Journal of Mathematical Chemistry, Vol. 43, No. 2, February 2008 (© 2006)
DOI: 10.1007/s10910-006-9214-6

Application of a novel ranking approach in QSPR-QSAR

Pablo R. Duchowicz*, Eduardo A. Castro, and Francisco M. Fernández

Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), División Química Teórica, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Diag. 113 y 64, Suc. 4, C.C. 16, (1900) La Plata, Argentina
E-mail: duchow@inifta.unlp.edu.ar

Received 27 August 2006; accepted 20 September 2006

In this study we present a simple algorithm based on the Partial Order Ranking (POR) technique which allows to rank a series of compounds according to their molecular descriptor values. A training set composed of 82 normal boiling points for structurally diverse organic compounds is analyzed by considering a pool of 1202 molecular descriptors obtained from the Dragon 5 software and two "flexible" type of variables. The predictive performance of the proposed approach is assessed by means of a test set of 82 "unknown" structurally related molecules.

KEY WORDS: Partial Order Ranking, QSPR-QSAR, molecular descriptor, normal boiling point

AMS Subject Classification: 78P05, 06P12

1. Introduction

The prediction of physicochemical and biological data of substances through the application of Quantitative Structure Property-Activity Relationships Theory (QSPR-QSAR) has acquired an increasing importance in the last decades. This is specially so when the experimental values of an endpoint can not be determined in the laboratory due to several circumstances, such as economical reasons or simply because the measurements demand too much time. The QSPR-QSAR studies are considered to be the most effective computational approaches for the estimation of different type of properties [1].

Although there is a great number of definitions for molecular descriptors available in the literature, it is well known that a single variable is unable to carry all the information on molecular structure, and this leads to the employment of more parameters in the QSPR-QSAR relationship. Nowadays, different standard statistical methods constitute a common practice for the model design,

*Corresponding author.

such as Multivariable Linear Regression (MLR) [2], Principal Component Analysis (PCA) [3], Partial Least Squares (PLS) [4], Genetics Algorithms [5–7] or Artificial Neural Networks (ANN) [8]. However, all of these elaborated techniques require the knowledge of a specific functional form of the model (linear or non-linear) and also optimized regression parameters to be present in the equation which, however, may not lead to the best results. QSPR-QSAR studies are usually based on such complex statistical analyzes and sophisticated local and global descriptor definitions.

The Partial Order Ranking (POR) approach provides with an interesting alternative and simplified approach to establish the desired structure–property connection, as it does not depend on Statistics. It represents a parameter-free technique that avoids the definition of analytical functional relationships and the use of optimized parameters. A consequence of this is that it also constitutes an obvious advantage to remedy the lack of availability of experimental data, one of the main drawbacks in statistical procedures.

The aim of this research is to introduce a new algorithm developed recently in our group that considers different sorting ideas. The development of an efficient and practical technique for performing the best quality predictions of a given endpoint is not an easy task. Many programs have to be written to test things and to decide whether the searched algorithm tend to reproduce the "tendencies" in the numerical data. For the case of ranking-based algorithms, this design involves two main objectives:

(a) locating the upper and lower limits of the experimental property interval within which a compound is to be predicted.
(b) performing the prediction in the selected interval in the best possible manner by resorting to interpolation formulae.

Clearly, if the algorithm is unable to position a compound X in an adequate interval, then the predictions performed in (b) will result of poor quality. Step (a) is a key step. A good interval can be understood as one that is able to position compounds from both the training and test sets with accuracy. The smaller the length of the interval is, the better the predictions performed will be in (b). The present study concerns with step (a) by using one or more molecular descriptors for ranking.

2. Molecular descriptors and data set

All the structures of the compounds were preoptimized by means of the Molecular Mechanics Force Field (MM+) included in Hyperchem version 6.03 [9]. Since various molecules contain sulfur atoms, final refined molecular structures were obtained using the semiempirical method PM3 (Parametric Method-3). We chose a gradient norm limit of 0.01 kcal/Å for the geometry optimization.

Several types of molecular descriptors were derived, such as constitutional, topological, geometrical, charge, GETAWAY (GEometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), molecular walk counts, BCUT descriptors, 2D-Autocorrelations, aromaticity indices, Randic molecular profiles, radial distribution functions, functional groups and atom centered fragments, by means of the software Dragon version 5 available in the Web for evaluation [10]. We excluded the empirical and property-based descriptors. In addition, two flexible molecular descriptors were added to this pool of variables: the so-called Correlation Weighting of Atomic Orbitals with Extended Connectivity of Zero- and First-Order (DCW^1 and DCW^2), based in the Graph of Atomic Orbitals (GAO) [11].

The data set employed consists on a representative set of 200 normal boiling points (NBP) of organic molecules studied earlier [12]. In this set it is found that 36 compounds do not obey the Similarity Principle [13], that is, NBP is a property that includes degenerated values and assigns the same number to several substances, even though different structures are involved. This type of conflicting molecules were removed from the set, thus leading to 164 molecules to be analyzed.

Since there are many molecules for calibrating the model, we decided to partition the set into two subsets composed of 82 structures, one for training the model and the other for testing its predictive performance. Notice that, as the POR relationship does not depend on regression coefficients, the size of both subsets can be the same, contrary to the case appearing usually in regression-based analyzes when dealing with a great number of data [14]. The compounds belonging to both the training and test series were chosen in such a way to have a representative sample of experimental NBP values in both subsets, and are shown in Tables 1 and 2. In other words, the members of each series were chosen in such a way that the experimental values of the compounds of the test set can be interpolated in the training set.

3. Principles of POR's based QSPR-QSAR

The methodology of POR [15] involves an extremely simple principle from the mathematical point of view. Consider a subset $\mathbf{d} = \{d_1, \ldots, d_i\}$ containing $i = 1, \ldots, d$ molecular descriptors, usually called an information basis (IB). If a compound A is characterized with the subset $\mathbf{d}$(A), and a compound B with the subset $\mathbf{d}$(B), then two compounds A and B exhibiting an experimental property p can be compared (ranked) through comparison (ranking) of their single descriptor values according the binary relation "$\leq$". That is to say,

$$p_{\mathrm{B}} \leq p_{\mathrm{A}} \leftrightarrow d_i(\mathrm{B}) \leq d_i(\mathrm{A}) \quad \text{for all } i = 1, \ldots, d \tag{1}$$

Table 1
Experimental values of NBP (°C) for the training set (82 compounds).

ID	Compound name	Exp.
1	1,4-pentadiene	26
2	Ethene-ethoxy-	33
3	methane.isocyanate-	37
4	1-propene-3-chloro-	44
5	2-propenal	53
6	3-nitrosostyrene	56
7	pyridine.4-ethenyl-	62
8	1-propanamine-2-methyl-	64
9	1,6,10-dodecatrien-3-ol-3,7,11-trimethyl-	68
10	silane-dichloro-dimethyl-	70
11	thiirane-methyl-	72
12	carbonochloridic acid-phenyl ester	74
13	1-butanamine	78
14	acetonitrile	81
15	2-octenal-(E)-	84
16	trimethyl-2-hydroxyethylsilane	90
17	Trifluoroaniline	92
18	*N*-octyltriethoxysilane	98
19	2-propenoic acid-2-methyl ester	100
20	2-butanol-2-methyl-	102
21	piperidine	106
22	(propargyloxy)trimethylsilane	110
23	1-penten-3-ol	114
24	5-hydroxydecanoic acid-lactone	117
25	butanoic acid ethyl ester	120
26	1-propanamine-3-(triethoxysilyl)-	122
27	ethanol-2-methoxy-	125
28	hexaethyldisiloxane	129
29	benzene-chloro-	132
30	propanoic acid-3-bromo-ethyl ester	135
31	piperazine-1-methyl-	138
32	3-ethyl-3-pentanol	141
33	quinoline-8-methyl-	143
34	1,4-oxathiane	147
35	Dibenzofuran	154
36	1H-indole-3-acetonitrile	157
37	urea-allyl-	163
38	2,4,6,8,10-pentamethylcyclopentasiloxane	168
39	2-propanol-1-3-dichloro-	174
40	*N*, *N*-diethylformamide	176
41	tri-*n*-butylphosphate	180
42	benzeneamine-2-fluoro-	182
43	benzene-1,4-diethyl-	184
44	di-*tert*-butyldichlorosilane	190
45	benzenemethanethiol	194

Table 1
Continued.

ID	Compound name	Exp.
46	benzene-[(phenylmethyl)thio]	197
47	benzene-1,1′-methylene bis 4-isocyanato-	200
48	2-bromopropanoic acid	203
49	4-chloro-3-cyanobenzotrifluoride	210
50	1-chloro-4-isopropenylbenzene	214
51	benzene methanol-4-methyl-	217
52	malononitrile	220
53	propanenitrile-3-(triethoxysilyl)-	224
54	benzene-1-methyl-3-nitro-	230
55	4-bromobenzoic acid nitrile	235
56	dimethylsulfone	238
57	ethanone-2-chloro-1-phenyl-	244
58	tabun	246
59	benzoic acid	249
60	indole	253
61	2, 2′-dimethylbiphenyl	258
62	naphthalene-1,5-dimethyl-	265
63	benzenamine-2,4,5-trichloro-	270
64	2, 2′-bipyridine	273
65	phenol-4-nitro-	279
66	hydroquinoline	285
67	benzenemethanol-a-phenyl-	297
68	*N*-(3-tolyl)acetic acid amid	303
69	hexanedioic acid-dibutyl ester	305
70	4-aminophenylacetic acid nitrile	312
71	ethanone-1,2-diphenyl-	320
72	tetraethylenepentamine	340
73	methanone-(4-bromophenyl)phenyl-	350
74	triphenylmethane	359
75	phenothiazine	371
76	triphenylchlorosilane	378
77	1,2-benzofluorene	407
78	benz[a]anthracene	438
79	benz[b]fluoranthene	481
80	perylene	495
81	dibenz[a-j]anthracene	531
82	dibenz[b-def]chrysene	596

The demand "for all i" to set up the order relation is called "The Generality Principle," and this condition transforms Partial Ordering into a vectorial approach. Each molecule is characterized with a vector whose elements are its attribute values [16]. When the inequality of equation (1) is true, then it is said

Table 2
Experimental values of NBP (°C) for the test set (82 compounds).

ID	Compound name	Exp.
1	ethylene-1,1-dichloro-	30
2	ethoxyethane	35
3	methyl propyl ether	39
4	cyclopentane	50
5	silane-ethenyltrimethyl-	55
6	silane-fluorothimethyl-	57
7	propanal-2-methyl-	63
8	aziridine-2-methyl-	66
9	cyclopropyl bromide	69
10	ethane-1-bromo-2-fluoro-	71
11	benzene methanamine-2-fluoro-	73
12	silane-ethoxytrimethyl-	75
13	pyridine-2-ethenyl-	79
14	diazinon	83
15	3,4-difluorophenol	85
16	2-pentanamine	91
17	propanenitrile	97
18	1,1,3,3-tetramethyldisilalazone	99
19	ethyl-3-(trifluoromethyl)benzoate	101
20	3-butyn-2-ol-2-methyl-	104
21	decanoic acid-methyl ester	108
22	propyldimethylchlorosilane	113
23	silane-chloro(chloromethyl)dimethyl-	115
24	pentanol-2-methyl-	119
25	silane-dichloro(chloromethyl)methyl-	121
26	ethane-1,1-dichloro-1-nitro-	124
27	diisopropylethylamine	127
28	silane-tetraethyl-	130
29	1,3-propanediamine-*N*, *N*-dimethyl-	133
30	2-propanone-1-bromo-	137
31	acetylacetone	140
32	1-butanol-3-mathyl-.acetate	142
33	phenol-3,4-dichloro-	145
34	1,4-bis(3-aminopropyl)piperazine	150
35	piperazine-2-methyl-	155
36	benzene-1-chloro-3-methyl-	160
37	acetamide-*N*, *N*-dimethyl-	165
38	limonene	170
39	benzene-1-ethenyl-4-methyl-	175
40	benzyl chloride	177
41	cyclopentane-pentyl-	181
42	octadecanoic acid	183
43	pentanoic acid	185
44	cyclohexane-1,2-dichloro-(trans)	193
45	phenol-2-chloro-4-methyl-	195
46	thiophenecarboxaldehyde	198

Table 2
Continued.

ID	Compound name	Exp.
47	octane-1-bromo-	201
48	benzene-1-fluoro-4-nitro-	205
49	benzoic acid-2-hydroxy	211
50	2-propenoic acid-2-ethylhexyl ester	215
51	phenol-4-ethyl-	218
52	*meta*-methoxybenzenethiol	223
53	4-methylbenzoicacidchloride	225
54	phenol-3,5-dichloro-	233
55	octanoic acid	237
56	mequinol	243
57	1,4-benzenedicarboxaldehyde	245
58	3-ethoxyaniline	248
59	2-propen-1-ol-3-phenyl-	250
60	phenol-2,6-bis(1-methylpropyl)-	255
61	benzoic acid-3-methyl	263
62	pyridine-3-phenyl-	269
63	benzene-1,3,5-tribromo-	271
64	1, 1′-biphenyl-2-chloro-	274
65	benzene-1-chloro-4-phenoxy-	284
66	hexanedinitrile	295
67	coumarin	298
68	acetamide-N-phenyl-	304
69	(Z)-stilbene	307
70	phenyl-4-pyridyl ketone	315
71	hexylresorcinol	333
72	phenanthridine	349
73	carbazole	355
74	pyrene	360
75	phosphoric acid-2-ethylhexyldiphenyl ester	375
76	2,3-benzofluorene	402
77	triphenylene	425
78	chrysene	448
79	benzo[e]pyrene	492
80	Picene	525
81	benzo[ghi]perylene	542
82	1,2,9,10-dibenzopyrene	595

that compound A is ranked higher than compound B (A dominates B), and that at least one descriptor for A is higher than the corresponding descriptor for B, and no descriptor for A is lower than the respective descriptor for B. If equation (1) is false, then both A and B are incomparable and can not be assigned a mutual order. Obviously, if all the descriptors for A are equal to the

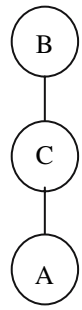

Figure 1. Representation of the two neighbors of C.

corresponding descriptors for B, the two compounds will have identical order (rank) and will be considered as "equivalent" rather than "identical," belonging to the same equivalence class. In consequence, the binary relation "$\leq$" becomes a quasi-order.

Condition (1) gives rise to a net of comparisons among the N compounds of the training set. Notice that the d descriptors participating in the POR model need to have the same designations, i.e. "high" and "low," and it may be necessary to multiply some of these descriptors by –1 in order to achieve identical designations. Furthermore, in POR's based QSPR-QSAR analyzes, the set of compounds under study has to follow the Similarity Principle, and two molecules do not belong to the same equivalence class if they exhibit different property values. Therefore, the absence of equivalent molecules leads to the conclusion that, whenever the property does not involve degenerated values for different structures, a given compound C can be positioned in the net of comparisons of the model if and only if there exist two neighbors A and B with a lower and higher rank than C, respectively. This situation is depicted in figure 1.

4. The algorithm

Our proposed ranking approach is also valid for a single descriptor model net [17]. Consider a training set **a** composed of N compounds. If we apply the condition (1) to this set then it will generate two different subsets $\mathbf{a}_1$ and $\mathbf{a}_2$: in $\mathbf{a}_1$ all the compounds will satisfy (1) and therefore this subset is called "ranking subset." The second subset $\mathbf{a}_2$ will contain incomparable compounds that do not follow the rule. Notice that, if $\mathbf{a}_2$ is an empty set, then there will exist a total order in **a**. It is still possible to order each compound Z belonging to $\mathbf{a}_2$ by searching among the $N-1$ compounds of set **a** (except Z) the corresponding two associated neighbors, allowing thus to generate new ranking subsets containing each one three compounds.

The first step of the algorithm consists on searching a subset of molecular descriptors $\mathbf{D}'$ (containing D' descriptors) from the pool **D** (with D total

available descriptors), in such a way that these variables are able to generate $N - 2$ ranking subsets for $N - 2$ compounds of the training set according to the inequality (1). It is not possible to generate ranking subsets for the two compounds that have the highest (p_{max}) and lowest (p_{min}) value of the observed property in this set. Obviously, each descriptor in $\mathbf{D}'$ need to have the highest and lowest numerical values for the compounds exhibiting p_{max} and p_{min}, respectively. Otherwise, there would not exist ranking subsets for all these $N - 2$ compounds. The number D' is dependent upon the property, the training set of compounds considered, and the molecular descriptor set under investigation.

A second aspect to have in mind is that the subset of descriptors **d** participating in the POR model must be able to identify and characterize each molecule from the training set independently. Therefore, those descriptors of $\mathbf{D}'$ having lower degeneracy would meet better this specific requirement when establishing the model net. Needless to say that different combinations of d descriptors may result equally suitable to describe satisfactorily the property intervals of the training set.

In order to apply the ranking subsets intervals obtained from the training set to predict the property intervals of the "unknown" compounds belonging to the test set, the algorithm considers the concept of average ranks. A compound X not pertaining to the training series is able to have its descriptor values lying in more than one ranking subset. Therefore, average upper and lower property limits for X have to be calculated along these subsets. The average limits are then translated to the nearest-lying experimental property values for two compounds of the training set.

5. Results and discussion

In the present study, D = 1201 and D′ results in 63 descriptors for the training set of compounds shown in Table 1. Since most of the molecular descriptors of **D** have high orientation, and our main intention is to show the performance of the algorithm proposed, the few descriptors with low orientation were not considered in the analysis. Also, it is not our purpose here to interpretate the models found in structural terms.

It is possible to try all possible combinations between the descriptors from $\mathbf{D}'$ to analyze the property intervals derived from the ranking subsets. However, the less degenerated the descriptors employed, the better the discrimination among the compounds in the model. The descriptor search for the model nets is performed in such a way that the variables are able to provide narrow prediction intervals for the compounds of the training set, and allow simultaneously to assign a position to the highest number of compounds belonging to the test set according with their descriptor values.

Table 3
Illustrative example with experimental intervals for ten ranking subsets obtained with the descriptor DCW^1.

$\mathbf{a}_1$		$\mathbf{a}_2$		$\mathbf{a}_3$		$\mathbf{a}_4$		$\mathbf{a}_5$		$\mathbf{a}_6$		$\mathbf{a}_7$		$\mathbf{a}_8$		$\mathbf{a}_9$		$\mathbf{a}_{10}$	
ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP
1	26	2	33	4	44	10	70	14	81	4	44	8	64	8	64	4	44	5	53
2	33	4	44	10	70	14	81	16	90	8	64	13	78	11	72	5	53	21	106
17	92	14	81	19	100	19	100	52	220	11	72	21	106	21	106	7	62	27	125
19	100																		
52	220																		
53	224																		
55	235																		
57	244																		
60	253																		
69	305																		
71	320																		
73	350																		
75	371																		
76	378																		
78	438																		
80	495																		
81	531																		
82	596																		

5.1. *Best single descriptor-model net*

Among the 63 available descriptors, the flexible variable DCW^1 is the least degenerated of all, and also satisfy the conditions mentioned previously. Table 3 provides an illustrative example with experimental intervals for ten ranking subsets obtained with the descriptor DCW^1, where the designation for each compound and its associated experimental NBP is given. The resulting 80 ranking subsets for the training set lead to the experimental NBP intervals for each compound indicated in Table 4. As mentioned in the introduction section, we are concerned here only in that the algorithm enables to position the compounds in a certain property interval from the net, leaving the prediction stage (b) to be discussed in a next publication. As can be observed from Table 4, all the experimental intervals assign the correct position for the training molecules once these are removed successively from the set (taking one at each time) and relocating them in the ranking subsets net. This is a sort of leave-one-out cross validation technique [18, 19], constituting a common practice in all regression analyzes.

Now, as a further step to test the predictive power of the one descriptor model, the ranking subsets are applied to predict the intervals for the 82 com-

Table 4
Experimental intervals for the training set derived from the ranking subsets of model nets including different number of descriptors.

ID	Exp.	One descriptor	Two descriptors	Three descriptors
1	26	$p_{\min}$	$p_{\min}$	$p_{\min}$
2	33	26–92	26–92	26–92
3	37	33–224	33–141	26–141
4	44	33–81	33–81	26–81
5	53	44–62	44–62	44–62
6	56	37–182	37–182	37–168
7	62	53–125	53–210	53–210
8	64	44–72	44–72	44–72
9	68	56–129	6–129	56–129
10	70	44–100	44–90	44–90
11	72	64–106	64–210	64–210
12	74	68–214	68–214	37–214
13	78	64–106	44–106	44–110
14	81	70–100	44–90	44–90
15	84	56–180	37–180	37–180
16	90	81–220	81–106	81–110
17	92	33–100	33–210	33–210
18	98	84–168	84–168	84–168
19	100	92–220	81–106	81–106
20	102	62–110	53–110	53–110
21	106	53–125	53–110	53–110
22	110	102–210	102–210	102–210
23	114	110–224	53–224	53–224
24	117	114–224	114–182	114–168
25	120	110–138	53–138	53–224
26	122	120–224	120–224	120–168
27	125	102–163	53–163	53–163
28	129	74–190	74–190	74–190
29	132	98–214	98–214	37–214
30	135	98–214	84–214	37–214
31	138	122–224	122–182	53–168
32	141	37–176	37–182	37–180
33	143	129–197	129–320	74–320
34	147	110–224	110–224	102–224
35	154	143–258	143–371	74–371
36	157	154–350	143–350	74–350
37	163	110–203	125–210	125–210
38	168	98–214	98–190	98–596
39	174	143–279	135–279	37–279
40	176	56–180	37–180	37–180
41	180	98–214	98–214	84–190
42	182	176–214	176–214	176–214
43	184	129–235	74–244	74–244
44	190	129–235	129–320	129–359
45	194	190–235	132–235	132–235

Table 4
Continued.

ID	Exp.	One descriptor	Two descriptors	Three descriptors
46	197	143–320	184–320	194–320
47	200	154–303	174–350	174–359
48	203	147–224	163–224	53–224
49	210	163–224	92–320	92–320
50	214	129–235	129–253	74–217
51	217	184–305	184–320	184–320
52	220	100–224	81–224	81–224
53	224	220–235	37–235	37–359
54	230	143–320	184–320	194–320
55	235	224–244	184–253	194–249
56	238	217–320	135–320	37–320
57	244	235–253	235–249	235–320
58	246	197–312	135–312	176–312
59	249	217–305	244–270	235–320
60	253	244–305	244–320	132–320
61	258	154–303	174–350	174–359
62	265	246–312	246–350	246–359
63	270	238–320	238–320	238–320
64	273	154–285	174–285	174–297
65	279	174–285	174–285	174–285
66	285	258–303	273–350	279–350
67	297	157–371	157–371	157–359
68	303	157–350	279–350	279–371
69	305	253–320	135–320	135–359
70	312	174–350	174–350	174–350
71	320	305–350	210–350	210–359
72	340	297–378	84–378	84–495
73	350	320–371	320–371	303–359
74	359	297–378	297–378	320–438
75	371	350–378	350–378	303–378
76	378	371–438	371–438	371–438
77	407	378–495	378–495	371–495
78	438	378–495	378–495	359–495
79	481	407–531	407–531	407–596
80	495	438–531	438–531	438–596
81	531	495–596	495–596	481–596
82	596	*p*max	*p*max	*p*max

pounds of the test set with "unknown" NBP. The results are shown in Table 5. Two of these compounds (1 and 82) are not able to be predicted in the present situation, as these molecules have descriptor values out of the training set intervals. It has to be noticed, however, that if some of the compounds from the training set were excluded and recalculated the ranking subsets after that, then

it would be also possible to predict the property intervals for these compounds. It can be appreciated that 29 compounds have their property intervals well predicted, and that many others have experimental NBP close to the predicted intervals.

5.2. *Two descriptors-model net*

A higher number of descriptors involved in the POR model would tend to characterize better the intervals for the training compounds, especially when average ranks are employed, tending to generate shorter property intervals. This causes compounds of the test set to have their descriptor values lying in less number of ranking subsets, and therefore avoiding a great uncertainty when locating them inside the net. On the contrary, if such characterization is excessive, it would result much more difficult to position the compounds of the test set in the model net. For the training set considered here, two descriptors from $\mathbf{D}'$ that differentiate the best among the compounds are DCW^1 again and Q_{index}, the topological descriptor "quadratic index" [20].

As can be seen from Table 4, all the training compounds are assigned a correct interval for two descriptors involved when practicing the leave-one-out procedure, and these tend to be narrower when compared with those encountered for a single descriptor model. The first ten ranking subsets for the two descriptors model are included in Table 6. When applying this model to predict the test set molecules, it is found that there are more compounds having descriptor values out of the intervals given by the net (12 molecules in Table 5). It is also noted that 32 compounds have their property intervals well predicted, and that many others have experimental values close to the predicted intervals. This model with DCW^1 and Q_{index} results of better quality, although it predicts a smaller number of test compounds.

5.3. *Three descriptors-model net*

The model containing three descriptors corresponds to the worst case of the three presented. It is composed of the descriptors DCW^1, $RDF050v$ [21], and Q_{index}, with $RDF050v$ being a Radial Distribution Function (5.0, weighted by atomic Van der Waals volumes). Table 3 reveals that the length of the intervals has increased for some compounds of the training series. From Table 5 it is also appreciated that 27 compounds can not be predicted in the test set.

6. Conclusions

We introduced a novel algorithm based on partial ordering ideas that is able to assign properly experimental endpoint intervals to 82 training compounds,

Table 5
Experimental intervals for the test set derived from the ranking subsets of different model nets.

ID	Exp.	One descriptor	Two descriptors	three descriptors
1	30	–	–	–
2	35	44–64	44–78	44–78
3	39	64–78	44–78	220–224
4	50	33–44	33–92	–
5	55	33–44	–	–
6	57	26–33	26–33	26–37
7	63	78–106	78–106	53–120
8	66	102–125	62–210	53–138
9	69	102–125	62–210	53–138
10	71	125–163	125–163	53–138
11	73	125–163	62–210	62–210
12	75	44–70	44–70	–
13	79	147–224	114–117	114–117
14	83	78–106	–	–
15	85	147–224	–	–
16	91	78–106	78–106	53–120
17	97	44–64	44–78	44–78
18	99	78–106	78–106	78–110
19	101	98–168	210–320	210–320
20	104	176–180	176–180	37–132
21	108	122–224	37–176	–
22	113	102–125	106–110	106–110
23	115	102–125	106–110	106–110
24	119	78–106	78–106	78–110
25	121	110–147	110–210	53–138
26	124	147–224	120–138	53–138
27	127	147–224	147–224	147–224
28	130	147–203	147–224	147–224
29	133	102–125	53–120	78–110
30	137	122–224	37–176	37–141
31	140	184–217	135–238	–
32	142	78–106	78–106	78–110
33	145	230–320	238–320	238–320
34	150	147–203	–	–
35	155	110–147	110–210	53–138
36	160	135–214	135–214	37–74
37	165	78–106	78–106	53–62
38	170	102–125	72–210	–
39	175	147–224	–	–
40	177	168–214	180–214	132–194
41	181	102–125	62–210	–
42	183	217–249	135–238	–
43	185	141–176	37–56	37–74
44	193	147–224	114–117	114–117
45	195	184–217	184–217	132–253

Table 5
Continued.

ID	Exp.	One descriptor	Two descriptors	three descriptors
46	198	98–168	180–214	37–74
47	201	176–180	37–84	–
48	205	176–180	182–214	138–168
49	211	303–350	303–350	–
50	215	176–180	84–180	176–180
51	218	230–320	238–320	238–320
52	223	217–238	184–230	74–143
53	225	197–246	–	–
54	233	230–320	238–320	238–320
55	237	168–214	84–135	–
56	243	238–270	238–320	238–270
57	245	190–194	74–184	–
58	248	246–265	246–265	246–265
59	250	230–320	238–320	238–320
60	255	297–359	–	–
61	263	230–320	238–320	238–320
62	269	238–320	238–320	–
63	271	154–273	279–285	–
64	274	154–273	174–258	–
65	284	157–297	157–297	–
66	295	78–106	78–106	220–224
67	298	184–217	190–320	74–143
68	304	154–273	174–273	174–273
69	307	154–273	174–273	174–273
70	315	246–265	246–265	246–265
71	333	157–297	–	–
72	349	359–378	359–378	371–378
73	355	157–297	154–371	303–371
74	360	359–378	–	359–438
75	375	297–359	297–359	–
76	402	359–378	–	–
77	425	407–495	407–495	407–495
78	448	407–495	407–495	407–495
79	492	407–495	407–495	407–495
80	525	531–596	531–596	495–596
81	542	531–596	531–596	481–596
82	595	–	–	–

and further analyzed the predictability of the POR nets established by using a test set with 82 “unknown” compounds. The models including different number of molecular descriptors for characterizing the net tend to be predictive on this test set.

Table 6
Illustrative example with experimental intervals for ten ranking subsets obtained with the descriptors DCW^1 and Q_{index}.

$\mathbf{a}_1$		$\mathbf{a}_2$		$\mathbf{a}_3$		$\mathbf{a}_4$		$\mathbf{a}_5$		$\mathbf{a}_6$		$\mathbf{a}_7$		$\mathbf{a}_8$		$\mathbf{a}_9$		$\mathbf{a}_{10}$	
ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP	ID	NBP
1	26	2	33	4	44	4	44	14	81	14	81	14	81	4	44	4	44	8	64
2	33	4	44	10	70	14	81	19	100	16	90	52	220	8	64	13	78	11	72
17	92	14	81	16	90	16	90	21	106	21	106	53	224	11	72	21	106	49	210
49	210																		
71	320																		
73	350																		
75	371																		
76	378																		
78	438																		
80	495																		
81	531																		
82	596																		

It is very important to have the opportunity to assess the predictive performance of the training model via an external test set of molecules. The POR methodology includes an analogue of the leave-one-out cross validation technique usually employed in regression analyzes since, once located the interval where a compound X is to be predicted, this is performed with its neighbor compounds without taking into account the experimental value of X. However, this kind of leave-one-out procedure does not guarantee that the descriptors involved in the POR model are able to rank compounds from a test set.

Acknowledgment

P.R.D. would like to thank to Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) for a postdoctoral research fellowship.

References

[1] C. Hansch and A. Leo, *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology* (American Chemical Society, Washington D. C., 1995).
[2] E.R. Malinowski, *Factor Analysis in Chemistry* (Wiley, New York, 1991).
[3] H. Hotelling, J. Educ. Psychol. 24 (1933) 417.
[4] S. Wold, M. Sjostrom and L. Eriksson, *Encyclopedia of Computational Chemistry* (Chichester, 1998).
[5] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, Massachusetts, 1989).
[6] R. Wehrens and L.M.C. Buydens, TrAC, Trends Anal. Chem. 17 (1998) 193.

[7] R. Leardi, J. Chemom. 8 (1994) 65.
[8] J. Zupan, *Encyclopedia of Computational Chemistry* (Chichester, 1998).
[9] HYPERCHEM (Hypercube), available from <http://www.hyper.com>.
[10] DRAGON 5.0 Evaluation Version, available from <http://www.disat.unimib.it/chm>.
[11] M.P. Gonzalez, A.A. Toropov, P.R. Duchowicz and E.A. Castro, Molecules 9 (2004) 1019.
[12] P.R. Duchowicz, E.A. Castro, F.M. Fernández and M.P. Gonzalez, Chem. Phys. Lett. 412 (2005) 376.
[13] M.A. Johnson and G.M. Maggiora, *Concepts and Applications of Molecular Similarity* (Wiley, New York, 1990).
[14] D.M. Hawkins, S.C. Basak and D. Mills, J. Chem. Inf. Model. 43 (2003) 579.
[15] G. Birkhoff, *Lattice Theory* (American Mathematical Society, Rhode Island, 1948).
[16] E. Halfon and M.G. Reggiani, Environ. Sci. Technol. 20 (1986) 1173.
[17] E.A. Castro, F.M. Fernández and P.R. Duchowicz, J. Math. Chem. 37 (2005) 433.
[18] P.A. Lachenbruch and M. Mickey, Technometrics 10 (1968) 1.
[19] S. Geisser, J. Am. Stat. Assoc. 70 (1975) 320.
[20] A.T. Balaban, Theor. Chim. Acta 53 (1979) 355.
[21] M.C. Hemmer, V. Steinhauer and J. Gasteiger, Vibrat. Spect. 19 (1999) 151.