for the toxicity and bioaccumulation factor of nitrobenzene derivatives towards the guppy (*Poecilia reticulata*). *Aquat. Toxicol.* **1987**, *10*, 115– 129.
[21] J. W. Deneer, C. J. van Leeuwen, W. Seinen, J. L. Maas–Diepeveen, and J. L. M.Hermens, A QSAR study of the toxicity of nitrobenzene derivatives towards *Daphnia magna*, *Chlorella pyrenoidosa* and *Photobacterium phosphoreum*. *Aquat. Toxicol.* **1989**, *15*, 83–98.
[22] K. Rose and L. H. Hall, E–State Modeling of Fish Toxicity Independent of 3–D Structure Information, *SAR QSAR Environ. Res.* **2003**, *14*, 113–129.
[23] A. R. Katritzky, D. B. Tatham, and U. Maran, Theoretical Descriptors for the Correlation of Aquatic Toxicity of Environmental Pollutants by Quantitative Structure–Toxicity Relationships, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1162–1176.
[24] A. A. Toropov and T. W. Schultz, Prediction of Aquatic Toxicity: Use of Optimization of Correlation Weights of Local Graph Invariants, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 560–567.

[25] J. R. Seward and T. W. Schultz, QSAR Analyses of the Toxicity of Aliphatic Carboxylic Acids and Salts to *Tetrahymena pyriformis*, *SAR QSAR Environ Res*. **1999**, *10*, 557–567.
[26] K. Roy and G. Ghosh, Introduction of Extended Topochemical Atom (ETA) Indices in the Valence Electron Mobile (VEM) Environment as Tools for QSAR/QSPR Studies, *Internet Electron*. *J*. *Mol*. *Des*. **2003**, *2*, 599–620, http://www.biochempress.com.
[27] K. Roy and G. Ghosh, QSTR with Extended Topochemical Atom Indices. 2. Fish Toxicity of Substituted Benzenes, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **2004**, *44*, 559–567.
[28] E. Estrada and E. Uriate, Quantitative Structure Activity Relationships Using TOPS–MODE. 1. Nitrobenzene Toxicity to *Tetrahymena pyriformis*. *SAR QSAR Environ*. *Res*. **2002**, *12*, 309–324.
[29] T. W. Schultz, T. I. Netzeva, and M. T. D. Cronin, Selection of Data Sets for QSARs: Analysis of *Tetrahymena* Toxicity from Aromatic Compounds, *SAR QSAR Environ*. *Res*. **2003**, *14*, 59–81.
[30] M. V. Diudea and O. Ursu, TOPOCLUJ (Copyright Babes–Bolyai Univ. Cluj), **2002**
[31] A.T. Balaban, I. Motoc, D. Bonchev, and O. Mekenyan, Topological Indices for Structure–Activity Correlations, *Top*. *Curr*. *Chem*. **1993**, *114*, 21–55.
[32] D. H. Rouvray, The Challenge of Characterizing Branching in Molecular Species *Discr*. *Appl*. *Math*. **1988**, *19*, 317–338.
[33] M. Randić, Design of Molecules with Desired Properties. A Molecular Similarity Approach to Property Optimization, in *Concepts and Applications of Molecular Similarity*, M. A. Johnson and G. M. Maggiora, Eds. John Wiley & Sons, Inc., **1990**, Chap. 5, pp.77–145
[34] M. V. Diudea, Molecular Topology. 16. Layer Matrixes in Molecular Graphs, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **1994**, *34*, 1064–1071.
[35] M. Randić, Generalized Molecular Descriptors, *J*. *Math*. *Chem*. **1991**, *7*, 155–168.
[36] M. V. Diudea, Cluj Matrix Invariants, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **1997**, *37*, 300–305
[37] M. V. Diudea, I. Gutman, Wiener–Type Topological Indices, *Croat*. *Chem*. *Acta*, **1998**, *71*, 21–51.
[38] M. V. Diudea and B. Parv, I. Gutman, Detour–Cluj Matrix and Derived Invariants, *J*. *Chem*. *Inf*. *Comput*. *Sci*. **1997**, *37*, 1101–1108.
[39] M. T. D. Cronin, The current status and future applicability of quantitative structure–activity relationships (QSARs) in predicting toxicity. *ATLA* **2002**, *30*, Supplement 2, 81–84.
[40] A. P. Worth, ECVAM's activities on computer modelling and integrated testing. *ATLA* **2002**, *30*, Supplement 2, 133–137.
[41] M. T. D. Cronin, B. W. Gregory, and T. W. Schultz, Quantitative Structure–Activity Analyses of Nitrobenzene Toxicity to *Tetrahymena pyriformis*, *Chem. Res. Toxicol*. **1998**, *11*, 902–908.
[42] M. T. D. Cronin, N. Manga, J. R. Seward, G. D. Sinks, and T. W. Schultz, Parametrization of Electrophilicity for the Prediction of the Toxicity of Aromatic Compounds, *Chem Res Toxicol*. **2001**, *14*, 498–505.
[43] M. T. D. Cronin and T. W. Schultz, Development of Quantitative Structure–Activity Relationships for the Toxicity of Aromatic Compounds to *Tetrahymena pyriformis*: Comparative Assessment of the Methodologies, *Chem*. *Res*. *Toxicol*. **2001**, *14*, 1284–1295.
[44] StatSoft, Inc. (2001), STATISTICA (data analysis software system), version 6, www.statsoft.com.
[45] HyperChem, release 4.5 for SGI, © 1991–1995, Hypercube, Inc.

Biographies

Adina Costescu is a Ph. D student of computational chemistry at University of Chemistry and Chemical Engineering. Her current research activities include: molecular similarity, SAR (Structure Activity Relationships), QSAR/QSPR (Quantitative Structure Activity/Property Relationships), advanced statistical methods and data analysis (PLS, PCA, PCR), advanced chemometrics methods.

Mircea V. **Diudea** is professor of Organic Chemistry and Molecular Topology at the "Babes–Bolyai" University, Cluj, Romania. His research topics include: topological matrices, topological indices, hyper branched structures, symmetry and similarity, QSPR/QSAR, fullerenes, nanotubes and tori, periodic nanostructures, aromaticity of nanostructures. He is member of the International Academy of Mathematical Chemistry.