The Monte Carlo Method Based on Eclectic Data as an Efficient Tool for Predictions of Endpoints for Nanomaterials – Two Examples of Application

ARTICLE *in* COMBINATORIAL CHEMISTRY & HIGH THROUGHPUT SCREENING · MARCH 2015

Impact Factor: 1.22 · DOI: 10.2174/1386207318666150305125044 · Source: PubMed

Andrey Toropov

Mario Negri Institute for Pharmacological Re…

Alla Toropova

Mario Negri Institute for Pharmacological Re…

Eduardo Castro

National Scientific and Technical Research C…

Danuta Leszczynska

Jackson State University

The Monte Carlo Method Based on Eclectic Data as an Efficient Tool for Predictions of Endpoints for Nanomaterials – Two Examples of Application

Andrey A. Toropov[*,1], Alla P. Toropova[1], Aleksandar M. Veselinović[2], Jovana B. Veselinović[2], Karel Nesmerak[3], Ivan Raska Jr[4], Pablo R. Duchowicz[5], Eduardo A. Castro[5], Valentin O. Kudyshkin[6], Danuta Leszczynska[7] and Jerzy Leszczynski[8]

[1]*IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy*

[2]*University of Niš, Faculty of Medicine, Department of Chemistry, Niš, Serbia*

[3]*Charles University in Prague, Faculty of Science, Department of Analytical Chemistry, Hlavova 8, 128 43 Prague 2, Czech Republic*

[4]*3rd Department of Medicine, Department of Endocrinology and Metabolism, First Faculty of Medicine, Charles University in Prague and General University Hospital in Prague, U Nemocnice 1, 12808 Prague 2, Czech Republic*

[5]*Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64, C.C. 16, Suc.4, (1900) La Plata, Argentina*

[6]*Institute of Polymer Chemistry and Physics, Uzbek Academy of Sciences, ul. Abdully Kadyri 7b, 100128 Tashkent, Uzbekistan*

[7]*Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, Mississippi 39217, USA*

[8]*Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch Street, Jackson, Mississippi 39217-0510, USA*

Abstract: The theoretical predictions of endpoints related to nanomaterials are attractive and more efficient alternatives for their experimental determinations. Such type of calculations for the "usual" substances (i.e. non nanomaterials) can be carried out with molecular graphs. However, in the case of nanomaterials, descriptors traditionally used for the quantitative structure – property/activity relationships (QSPRs/QSARs) do not provide reliable results since the molecular structure of nanomaterials, as a rule, cannot be expressed by the molecular graph. Innovative principles of computational prediction of endpoints related to nanomaterials extracted from available eclectic data (technological attributes, conditions of the synthesis, etc.) are suggested, applied to two different sets of data, and discussed in this work.

Keywords: CORAL software, optimal descriptor, Quasi-QSPR/QSAR.

INTRODUCTION

Nanomaterials have different and useful features in comparison to "ordinary" materials like unique physical and chemical properties, such as shape, size distribution, surface area and structure, overall charge, porosity, agglomeration rate, and surface chemistry [1, 2]. These unique properties make nanomaterials very applicable in vast areas like electronics, opto-electronics, biomedical, environmental, material and energy related areas, cosmetics, pharmaceuticals and catalysts [3-5]. However, nanomaterials can have adverse effects on humans and environment [6-10]. Nanomaterials present in commercial products may potentially cause systemic, cellular and/or genomic toxicities through acute or repeated exposure. For this reason, understanding the biological effects of exposure to nanomaterials is of great importance. Further, toxicological characterization of the immense number of structural combinations for future and already synthesized nanomaterials are extremely demanding (if not impossible) in terms of time, costs and experimental facilities. However, the application of *in silico* toxicology methods is a way to overcome this issue [11]. Unfortunately, since nanomaterials have structural heterogeneity, complexity and diversity, modeling their biological effects is a difficult task.

Eclecticism as a philosophical approach can be interpreted as the choice of elements from different systems without consideration of possible contrasts between the systems. Therefore, eclecticism is based on one of the most

*Address correspondence to this author at the IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy;
Tel: ??????????????????; Fax: ??????????????????;
E-mail: andrey.toropov@marionegri.it

1386-2073/15 $58.00+.00

important defining characteristics of the scientific process - the selection of those elements, properties, characteristics which define the empirical object [12].

For over 50 years, various computational chemistry approaches have been used to predict molecular structures, properties and mechanisms of various processes. Among such techniques, quantitative structure – property/activity relationships (QSPR/QSAR) are an epistemological tool widely used in the natural sciences. The QSPR/QSAR analyses provide possibility to comprehend physical [13], physicochemical [14], biochemical [15], and medicinal [16] phenomena related to the considered species. Though such approaches are well established for classical molecules, nevertheless, in spite of the increasing number of publications related to the "nano-QSPR/QSAR", the knowledge about the nanomaterials remains eclectic and is not systematic. Under such circumstances, a possible way to define rational approach for the QSPR/QSAR analyses of nanomaterials is to suggest eclectic mathematical formulae which are able to address this task [17, 18]. The Monte Carlo technique is a possible way to construct a model as a mathematical function of various eclectic data [16, 19, 20]. The majority of traditionally used molecular descriptors are calculated with the molecular graph [21-24]. Simplified molecular input-line entry system (SMILES) can be an alternative of the molecular graph in the QSPR/QSAR analyses [25, 26]. In the case of the nanomaterials the application of the molecular graph as well as SMILES to build up a predictive model for some endpoint often is impossible owing to absence of detailed molecular structure for nanomaterials. However, data on various nano particles can be represented by special strings which are encoded data on physicochemical and biochemical conditions of impact of the nanoparticles. These SMILES-like strings can be named "quasi SMILES", since they represent conditions in contrast of traditional SMILES which represent the molecular structures.

There are different methodologies used for classical molecules and nanomaterials. The paradigm for traditional QSPR/QSAR analyses could be expressed as:

Endpoint = F(Molecular Structure)

In the case of the nanomaterials the paradigm can be modified to read as follows:

Endpoint = F(Available Eclectic Data)

The available eclectic data can be (i) the molecular structure of substances which are involved in phenomenon under consideration; (ii) presence/absence of irradiation (photo-inducing); (iii) concentration of nanomaterial and/or substances with small molecules involved in phenomenon under consideration; and (iv) any other circumstances which are able to have influence on the considered phenomenon.

The current work examines proposed innovative nano-QSPR/QSAR approach for two different set of data. The first case covers photocatalytic decolorisation rate constants related to doped-titania nanopowder photocatalysts (quasi QSPR), while the second relates to influence of fullerene on bacterial reverse mutation (quasi QSAR). These approaches are based on special SMILES-like strings of symbols which encoded conditions of impact of nano particle for physicochemical and biological phenomena. Thus, these models should be named as quasi QSPR/QSAR.

METHOD

Data Set 1

Experimental data on photocatalytic decolorisation rate constants DRC (10^{-5}/s) of methylene blue dye for 17 different doped-titania nanopowder photocatalysts for three concentrations were taken from the literature [27]. The experimental data set of dopants (at three concentrations) was randomly distributed into the training, test, and validation sets in order to develop a reliable quasi QSPR model. Table **1** lists the codes of various attributes of doped-titania nanopowders. No information on the external validation set was used to build up the model.

Data Set 2

The numerical data on the bacterial reverse mutation test that was conducted using Escherichia coli strain WP2 uvrA/pKM101 in the presence and absence of metabolic activation under dark condition and irradiation are taken from the literature [28]. Table **2** contains list of considered eclectic factors. In Table **2**, twenty possible characteristics of impact of fullerene, determined by different combinations of the eclectic factors are listed.

Distribution into the Training, Test, and Validation Sets

The principles of the distribution of available data into the training, test, and validation sets are the following: (i) they are random; but (ii) the range of endpoints for the training, test, and validation sets are as similar as possible.

Optimal Descriptors Calculated with Eclectic Data

The optimal descriptors have been calculated as follows:

$$DCW(T,N) = \Sigma CW(C_k) \quad (1)$$

where C_k is the SMILES attribute that comprises one symbol or two symbols which should be examined as one (Tables **1** and **2**). $CW(x)$ represents the correlation weight for an attribute x; the T is the threshold to divide attributes into two categories rare (noise) or not rare; the N is the number of epochs of the Monte Carlo optimization. Correlation weights are calculated for not rare attributes by the Monte Carlo optimization that gives maximum number of correlation coefficient between $DCW(T, N)$ and endpoint for the calibration set. The preferable values for the T^* and N^*, which gives the best statistics for the calibration set, should be defined at the preliminary phase of the QSAR analysis [29]. Having T^*, N^*, and $CW(x)$ which give maximum number of the correlation coefficient for the calibration set, one can define (using data from the training set) a model that could be described as:

$$EP = C_0 + C_1 \times DCW(T^*, N^*) \quad (2)$$

The predictability of the model for the endpoint (EP) should be checked with external validation set.

Table 1. Correlation weights of various dopants and their concentrations.

C_k	Split 1	Split 2	Split 3	Split 4	Split 5
	$CW(C_k)$	$CW(C_k)$	$CW(C_k)$	$CW(C_k)$	$CW(C_k)$
A	1.11217	1.18312	1.04657	1.09070	1.21578
B	1.04140	1.08472	0.95879	1.01400	1.15206
C	0.99112	1.04580	0.94555	0.97849	1.11740
Ag	1.19946	1.17897	1.13619	1.14847	1.13837
Ce	0.89122	0.92616	0.85408	0.89866	0.87587
Co	0.90180	0.94509	0.90437	0.89217	0.85969
Cr	0.87679	0.90454	0.85386	0.93856	0.87949
Fe	0.97044	0.99737	0.95842	0.95851	0.92321
Er	1.03573	1.03709	1.01230	0.97586	0.97912
Ga	0.95248	0.97525	0.94613	0.96023	0.91596
Gd	1.03921	1.02869	1.00842	0.99181	0.90248
La	0.98430	0.98916	0.95074	0.97549	0.95439
Nd	1.02463	1.01545	0.89599	0.94700	0.95052
Mn	0.93889	0.95268	0.97119	0.98785	0.95453
Ni	0.96193	1.00472	0.95436	0.97768	0.92608
V	0.83771	0.85484	0.84576	0.85968	0.82679
Pr	0.97063	1.04703	1.02104	1.03384	1.00110
Y	0.96347	1.05303	1.07993	0.97083	0.99093
Sr	1.37462	1.42675	1.30025	1.30003	1.22169
Zn	1.30390	1.22917	1.20492	1.21696	1.15176

An example of the *DCW*(1, 4) and DRC (10^{-5}/s) calculation for general code "ALa"

C_k	$CW(C_k)$
A	1.1122
La	0.9843
DCW(1,3)	***2.0965***

DRC (10^{-5}/s) = -93.1358 + 51.3416 * DCW(1,4) = 14.5019

Table 2. List of attributes of fullerene C60 nanoparticles exposure and their codes.

Attribute	Code of Attribute (C_k)
Dark or Irradiation	0 = Dark 1 = Irradiation
Mix S9	+ = with Mix S9 – = without Mix S9
Dose (g/plate)	A = 50 B = 100 C = 200 D = 400 E = 1000

RESULTS AND DISCUSSION

Data Set 1

Using the above approach, one-variable model has been developed. The models for rate constants DRC (10^{-5}/s) are expressed as follows:

Split 1

$$\text{DRC}\ (10^{-5}/\text{s}) = -93.13\ (\pm 2.45) + 51.34\ (\pm 1.21) * \text{DCW}(1,4) \quad (3)$$

$n = 32$, $r^2 = 0.8959$, $q^2 = 0.8606$, $s = 2.66$, $F = 258$, ${}^{c}R_p^2$=0.8735 (training set)

$n = 10$, $r^2 = 0.8645$, $s = 3.63$, ${}^{c}R_p^2$=0.8449 (test set)

$n = 9$, $r^2 = 0.6880$, $s = 4.37$ (validation set)

Split 2

DRC (10^{-5}/s) =-115.83 (± 1.17) + 60.66 (± 0.55) * DCW(1,3) (4)

n = 33, r^2 = 0.9118, q^2 = 0.9014, s = 2.37, F = 320, $^cR_p^2$=0.8872 (training set)

n = 9, r^2 = 0.8165, s = 6.02, $^cR_p^2$=0.7619 (test set)

n = 9, r^2 = 0.9660, s = 3.23 (validation set)

Split 3

DRC (10^{-5}/s) =-100.6 (± 2.22) + 57.58 (± 1.15) * DCW(1,5) (5)

n = 32, r^2 = 0.9002, q^2 = 0.8726, s = 2.76, F = 270, $^cR_p^2$=0.8865 (training set)

n = 11, r^2 = 0.8364, s = 3.88, $^cR_p^2$=0.7314 (test set)

n = 8, r^2 = 0.8253, s = 2.91 (validation set)

Split 4

DRC (10^{-5}/s) =-120.53 (± 1.03) + 65.76 (± 0.52) * DCW(1,4) (6)

n = 33, r^2 = 0.9256, q^2 = 0.9172, s = 2.09, F = 386, $^cR_p^2$=0.8988 (training set)

n = 9, r^2 = 0.8059, s = 5.05, $^cR_p^2$=0.7470 (test set)

n = 9, r^2 = 0.8688, s = 3.90 (validation set)

Split 5

DRC (10^{-5}/s) =-137.83 (± 3.02) + 70.75 (± 1.44) * DCW(1,4) (7)

n = 31, r^2 = 0.8957, q^2 = 0.8676, s = 2.91, F = 249, $^cR_p^2$=0.8844 (training set)

n = 12, r^2 = 0.6226, s = 3.86, $^cR_p^2$=0.6074 (test set)

n = 8, r^2 = 0.9100, s = 2.95 (validation set)

Table **1** contains the correlation weights for calculation with Eqs. 3-7. Table **3** contains the results of calculations with Eqs. 3-7 and distribution of available data into the training, calibration, and validation sets. Fig. (**1**) represents these models graphically.

Data Set 2.

One-variable models for the mutagenicity (*Escherichia coli* strain WP2 uvrA/pKM101 in the presence and absence of metabolic activation under dark condition and irradiation) of fullerene are the following:

Split 1

WP2uvrA/pKM101 = 28.26 (±12.20) + 41.43 (± 4.73) * DCW(3,20) (8)

n = 10, r^2 = 0.7487, q^2=05226, s = 18.1, F = 24, $^cR_p^2$=0.7001 (training set)

n = 5, r^2 = 0.8018, s = 25.9, $^cR_p^2$=0.6485 (calibration set)

n = 5, r^2 = 0.7480, s = 19.2 (validation set)

Split 2

WP2uvrA/pKM101 = 85.64 (± 8.81) + 17.64 (± 4.62) * DCW(2,9) (9)

n =6, r^2 = 0.8246, q^2=0.2665, s = 18.1, F = 19, $^cR_p^2$=0.6817 (training set)

n = 7, r^2 = 0.7109, s = 19.1, $^cR_p^2$=0.6828 (calibration set)

n = 7, r^2 = 0.7355, s = 13.1 (validation set)

Split 3

WP2uvrA/pKM101 =-1720.1 (±305.3) + 907.858 (±149.7) * DCW(4,13) (10)

n = 9, r^2 = 0.5870, q^2=0.2290, s = 24.2, F = 19, $^cR_p^2$=0.5345 (training set)

n = 6, r^2 = 0.7056, s = 17.2, $^cR_p^2$=0.6750 (calibration set)

n = 7, r^2 = 0.8967, s = 18.7 (validation set)

Split 4

WP2uvrA/pKM101 = 88.00 (± 6.00) + 15.38 (± 2.09) * DCW(3,8) (11)

n = 10, r^2 = 0.6521, q^2=0.3584, s = 14.4, F = 15, $^cR_p^2$=0.5987 (training set)

n = 5, r^2 = 0.8929, s = 35.7, $^cR_p^2$=0.7204 (calibration set)

n = 5, r^2 = 0.7198, s = 14.4 (validation set)

Split 5

WP2uvrA/pKM101 = 89.63 (± 4.54) + 15.43 (± 2.23) * DCW(2,9) (12)

n = 8, r^2 = 0.6551, q^2=0.3554, s = 24.6, F = 11, $^cR_p^2$=0.6189 (training set)

n = 6, r^2 = 0.7717, s = 15.3, $^cR_p^2$=0.6373 (calibration set)

n = 6, r^2 = 0.7860, s = 21.7 (validation set)

The statistical characteristics of models calculated with Eqs. 3-7 and Eqs.8-12 are the following: (i) the number of quasi SMILES in a set (n); (ii) the correlation coefficient (r^2); (iii) root-mean-square error (s); (iv) Fischer F-ratio (F); Y-randomization parameter [30] ($^cR_p^2$). A model is not chance correlation if $^cR_p^2$ is larger than 0.5 [30].

Table **4** contains the correlation weight for calculation with Eqs. 8-12. Fig. (**2**) represents the obtained model graphically.

In addition, in Table **5** the experimental and calculated values of the mutagenicity together with the distribution of the available data into the training, calibration, and validation sets are presented.

Having data on several runs of the Monte Carlo optimization one can extract four categories of molecular features: (i) features which have only positive values of the correlation weights - these can be classified as promoters of

Table 3. Experimental and calculated photocatalytic decolorisation rate constants, DRC (10^{-5}/s) and quasi SMILES which are used to represent the nanoparticles.

No.	Split					Quasi-SMILES	DRC (10^{-5}/s)	DRC (10^{-5}/s) Calculation				
	1*	2	3	4	5		Expr	Eq. 3	Eq.4	Eq. 5	Eq. 6	Eq. 7
1	E	T	T	T	E	AAg	24.90	25.5471	27.4737	25.0501	26.7233	28.7315
2	T	I	T	T	T	ALa	14.60	14.5019	15.9580	14.3717	15.3469	15.7141
3	E	T	T	T	T	APr	23.40	13.7985	19.4691	18.4198	19.1844	19.0194
4	T	I	T	I	T	ASr	26.90	34.5400	42.5059	34.4973	36.6909	34.6271
5	I	E	I	I	T	AZn	28.40	30.9089	30.5192	29.0081	31.2275	29.6793
6	T	T	I	T	T	AV	5.72	6.9742	7.8092	8.3274	7.7304	6.6865
7	T	T	T	E	T	AEr	22.10	17.1407	18.8661	17.9167	15.3712	17.4643
8	I	T	T	I	T	AY	22.00	13.4309	19.8328	21.8109	15.0404	18.2999
9	T	T	I	T	T	AMn	14.60	12.1688	13.7447	11.2192	13.4731	15.4403
10	T	T	E	I	I	ANd	17.10	16.5712	17.5531	15.5493	16.1599	15.7246
11	T	I	E	T	T	AFe	12.70	13.7889	16.4560	14.8144	14.2304	13.5083
12	T	T	T	T	I	AGd	21.50	17.3193	18.3565	17.6935	16.4201	12.0414
13	I	E	I	T	T	ACo	8.29	10.2645	13.2844	11.7022	9.8674	9.0143
14	T	E	T	T	I	ANi	12.80	13.3518	16.9019	14.5807	15.4912	13.7116
15	T	T	T	E	T	ACr	6.24	8.9806	10.8245	8.7934	12.9181	10.4150
16	T	T	E	T	T	ACe	12.50	9.7214	12.1359	8.8062	10.2939	10.1591
17	T	E	T	I	E	AGa	13.30	12.8665	15.1143	14.1062	14.3436	12.9950
18	T	T	E	T	T	BAg	23.90	21.9134	21.5040	19.9957	21.6791	24.2226
19	E	T	T	T	I	BLa	9.63	10.8667	9.9883	9.3173	10.3027	11.2052
20	T	E	I	I	I	BPr	10.70	10.1648	13.4994	13.3654	14.1402	14.5105
21	T	E	T	T	T	BSr	31.60	30.9063	36.5362	29.4429	31.6466	30.1182
22	I	T	T	T	E	BZn	24.20	27.2752	24.5495	23.9537	26.1833	25.1704
23	E	T	T	E	T	BV	2.89	3.3405	1.8395	3.2730	2.6861	2.1776
24	T	I	I	I	T	BEr	11.40	13.5070	12.8964	12.8623	10.3270	12.9554
25	E	T	E	I	E	BY	15.50	9.7973	13.8631	16.7565	9.9961	13.7910
26	T	T	T	E	E	BMn	5.83	8.5352	7.7750	6.1648	8.4289	10.9314
27	T	E	T	E	T	BNd	12.10	12.9375	11.5834	10.4949	11.1156	11.2157
28	T	T	T	T	E	BFe	10.80	10.1552	10.4863	9.7600	9.1862	8.9994
29	E	T	T	T	T	BGd	8.62	13.6856	12.3868	12.6390	11.3759	7.5325
30	T	T	T	T	I	BCo	6.98	6.6308	7.3147	6.6478	4.8232	4.5054
31	I	T	T	E	T	BNi	9.21	9.7181	10.9322	9.5263	10.4469	9.2027
32	T	T	E	E	T	BCr	8.06	5.3469	4.8548	3.7390	7.8739	5.9061
33	T	I	T	T	T	BCe	4.57	6.0877	6.1662	3.7518	5.2496	5.6502
34	E	T	T	T	I	BGa	9.85	9.2328	9.1446	9.0518	9.2994	8.4861
35	T	T	T	T	I	CAg	19.30	19.3319	19.1425	19.2332	19.3440	21.7709
36	I	I	E	T	T	CLa	10.20	8.2853	7.6267	8.5548	7.9675	8.7534
37	E	T	T	T	T	CPr	7.34	7.5834	11.1378	12.6029	11.8050	12.0587
38	T	T	T	E	T	CSr	33.80	28.3249	34.1747	28.6804	29.3115	27.6665
39	T	I	I	T	T	CZn	25.20	24.6938	22.1880	23.1911	23.8482	22.7187
40	T	T	T	T	I	CV	2.47	0.7591	-0.5221	2.5105	0.3510	-0.2741

(Table 3) contd…..

No.	Split					Quasi-SMILES	DRC (10^{-5}/s)	DRC (10^{-5}/s) Calculation				
	1*	2	3	4	5		Expr	Eq. 3	Eq.4	Eq. 5	Eq. 6	Eq. 7
41	T	T	T	T	T	CEr	7.71	10.9255	10.5349	12.0998	7.9918	10.5036
42	T	T	I	T	T	CY	7.17	7.2158	11.5016	15.9940	7.6610	11.3392
43	I	I	E	T	I	CMn	4.59	5.9537	5.4135	5.4022	6.0938	8.4797
44	T	E	T	T	I	CNd	8.10	10.3561	9.2218	9.7324	8.7805	8.7640
45	T	T	T	E	I	CFe	8.45	7.5738	8.1247	8.9975	6.8511	6.5476
46	T	I	I	T	E	CGd	7.00	11.1042	10.0253	11.8765	9.0408	5.0807
47	I	E	I	I	E	CCo	5.11	4.0494	4.9532	5.8852	2.4881	2.0536
48	I	T	T	T	I	CNi	10.50	7.1367	8.5706	8.7638	8.1118	6.7509
49	I	T	T	T	T	CCr	5.46	2.7655	2.4932	2.9765	5.5387	3.4543
50	T	T	T	T	T	CCe	2.70	3.5063	3.8047	2.9893	2.9145	3.1985
51	E	T	I	T	T	CGa	6.02	6.6513	6.7831	8.2893	6.9642	6.0344

*) T=training set; I=test set (internal); E=validation set (external).

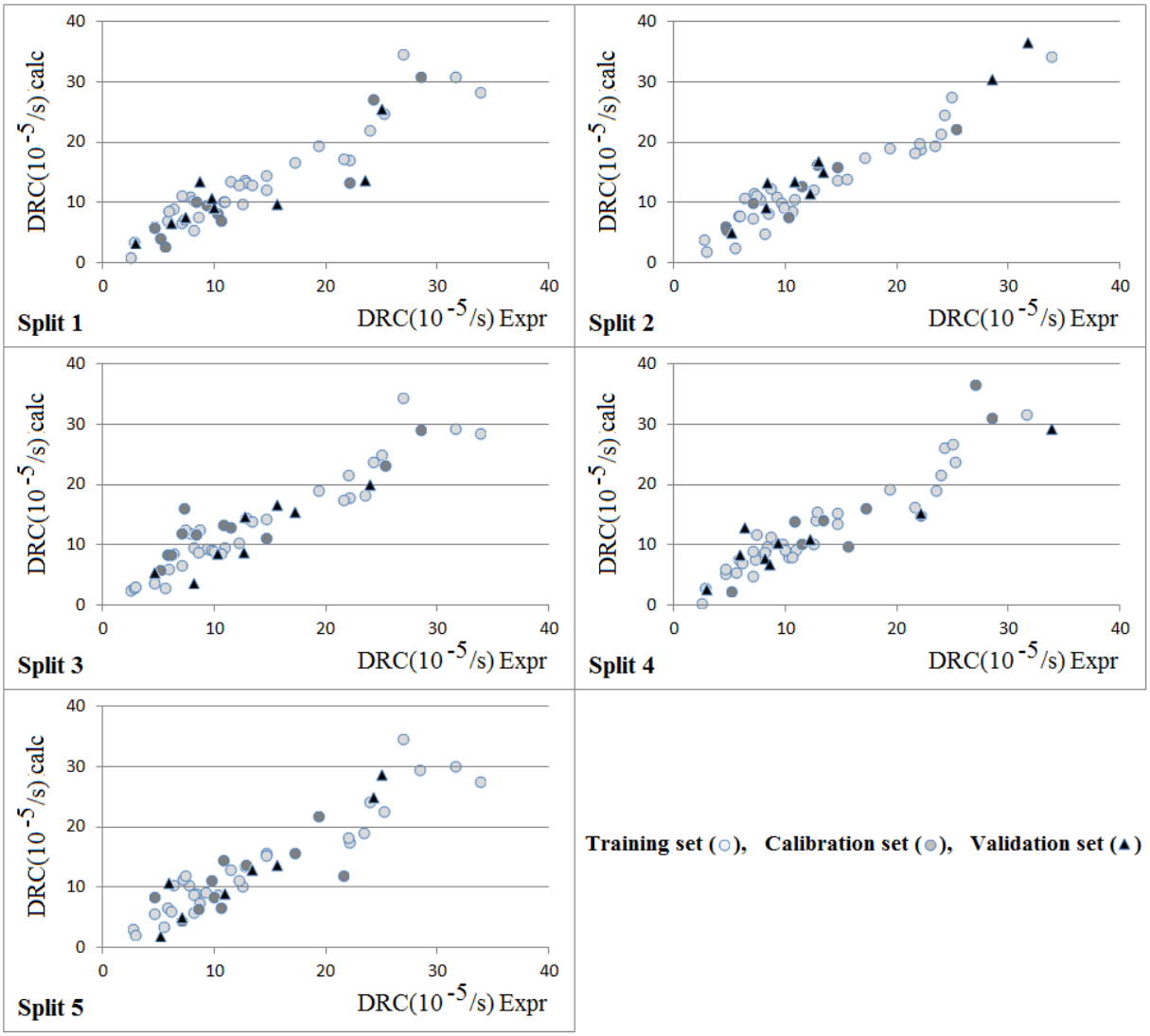

Fig. (1). Graphic representations of model for photocatalytic decolorisation rate constants, DRC (10^{-5}/s) calculated with Eqs.3-7.

Table 4. Correlation weights for calculation of the mutagenicity (*Escherichia coli* strain WP2 uvrA/pKM101) of fullerene.

	Split 1	Split 2	Split 3	Split 4	Split 5
C_k	$CW(C_k)$	$CW(C_k)$	$CW(C_k)$	$CW(C_k)$	$CW(C_k)$
+	0.82473	0.42477	0.95219	0.19521	0.01264
-	2.02643	2.47844	1.00393	2.59809	2.98026
0	1.83643	2.29125	1.08846	1.82572	2.63213
1	1.02221	0.15010	1.05207	0.53736	0.64512
A	0.0	0.0	0.0	0.0	0.01132
B	-0.65016	0.0	0.0	0.19987	0.0
C	-0.32615	0.0	0.0	0.0	0.00888
D	0.0	0.10014	0.0	0.0	0.45180
E	0.0	1.60026	0.0	0.0	0.0

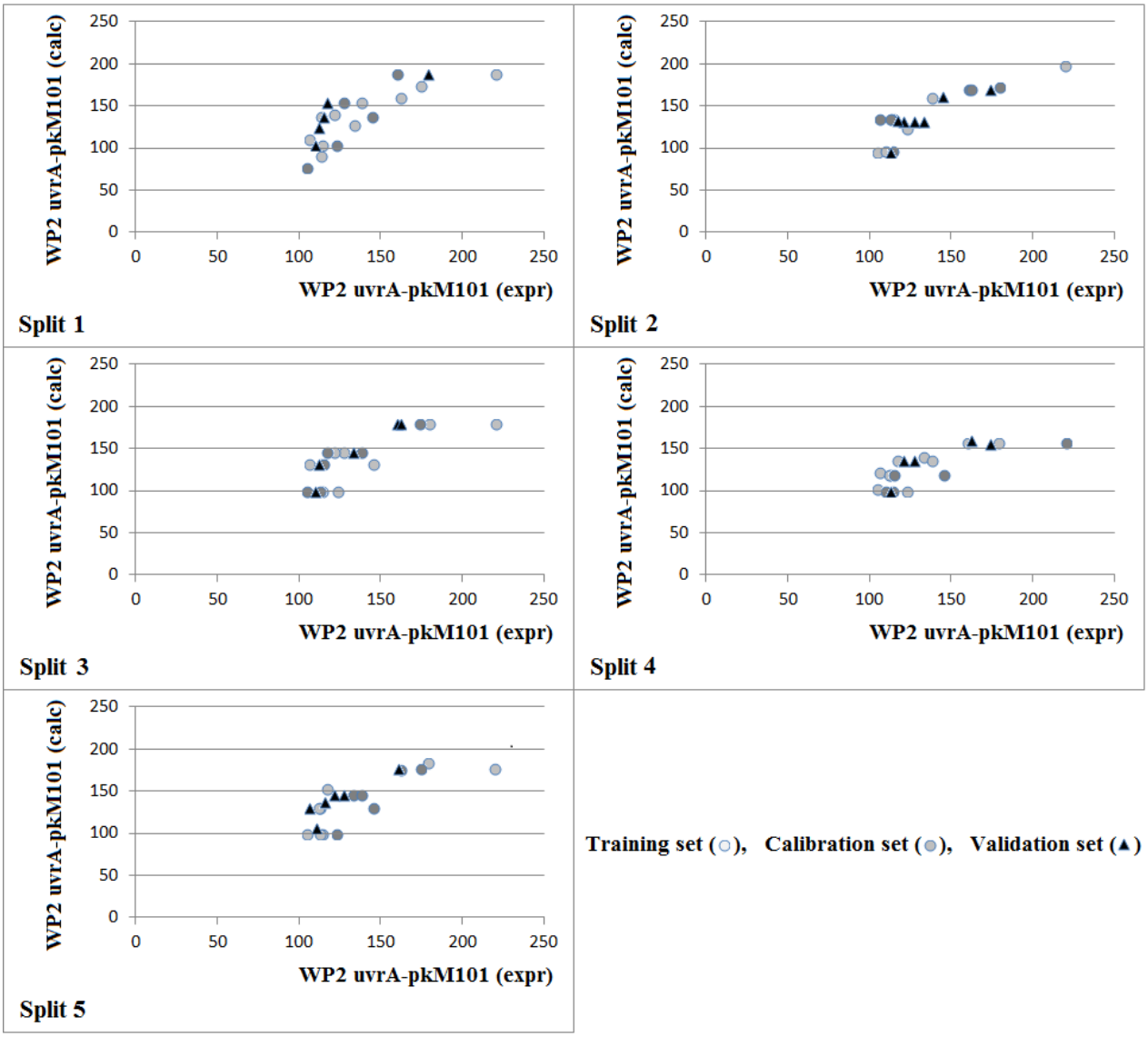

Fig. (2). Graphical representation of models for mutagenicity (*Escherichia coli* strain WP2 uvrA/pKM101) of fullerene calculated with Eqs. 8-12.

Table 5. Three splits of data into the training, calibration, and validation sets; experimental and calculated values on the bacterial reverse mutation test WP2uvrA/pKM101.

No.	Split					Quasi-SMILES	Mutagenicity Expr	Mutagenicity Calculation				
	1*	2	3	4	5			Eq.8	Eq. 9	Eq. 10	Eq.11	Eq.12
1	T	I	T	T	T	0+A	113.0	138.5293	133.5675	132.4706	119.1024	130.6360
2	T	I	T	T	E	0+B	106.0	111.5896	133.5675	132.4706	122.1779	130.4612
3	E	I	E	T	T	0+C	112.0	125.0150	133.5675	132.4706	119.1024	130.5983
4	E	T	I	I	E	0+D	115.0	138.5293	135.3343	132.4706	119.1024	137.4352
5	I	E	T	I	I	0+E	145.0	138.5293	161.8008	132.4706	119.1024	130.4612
6	I	I	E	T	E	0-A	160.0	188.3225	169.8004	179.4440	156.0775	176.4445
7	T	I	E	E	I	0-B	162.0	161.3827	169.8004	179.4440	159.1531	176.2697
8	T	E	I	E	I	0-C	174.0	174.8081	169.8004	179.4440	156.0775	176.4069
9	E	I	T	T	T	0-D	179.0	188.3225	171.5672	179.4440	156.0775	183.2438
10	T	T	T	I	T	0-E	220.0	188.3225	198.0337	179.4440	156.0775	176.2697
11	T	I	T	I	T	1+A	114.0	104.7914	95.7915	99.4383	99.2771	99.9642
12	I	T	I	T	T	1+B	105.0	77.8516	95.7915	99.4383	102.3527	99.7894
13	T	E	I	E	T	1+C	113.0	91.2770	95.7915	99.4383	99.2771	99.9266
14	E	T	E	I	E	1+D	110.0	104.7914	97.5583	99.4383	99.2771	106.7635
15	I	T	T	T	I	1+E	123.0	104.7914	124.0247	99.4383	99.2771	99.7894
16	I	E	T	E	E	1-A	127.0	154.5845	132.0243	146.4117	136.2523	145.7728
17	T	E	E	T	I	1-B	133.0	127.6447	132.0243	146.4117	139.3279	145.5980
18	T	E	T	E	E	1-C	121.0	141.0701	132.0243	146.4117	136.2523	145.7351
19	E	E	I	T	T	1-D	117.0	154.5845	133.7911	146.4117	136.2523	152.5720
20	T	T	I	T	I	1-E	138.0	154.5845	160.2576	146.4117	136.2523	145.5980

*) T=training set; I=test set (internal); E=validation set (external).

an endpoint increase; (ii) features which have only negative values of the correlation weights - these can be classified as promoters of the endpoint decrease; (iii) features which have both positive and negative correlation weights for different runs of the Monte Carlo optimization - these can be classified as features with unclear role; finally, (iv) features which are blocked according to the used threshold. Therefore, the used approach gives models which can have a mechanistic interpretation. Thus, the described models are calculated in accordance with OECD principles [31].

Table **6** contains statistical checking up of the models for DRC (10^{-5}/s) and WP2uvrA/pKM101 according to criteria of reliability of QSPR/QSAR models taken in the literature [32, 33].

In the two considered cases it is not clear how one can involve traditional descriptors to build up predictive models for two endpoints related to nanomaterials. Our study indicates that the proposed approach provides reasonable accurate predictions. One should also mentioned that the recently updated CORAL software (http://www.insilico.eu/coral) can estimate the quality of different distribution into the visible training and calibration sets and invisible validation set from statistical (probabilistic) point of view, i.e. automatically define the domain of applicability for the above-mentioned different distributions.

It is to be noted, that the described approach has been tested as a tool to build up quasi-QSAR for the bacterial reverse mutation test (TA100) for fullerene under the same conditions which are examined in this work (for WP2uvrA/pKM101) [34] and for membrane damage induced by impacts of TiO_2 and ZnO nanoparticles under various conditions [19, 35].

CONCLUSION

Two set of data related to nanomaterials have been considered in this study. The Monte Carlo technique based on the eclectic data was used to build up predictive models for the photocatalytic decolorisation rate constants DRC (10^{-5}/s) and for the mutagenicity of fullerene (*Escherichia coli* strain WP2 uvrA/pKM101) under various (eclectic) conditions. In both cases reasonably accurate results were obtained. We believe that this approach can be used for the "quasi-QSPR/QSAR analyses" of other endpoints related to nanomaterials.

Table 6. The statistical characteristics of models for the external validation sets.

Endpoint	Split	The Statistical Quality for Validation Set*
DRC (10^{-5}/s)	1	$n = 9$ $r^2 = 0.6880$ $r_0^2 = 0.6232$ $r'^2_0 = 0.6854$ $\frac{r^2 - r_0^2}{r^2} = 0.0942 < 0.1$ $\frac{r^2 - r'^2_0}{r^2} = 0.0038 < 0.1$ $k = 1.0605(0.85 < k < 1.15)$ $k' = 0.8641(0.85 < k' < 1.15)$ $r_m^2 = 0.6528 > 0.5$ $\overline{r_m^2} = 0.5828 > 0.5, \Delta r_m^2 = 0.1400 < 0.2$
	2	$n = 9$ $r^2 = 0.9660$ $r_0^2 = 0.9652$ $r'^2_0 = 0.9625$ $\frac{r^2 - r_0^2}{r^2} = 0.0008 < 0.1$ $\frac{r^2 - r'^2_0}{r^2} = 0.0036 < 0.1$ $k = 0.8671(0.85 < k < 1.15)$ $k' = 1.1426(0.85 < k' < 1.15)$ $r_m^2 = 0.9393 > 0.5$ $\overline{r_m^2} = 0.9242 > 0.5, \Delta r_m^2 = 0.0302 < 0.2$
	3	$n = 8$ $r^2 = 0.8253$ $r_0^2 = 0.8253$ $r'^2_0 = 0.7935$ $\frac{r^2 - r_0^2}{r^2} = 0.0000 < 0.1$ $\frac{r^2 - r'^2_0}{r^2} = 0.0386 < 0.1$ $k = 0.8960(0.85 < k < 1.15)$ $k' = 1.0810(0.85 < k' < 1.15)$ $r_m^2 = 0.8224 > 0.5$ $\overline{r_m^2} = 0.7502 > 0.5, \Delta r_m^2 = 0.1443 < 0.2$
	4	$n = 9$ $r^2 = 0.8688$

(Table 6) contd…..

Endpoint	Split	The Statistical Quality for Validation Set*
DRC (10^{-5}/s)	4	$r_0^2 = 0.8043$ $r'^2_0 = 0.8558$ $\frac{r^2 - r_0^2}{r^2} = 0.0742 < 0.1$ $\frac{r^2 - r'^2_0}{r^2} = 0.0149 < 0.1$ $k = 1.0819(0.85 < k < 1.15)$ $k' = 0.8751(0.85 < k' < 1.15)$ $r_m^2 = 0.7700 > 0.5$ $\overline{r_m^2} = 0.7091 > 0.5, \Delta r_m^2 = 0.1219 < 0.2$
	5	$n = 8$ $r^2 = 0.9100$ $r_0^2 = 0.9024$ $r'^2_0 = 0.8768$ $\frac{r^2 - r_0^2}{r^2} = 0.0084 < 0.1$ $\frac{r^2 - r'^2_0}{r^2} = 0.0365 < 0.1$ $k = 1.0376(0.85 < k < 1.15)$ $k' = 0.9363(0.85 < k' < 1.15)$ $r_m^2 = 0.8305 > 0.5$ $\overline{r_m^2} = 0.7873 > 0.5, \Delta r_m^2 = 0.0864 < 0.2$
Mutagenicity, WP2uvrA/ pKM101	1	$n = 5$ $r^2 = 0.7475$ $r_0^2 = 0.7395$ $r'^2_0 = 0.7157$ $\frac{r^2 - r_0^2}{r^2} = 0.0106 < 0.1$ $\frac{r^2 - r'^2_0}{r^2} = 0.0425 < 0.1$ $k = 1.1155(0.85 < k < 1.15)$ $k' = 0.8868(0.85 < k' < 1.15)$ $r_m^2 = 0.6808 > 0.5$ $\overline{r_m^2} = 0.6475 > 0.5, \Delta r_m^2 = 0.0666 < 0.2$
	2	$n = 7$ $r^2 = 0.7355$ $r_0^2 = 0.7337$ $r'^2_0 = 0.6735$

(Table 6) contd…..

Endpoint	Split	The Statistical Quality for Validation Set*
Mutagenicity, WP2uvrA/ pKM101	2	$\frac{r^2-r_0^2}{r^2}=0.0025<0.1$ $\frac{r^2-r'^2_0}{r^2}=0.0843<0.1$ $k=1.0283(0.85<k<1.15)$ $k'=0.9658(0.85<k'<1.15)$ $r_m^2=0.7040>0.5$ $\overline{r_m^2}=0.6282>0.5,\Delta r_m^2=0.1517<0.2$
	3	$n=5$ $r^2=0.8967$ $r_0^2=0.8775$ $r'^2_0=0.8143$ $\frac{r^2-r_0^2}{r^2}=0.0214<0.1$ $\frac{r^2-r'^2_0}{r^2}=0.0919<0.1$ $k=1.0940(0.85<k<1.15)$ $k'=0.9096(0.85<k'<1.15)$ $r_m^2=0.7725>0.5$ $\overline{r_m^2}=0.7059>0.5,\Delta r_m^2=0.1331<0.2$
	4	$n=5$ $r^2=0.7198$ $r_0^2=0.7173$ $r'^2_0=0.6526$ $\frac{r^2-r_0^2}{r^2}=0.0036<0.1$ $\frac{r^2-r'^2_0}{r^2}=0.0934<0.1$ $k=1.0132(0.85<k<1.15)$ $k'=0.9789(0.85<k'<1.15)$ $r_m^2=0.6834>0.5$ $\overline{r_m^2}=0.6083>0.5,\Delta r_m^2=0.1503<0.2$
	5	$n=6$ $r^2=0.7860$ $r_0^2=0.7777$ $r'^2_0=0.7680$ $\frac{r^2-r_0^2}{r^2}=0.0229<0.1$ $\frac{r^2-r'^2_0}{r^2}=0.0105<0.1$ $k=1.1380(0.85<k<1.15)$ $k'=0.8745(0.85<k'<1.15)$ $r_m^2=0.7146>0.5$ $\overline{r_m^2}=0.6976>0.5,\Delta r_m^2=0.0340<0.2$

*) The statistical characteristics taken in the literature [32,33].

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Authors thank EC project PreNanoTox (contract 309666), the EC project NanoPUZZLES (Project Reference: 309837), J.L. and D.L. thank the National Science Foundation (NSF/CREST HRD-0833178, and EPSCoR (Award #: 362492-190200-01/NSFEPS-090378) for financial support. A.M.V. and J.B.V. would like to thank Ministry of Education and Science, Republic of Serbia (Project Number 172044) for financial support. P.R.D. acknowledges the financial support from the National Research Council of Argentina (CONICET) PIP11220100100151 project. P.R.D. and E.A.C. are the members of the scientific researcher career of CONICET. Authors also express their gratitude to Dr. L. Cappellini, Dr. G. Bianchi and Dr. R. Bagnati for valuable consultations in the computer science.

REFERENCES

[1] Nel, A. Xia, T.; Mädler, L.; Li, N. Toxic potential of materials at the nanolevel. *Science* **2006**, 311 (5761), 622-627.

[2] Nel, A.E.; Mädler, L.; Velegol, D.; Xia, T.; Hoek, E.M.; Somasundaran, P.; Klaessig, F.; Castranova, V.; Thompson, M. Understanding biophysicochemical interactions at the nano-bio interface. *Nat. Mater.* **2009**, 8 (7), 543-57.

[3] West, J.L.; Halas, N.J. Engineered nanomaterials for biophotonics applications: improving sensing, imaging, and therapeutics. *Annu. Rev. Biomed. Eng.* **2003**, 5, 285-92.

[4] Zhang, L.; Gu, F.X.; Chan, J.M.; Wang, A.Z.; Langer, R.S.; Farokhzad, O.C. Nanoparticles in medicine: Therapeutic applications and developments. *Clin. Pharmacol. Ther.* **2008**, 83 (5), 761-769.

[5] Bianco, A.; Kostarelos, K.; Prato, M. Applications of carbon nanotubes in drug delivery. *Curr. Opin. Chem. Biol.* **2005**, 9 (6), 674-679.

[6] Hoet, P.H.; Brüske-Hohlfeld, I.; Salata, O.V. Nanoparticles - known and unknown health risks. *J. Nanobiotechnology.* **2004**, 2 (1), 12.

[7] Lewinski, N.; Colvin, V.; Drezek, R. Cytotoxicity of nanopartides. *Small* **2008**, 4 (1), 26-49.

[8] Hansen, F.S.; Maynard, A.; Baun, A.; Tickner, J.A. Late lessons from early warnings for nanotechnology. *Nat. Nanotechnol.* **2008**, 3 (8), 444-447.

[9] Cattaneo, A.G.; Gornati, R.; Sabbioni, E.; Chiriva-Internati, M.; Cobos, E.; Jenkins, M.R.; Bernardini, G. Nanotechnology and human health: risks and benefits. *J. Appl. Toxicol.* **2010**, 30 (8), 730-744.

[10] Heinemann, M.; Schafer, H.G. Guidance for handling and use of nanomaterials at the workplace. *Hum. Exp. Toxicol.* **2009**, 28 (6-7), 407-411.

[11] Clark, K.A.; White, R.H.; Silbergeld, E.K. Predictive models for nanotoxicology: current challenges and future opportunities. *Regul. Toxicol. Pharmacol.* **2011**, 59 (3), 361-363.

[12] Køppe, S. A Moderate Eclecticism: Ontological and Epistemological Issues. *Integr. Psych. Behav.* **2012**, 46 (1), 1-19.

[13] Duchowicz, P.R.; Bucknum, M.J.; Castro, E.A. New molecular descriptors based upon the Euler equations for chemical graphs. *J. Math. Chem.* **2007**, 41 (2), 193-208.

[14] Duchowicz, P.R.; Castro, E.A. QSPR studies on aqueous solubilities of drug-like compounds. *Int. J. Mol. Sci.* **2009**, 10 (6), 2558-2577.

[15] Pramanik, S.; Roy, K. Modeling bioconcentration factor (BCF) using mechanistically interpretable descriptors computed from open source tool "PaDEL-Descriptor". *Environ. Sci. Pollut. Res*. **2014**, 21 (4), 2955-2965.

[16] Toropova, A.P.; Toropov, A.A. CORAL software: Prediction of carcinogenicity of drugs by means of the Monte Carlo method. *Eur. J. Pharm. Sci*. **2014**, 52 (1), 21-25.

[17] Burello., E.; Worth, A. P.: QSAR modeling of nanomaterials. WIREs Nanomed. Nanobiotechnol. **2011**, 3 (1), 298–306.

[18] Winkler, D.A.; Mombelli, E.; Pietroiusti A.; Tran L.; Worth, A.; Fadeel, B.; McCall, M.J.: Applying quantitative structure–activity relationship approaches to nanotoxicology: Current status and future potential. Toxicology **2013**, 313 (1), 15–23.

[19] Toropova, A.P.; Toropov, A.A. Optimal descriptor as a translator of eclectic information into the prediction of membrane damage by means of various TiO_2 nanoparticles. *Chemosphere*, **2013**, 93 (10), 2650-2655.

[20] Toropova, A.P.; Toropov, A.A.; Puzyn, T.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. Optimal descriptor as a translator of eclectic information into the prediction of thermal conductivity of micro-electro-mechanical systems. *J. Math. Chem.* **2013**, 51 (8), 2230-2237.

[21] Estrada, E.; Gutman, I. A topological index based on distances of edges of molecular graphs. *J. Chem. Inf. Comput. Sci.* **1996**, 36 (4), 850-853.

[22] Gutman, I.; Estrada, E. Topological indices based on the line graph of the molecular graph. *J. Chem. Inf. Comput. Sci.* **1996**, 36 (3), pp. 541-543.

[23] Gutman, I.; Rücker, C.; Rücker, G. On Walks in Molecular Graphs (2001) *J. Chem. Inf. Comput. Sci.* **2001**, 41 (3), 739-745.

[24] Toropov, A.A.; Gutman, I.; Furtula, B. Graph of atomic orbitals and the molecular structure-descriptors based on it. *J. Serbian Chem. Soc*. **2005**, 70 (4), 669-674.

[25] Toropov, A.A.; Leszczynska, D.; Leszczynski, J. QSPR study on solubility of fullerene C60 in organic solvents using optimal descriptors calculated with SMILES. *Chem. Phys. Lett.* **2007**, 441 (1-3), 119-122.

[26] Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. SMILES-based optimal descriptors: QSAR analysis of fullerene-based HIV-1 PR inhibitors by means of balance of correlations. *J. Comput. Chem.* **2010**, 31 (2), 381-392.

[27] Yang, Y.; Lin, T.; Weng, X.L.; Darr, J.A.; Wang, X.Z. Data flow modeling, data mining and QSAR in high-throughput discovery of functional nanomaterials. *Comput. Chem. Eng*, **2011**, 35 (4), 671-678.

[28] Shinohara, N.; Matsumoto, K.; Endoh, S.; Maru, J.; Nakanishi, J. and genotoxicity tests on fullerene C60 nanoparticles. *Toxicol. Lett*. **2009**, 191 (2-3), 289-296.

[29] Toropova, A.P.; Toropov, A.A.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. CORAL: Quantitative structure-activity relationship models for estimating toxicity of organic compounds in rats. *J. Comput. Chem*. **2011**, 32 (12), 2727-2733.

[30] Ojha, P.K.; Roy, K. Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection. *Chemometr. Intell. Lab. Syst*. **2011**, 109 (2), 146-161.

[31] OECD Principles For The Validation, For Regulatory Purposes, Of (Quantitative) Structure-Activity Relationship Models, **2004**, http://www.oecd.org/env/ehs/risk-assessment/37849783.pdf

[32] Golbraikh, A., Tropsha, A. Beware of q^2! (2002) *J. Mol. Graph. Modell*. **2002**, 20 (4), 269-276.

[33] Ojha, P.K.; Mitra, I., Das, R.N.; Roy, K. Further exploring r_m^2 metrics for validation of QSPR models. *Chemometr. Intell. Lab. Syst.* **2011**, 107 (1), 194-205.

[34] Toropov, A.A.; Toropova, A.P. Optimal descriptor as a translator of eclectic data into endpoint prediction: Mutagenicity of fullerene as a mathematical function of conditions. *Chemosphere* **2014**, 104, 262–264.

[35] Toropova, A.P.; Toropov, A.A.; Benfenati, E.; Puzyn, T.; Leszczynska, D.; Leszczynski, J. Optimal descriptor as a translator of eclectic information into the prediction of membranedamage:The case of a group of ZnO and TiO2 nanoparticles. *Ecotox. Environ. Safe.* **2014**, 108, 203–209

.

Received: July 25, 2014 **Revised: September 25, 20145** **Accepted: November 10, 2014**