A COMPARATIVE QSAR STUDY OF SVM AND PPR IN THE CORRELATION OF LITHIUM CATION BASICITIES

Alan R. Katritzky[a1,*], Yueying Ren[a2], Svetoslav H. Slavov[a3] and Mati Karelson[b,c]

[a] *Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida 32611, USA; e-mail: [1] katritzky@chem.ufl.edu, [2] renry02@st.lzu.edu.cn, [3] svetoslavslavov2000@yahoo.co.uk*

[b] *Department of Chemistry, University of Tartu, 2 Jakobi Street, Tartu 51014, Estonia; e-mail: mati.karelson@ttu.ee*

[c] *Department of Chemistry, Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia; e-mail: mati.karelson@ttu.ee*

Received September 26, 2008
Accepted January 9, 2009
Published online February 11, 2009

Dedicated to the memory of Professor Otto Exner.

Correlation of gas-phase lithium cation basicities (LCB) of 259 diverse compounds extends the published datasets utilizing multilinear, support vector machine (SVM) and projection pursuit regression (PPR) modeling. The best multiple linear regression (BMLR) method implemented in CODESSA was used to: (i) build multiparameter linear QSPR models and (ii) select set of descriptors for further treatment by the SVM and PPR. The external predictivity and the performance of each of the above methods was estimated and compared to those of the other techniques. The PPR method produced results superior to SVM, which in turn outperformed MLR. The physico-chemical interpretation of each of the descriptors provides new insight into the mechanism of LCB interactions.

Keywords: Quantitative structure-property relationships; Lithium cation basicity; CODESSA; Multiple linear regression; Support vector machines; Projection pursuit regression.

The present comprehensive study of the relationship of lithium cation basicity values to chemical structure amplifies previous work[1,2] by: (i) extending the dataset; providing (ii) new models with superior predictive power and (iii) new insight into the mechanism of LCB interactions.

During the last 60 years, proton affinity scales ranging from weak to very strong bases[3] have been constructed by computational chemistry, including high-level *ab initio* calculations[4]. Recent studies have concentrated on inter-

doi:10.1135/cccc2008191

actions with metal cations instead of protons[5], especially alkali-metal cations which are relatively easily produced under vacuum[6–11].

$$B_{(g)} + Li^+_{(g)} \rightarrow B\text{–}Li^+_{(g)} \qquad (1)$$

Similarly to H^+, the gas-phase lithium cation basicity (LCB) is defined as the negative Gibbs free energy associated with reaction (*1*). However, the coordination properties of Li^+ are quite different from those of H^+: a proton adds to a base, forming a polar covalent σ bond with a very extensive charge transfer, while the bonds formed by Li^+ are largely due to ion-dipole (electrostatic) interactions[12]. As a result, the basicities toward lithium cation are much smaller and cover a narrower range in the energy scale than gas phase basicities towards the proton.

In addition to the now classical methods for experimental determination of LCBs such as mass spectrometry[13] and ion cyclotron resonance (ICR)[14], more recent methods employing Fourier transform ion cyclotron resonance (FT-ICR)[15–17], equilibrium constant determination high pressure mass spectrometry (HPMS)[11,18], energy-resolved collision-induced dissociation (CID)[19] and photodissociation and radiative association kinetics[20,21] have become widely applicable.

Most current LCB estimations rely on *ab initio*[22,23] or density functional theory (DFT)[12,24–27] calculations. The *ab initio* Hartree–Fock calculations are usually performed by using 6-311+G(d,p) or higher basis sets and scaling factors less than 1. Compared to the Hartree–Fock, MP2, MP3 and configuration interaction (CI) methods, the DFT method at B3LYP level[28] in most cases provide results with higher accuracy while no scaling factors are necessary. Then, the Gibbs free energies and enthalpies at 298 K are calculated and the $\Delta\Delta G_{\mathrm{Li+}}$ values obtained are converted into absolute LCBs.

Applications of the *ab initio* and DFT methods for LCB correlation usually result in excellent correlations with $R^2 > 0.97$ [29], but these are somewhat limited due to the significant computational power required. Thus, the QSPR methods known for their ability to generate accurate estimations in relatively short time have gained increased popularity.

The first two QSPR models for the prediction of LCB were proposed almost simultaneously: (i) our group[1] proposed a six-descriptor linear model for 205 compounds with $R^2 = 0.801$ and $s = 2.963$; (ii) Jover et al.[2] proposed QSPR models based on a) MLR-CNN (multiple linear regression – cellular neural networks) and b) GA-CNN (genetic algorithm – cellular neural networks) nonlinear techniques for an extended dataset of 229 compounds with statistical parameters for the HMCNN (hybrid multiple component neural networks) training set R^2 = 0.905 and RMSE = 2.251, while the alternative (GA-CNN) produced superior results: R^2 = 0.954 and RMSE = 1.563.

Despite this work, the lithium cation basicity scales remain much more limited than the proton affinity scale. Thus, the extension of the LCB scale could greatly benefit from the existing theoretical methods for property predictions and estimations.

DATASET

Experimental lithium cation basicity (LCB) values for the 259 molecules (shown in Table I) were taken from the literature[12,16,21,24–27,30]. With only two exceptions (methoxyethanol and 1,2-dimethoxyethane) the dataset of ref.[2] appears to be an extension (including 26 new compounds) of the dataset of ref.[1] The names of the compounds, the literature source used and their experimental and predicted LCBs are listed in Table I. The LCB ranged from 17.9 to 54.7 kcal/mol, with a mean value of 35.33 kcal/mol; LCB data distribution is close to normal (Gaussian), see Fig. 1

COMPUTATIONAL PROCEDURE

The structures of the compounds were preoptimized employing the molecular mechanics force field (MM+) available in the HyperChem 7.5 [31]. Final refined molecular geometries were obtained using the semi-empirical meth-

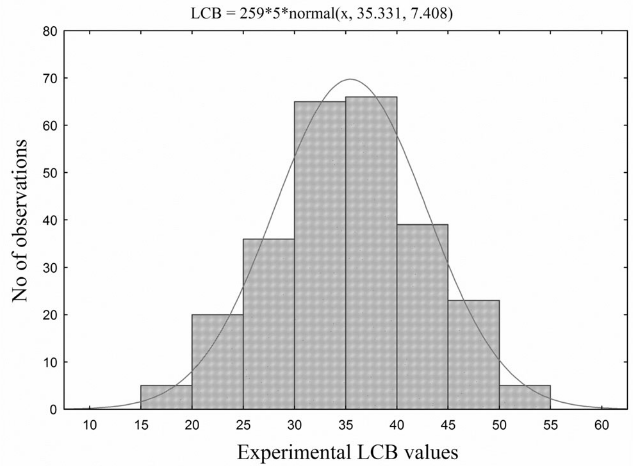

FIG. 1
Histogram and probability density function of the LCB values

TABLE I
Compounds together with their experimental and predicted LCBs

No.	Compounds	Ref.	Subset	Exp.	Predicted		
					BMLR[i]	SVM[j]	PPR[j]
1	trifluoromethylacetylene	a	A	17.9	20.1	19.3	21.8
2	sulfur dioxide	a	B	18.2	21.2	20.9	18.9
3	carbonyl fluoride	a	C	18.4	15.6	21.3	16.7
4	hexafluoroacetone	a	A	19.1	19.6	17.8	18.9
5	bis(trifluoromethyl) disulfide	a	B	19.2	11.4	19.9	20.1
6	perfluoro-*tert*-butyl alcohol	a	C	20.3	24.1	21.4	20.7
7	methanethiol	a	A	20.3	26.1	20.7	21.3
8	trifluoroacetonitrile	a	B	21.3	19.1	19.8	19.1
9	ethanethiol	a	C	21.4	27.0	21.5	21.8
10	trifluoroacetaldehyde	a	A	21.8	22.4	21.1	22.6
11	bis(perfluoroisopropyl)ketone	a	B	21.9	24.0	20.4	21.4
12	perfluoropyridine	a	C	22.3	25.0	22.8	23.3
13	propane-2-thiol	a	A	22.4	27.6	22.6	22.5
14	propane-1-thiol	a	B	22.5	28.7	23.9	23.8
15	dimethyl sulfide	a	C	23.4	27.6	24.3	24.5
16	2-methylpropane-1-thiol	a	A	23.7	29.7	25.5	25.6
17	chlorobenzene	d	B	23.7	28.9	28.7	27.1
18	1,1,1,3,3,3-hexafluoropropan-2-ol	a	C	23.8	26.3	23.8	24.9
19	perfluoro-*tert*-butylamine	a	A	23.8	26.5	24.2	23.3
20	2-methylpropane-2-thiol	a	B	23.8	28.4	23.8	23.4
21	1-butanethiol	a	C	24.0	29.7	25.6	25.7
22	bromobenzene	d	A	24.3	29.6	29.7	28.5
23	bis(difluoromethyl) ketone	a	B	24.6	26.3	24.6	24.9
24	water	a	C	24.7	28.9	26.3	27.4
25	ethyl methyl sulfide	a	A	25.0	29.6	26.8	26.5
26	formaldehyde	a	B	25.4	29.9	28.2	27.5
27	2,2,2-trifluoroethyl trifluoroacetate	a	C	25.7	27.3	26.9	28.0
28	tetrahydrothiophene	a	A	25.8	28.6	25.5	24.8
29	hydrogen cyanide	a	B	25.9	27.1	26.1	25.9
30	tetrahydrothiopyran	a	C	25.9	29.4	26.8	26.0
31	fluoroacetonitrile	a	A	26.2	29.1	28.5	28.9
32	1,1,1,3,3,3-hexafluoro-2-methoxypropane	a	B	26.2	26.4	26.7	27.4
33	malononitrile	a	C	26.3	33.2	31.7	32.3
34	diethyl sulfide	a	A	26.4	29.0	26.2	25.4
35	2,2,2-trifluoroethanol	a	B	26.5	28.1	27.9	28.0
36	benzenethiol	d	C	26.8	27.9	22.5	23.0
37	trichloroacetonitrile	a	A	26.8	27.0	25.7	25.6
38	benzene	a	B	26.9	27.1	26.4	27.4
39	1,1,1-trifluoroacetone	a	C	27.0	27.3	27.3	27.3
40	trichloroacetaldehyde	a	A	27.2	30.1	27.7	28.6

TABLE I
(*Continued*)

No.	Compounds	Ref.	Subset	Exp.	Predicted		
					BMLR[i]	SVM[j]	PPR[j]
41	1,1,1,5,5,5-hexafluoroacetylacetone	a	B	27.3	31.6	29.7	31.9
42	isobutyl methyl sulfide	a	C	27.4	29.8	30.9	30.7
43	(2,2,2-trifluoroethoxy)ethene	a	A	27.4	31.9	30.1	30.6
44	dichloroacetonitrile	a	B	27.7	30.4	28.3	28.7
45	phenol	d	C	28.1	32.4	30.5	30.8
46	methanol	a	A	28.5	30.5	28.3	28.5
47	pyrazine	a	B	28.6	33.9	33.2	33.8
48	methyl chloroformate	a	C	28.9	28.9	30.7	28.6
49	dipropyl sulfide	a	A	28.9	29.5	27.2	27.1
50	methyl trifluoroacetate	a	B	28.9	31.3	29.6	28.9
51	diisopropyl sulfide	a	C	28.9	27.4	28.6	28.7
52	bis(2,2,2-trifluoroethyl) ether	a	A	29.2	26.3	27.6	28.0
53	1,2-dihydrocyclobutabenzene	d	B	29.3	32.3	33.1	31.9
54	cyanogen bromide	a	C	29.4	24.5	29.5	25.8
55	chloroacetonitrile	a	A	29.4	31.3	30.2	31.8
56	4-(trifluoromethyl)pyridine	a	B	29.5	30.6	30.3	29.1
57	dimethyl ether	a	C	29.5	29.4	30.0	28.9
58	methyl 2,2,2-trifluoroethyl ether	a	A	29.6	27.1	28.7	28.0
59	toluene	d	B	29.7	32.6	33.1	32.3
60	pyrimidine	a	C	29.8	36.1	35.7	35.2
61	*S*-methyl trifluoroacetothioate	a	A	29.9	27.5	30.0	27.2
62	anisole	a	B	30.2	34.3	34.7	32.4
63	ammonia	a	C	30.2	30.0	28.6	30.4
64	1,4-dioxane	a	A	30.3	31.3	32.4	32.3
65	2,2,2-trichloroethanol	a	B	30.4	34.2	33.2	32.8
66	ethanol	a	C	30.4	32.1	30.3	30.5
67	naphthalene	a	A	30.5	30.1	30.5	30.2
68	dibutyl sulfide	a	B	30.6	32.7	31.9	31.9
69	di-*tert*-butyl sulfide	a	C	30.6	30.0	31.6	30.8
70	ethyl trifluoroacetate	a	A	30.6	30.8	29.8	29.2
71	ethylbenzene	a	B	31.1	35.0	35.3	33.2
72	methylamine	a	C	31.3	32.8	32.6	34.2
73	1-propanol	a	A	31.4	33.7	32.4	32.9
74	3-chloropyridine	a	B	31.6	34.5	34.1	34.9
75	acetaldehyde	a	C	31.8	32.5	31.5	32.1
76	trimethylamine	a	A	32.0	32.6	32.9	32.3
77	1,2,3-triazole	a	B	32.1	37.5	33.3	34.7
78	dimethylamine	a	C	32.1	32.5	32.5	33.5
79	2-propanol	a	A	32.3	33.0	31.7	31.9
80	methyl formate	a	B	32.4	31.5	32.7	31.0

TABLE I
(*Continued*)

No.	Compounds	Ref.	Subset	Exp.	Predicted		
					BMLR[i]	SVM[j]	PPR[j]
81	isobutyl alcohol	a	C	32.5	34.6	33.9	33.6
82	butylbenzene	a	A	32.6	38.0	37.2	36.5
83	acetic acid	a	B	32.7	34.0	32.7	33.7
84	tetrahydrofuran	a	C	32.7	31.5	32.8	32.1
85	1,2,4-triazole	a	A	32.7	37.9	34.3	36.0
86	methoxyacetonitrile	a	B	32.8	34.7	35.5	34.9
87	butan-1-ol	a	C	32.8	34.7	33.9	33.6
88	propanal	a	A	32.8	34.0	33.7	34.1
89	isoxazole	a	B	32.9	30.8	31.4	30.2
90	2-hydroxyethyl hydrogen sulfate	a	C	33.0	34.9	34.3	35.2
91	2,2-dimethyl-1-propanol	a	A	33.1	35.4	35.0	34.7
92	2,6-difluoropyridine	a	B	33.2	31.0	31.7	31.9
93	*tert*-butyl alcohol	a	C	33.3	33.7	32.8	32.9
94	diethyl ether	a	A	33.3	31.8	33.1	33.3
95	butanal	a	B	33.3	35.5	35.7	35.1
96	*sec*-butyl alcohol	a	C	33.3	34.6	33.7	33.5
97	tetrazole	a	A	33.3	38.9	32.5	35.3
98	thiazole	a	B	33.4	32.9	32.7	33.0
99	pyrazole	a	C	33.6	35.7	33.4	34.1
100	phenanthrene	a	A	33.7	32.0	32.8	33.5
101	pentanal	a	B	33.8	36.5	37.0	36.3
102	*S*-methyl thioacetate	a	C	33.8	32.1	31.8	31.7
103	anthracene	a	A	33.8	31.9	32.7	33.1
104	ethyl formate	a	B	33.9	32.8	33.7	33.3
105	trifluoroacetamide	a	C	33.9	31.3	33.2	33.1
106	acetonitrile	a	A	34.0	36.4	34.8	36.5
107	dimethyl sulfate	a	B	34.0	29.8	28.7	29.8
108	pyrene	a	C	34.2	32.7	33.4	34.3
109	*tert*-butyl methyl ether	a	A	34.2	32.9	34.4	34.4
110	1-methylpyrazole	a	B	34.3	39.0	39.7	37.7
111	butyl formate	a	C	34.3	36.6	36.8	36.6
112	propyl formate	a	A	34.3	35.1	35.2	34.9
113	(methylthio)acetonitrile	a	B	34.3	34.2	34.6	34.4
114	2-methyltetrahydrofuran	a	C	34.3	32.8	32.7	32.0
115	cyclohexanemethanol	a	A	34.3	33.3	34.8	34.4
116	hexanal	a	B	34.4	37.2	38.1	37.1
117	ethyl perfluoropivalate	a	C	34.5	41.0	39.9	39.5
118	*N*,*N*-dimethylcyanoformamide	a	A	34.5	32.5	33.0	35.1
119	heptanal	a	B	34.6	37.9	39.0	37.9
120	methoxytrimethylsilane	a	C	34.6	33.0	36.2	35.9

TABLE I
(*Continued*)

No.	Compounds	Ref.	Subset	Exp.	Predicted		
					BMLR[i]	SVM[j]	PPR[j]
121	coronene	e	A	34.7	39.2	36.8	38.0
122	azulene	a	B	34.7	30.8	31.6	30.7
123	phosphoryl chloride	a	C	34.7	28.1	31.7	31.9
124	dipropyl ether	a	A	34.8	34.0	35.5	35.8
125	2,5-dimethyltetrahydrofuran	a	B	35.0	33.7	35.1	35.1
126	pyridine	a	C	35.0	32.5	32.2	33.1
127	benzeneacetonitrile	a	A	35.1	36.2	36.4	36.0
128	3-methylpyrazole	a	B	35.1	38.6	36.9	37.3
129	2-fluoropyridine	a	C	35.1	31.2	30.5	30.7
130	hexafluoro-diacetamide	a	A	35.2	34.6	34.3	34.1
131	methyl acetate	a	B	35.2	32.4	34.4	33.6
132	1,1,1-trifluoro-2,4-pentanedione	a	C	35.3	35.1	37.3	36.2
133	acetone	a	A	35.3	32.9	33.3	33.5
134	propionitrile	a	B	35.3	32.3	32.2	33.5
135	*tert*-butyl ethyl ether	a	C	35.4	33.5	35.0	35.5
136	butanenitrile	a	A	35.4	33.9	34.3	34.4
137	benzonitrile	a	B	35.5	32.4	31.8	33.2
138	diisopropyl ether	a	C	35.5	33.0	34.5	34.9
139	2-hydroxyethyl hydrogen sulfite	a	A	35.6	32.9	34.9	34.5
140	isobutyronitrile	a	B	35.7	33.2	33.5	34.1
141	4-methylpyrazole	a	C	35.7	37.9	35.7	36.9
142	benzyl alcohol	a	A	35.8	37.0	35.8	37.6
143	valeronitrile	a	B	35.8	34.8	35.5	35.0
144	benzo[3,4]cyclobuta[3]phenylene	e	C	35.9	33.2	33.7	34.5
145	heptylbenzene	a	A	35.9	40.4	38.5	38.2
146	methyl 4-nitrophenyl sulfone	a	B	36.0	44.1	38.4	41.5
147	butan-2-one	a	C	36.0	34.8	35.7	35.2
148	ethyl acetate	a	A	36.0	33.8	35.5	34.7
149	methyl methanesulfonate	a	B	36.3	37.0	37.4	38.1
150	methyl propanoate	a	C	36.3	33.8	35.5	34.9
151	dimethyl terephthalate	b	A	36.3	41.8	38.8	41.6
152	*O*-methyl methanesulfonothioate	a	B	36.4	32.0	34.2	35.0
153	pivalonitrile	a	C	36.4	33.9	34.4	34.3
154	dibutyl ether	a	A	36.5	36.7	36.2	36.1
155	3-methylpyridine	a	B	36.5	35.4	36.7	37.1
156	1,2-oxathiolane-2,2-dioxide	a	C	36.7	37.2	37.4	37.7
157	pentan-3-one	a	A	36.7	35.1	36.2	35.9
158	octanonitrile	a	B	36.8	36.7	38.1	36.5
159	methyl benzoate	a	C	37.0	37.3	36.8	37.5
160	1,4-dimethylpyrazole	a	A	37.0	39.7	40.1	38.3

TABLE I
(*Continued*)

No.	Compounds	Ref.	Subset	Exp.	Predicted		
					BMLR[i]	SVM[j]	PPR[j]
161	dimethyl carbonate	a	B	37.0	31.8	33.7	32.7
162	1,1-diphenylethane	f	C	37.1	37.7	37.2	36.4
163	dimethyl sulfone	a	A	37.1	35.2	38.4	38.3
164	4,4,4-trifluorobutylamine	a	B	37.1	34.2	36.2	35.1
165	1-cyclopropylethan-1-one	a	C	37.4	35.4	36.5	36.6
166	formamide	a	A	37.5	36.4	37.9	37.3
167	2,4-dimethylpentan-3-one	a	B	37.5	37.2	38.8	37.2
168	dimethyl isophthalate	b	C	37.6	41.9	39.0	41.9
169	methyl phenyl sulfone	a	A	37.6	37.5	37.7	38.8
170	nonanonitrile	a	B	37.6	40.5	41.8	38.4
171	1,5-dimethylpyrazole	a	C	37.7	40.6	39.7	40.8
172	benzaldehyde	a	A	37.7	34.3	34.4	34.9
173	adamantane-1-carbonitrile	a	B	38.1	37.3	38.5	38.2
174	imidazole	a	C	38.1	36.1	35.2	35.8
175	1-(4-methylphenyl)ethan-1-one	a	A	38.1	40.7	41.8	43.2
176	1,3,5-trimethylpyrazole	a	B	38.3	41.1	43.2	39.3
177	dicyclopropylmethanone	a	C	38.4	35.9	37.1	37.1
178	3,4,5-trimethylpyrazole	a	A	38.7	38.3	37.2	38.0
179	ethyl pivalate	a	B	38.9	36.0	37.0	37.1
180	dimethylcyanamide	a	C	39.0	34.3	33.0	34.4
181	sulfolane	a	A	39.0	36.9	39.1	39.5
182	1,3,4,5-tetramethylpyrazole	a	B	39.0	40.0	41.5	38.8
183	circumcoronene	e	C	39.2	35.0	38.2	38.4
184	1-[4-(trifluoromethyl)phenyl]ethan-1-one	h	A	39.2	44.5	40.1	41.7
185	1-[3-(trifluoromethyl)phenyl]ethan-1-one	h	B	39.2	37.9	38.5	38.9
186	1,2-oxathiolane 2-oxide	a	C	39.2	38.0	38.7	38.9
187	methyl phenyl sulfone	a	A	39.3	40.6	39.7	40.8
188	angular benzo[3,4]cyclobuta[3]phenylene	e	B	39.4	33.2	33.8	34.9
189	*N*-methylformamide	a	C	39.6	39.2	40.3	40.8
190	*N*,*N*-dimethyltrifluoroacetamide	a	A	39.7	34.7	37.0	36.2
191	acetamide	a	B	39.9	39.2	40.5	40.3
192	*N*-methyl dimethylcarbamate	a	C	39.9	37.3	37.8	39.2
193	1-methyl-4-(methylsulfonyl)-benzene	a	A	40.2	43.3	39.6	42.1
194	1-methylimidazole	a	B	40.2	38.2	38.2	37.6
195	*N*,*N*-dimethyl-3-pyridinamine	a	C	40.6	43.7	47.0	44.3
196	diphenyl sulfone	a	A	40.6	39.8	40.8	42.1
197	1-(3-fluorophenyl)ethan-1-one	h	B	41.0	39.2	41.4	41.9
198	1-(3-chlorophenyl)ethan-1-one	h	C	41.1	40.5	41.8	42.9
199	pyridazine	a	A	41.4	33.7	32.5	33.5
200	*N*-methylacetamide	a	B	41.5	39.4	41.0	40.7

TABLE I
(*Continued*)

No.	Compounds	Ref.	Subset	Exp.	Predicted		
					BMLR[i]	SVM[j]	PPR[j]
201	isophorone	a	C	41.5	41.3	42.9	44.0
202	*N*,*N*-dimethylformamide	a	A	41.5	38.3	39.2	39.1
203	glycine	c	B	41.6	39.6	39.9	39.8
204	dimethyl sulfoxide	a	C	41.8	32.2	36.5	35.6
205	1,2-dimethylimidazole	a	A	41.8	39.8	40.6	38.8
206	1,2-diphenylethane	f	B	42.0	36.4	36.1	36.6
207	*N*,*N*-dimethyl-4-pyridinamine	a	C	42.0	43.8	47.1	44.1
208	tetramethylguanidine	a	A	42.4	42.5	43.3	42.6
209	1-(4-chlorophenyl)ethan-1-one	h	B	42.5	42.6	43.7	44.1
210	1-(4-fluorophenyl)ethan-1-one	h	C	42.5	40.5	41.7	42.8
211	dimethyl phosphate	a	A	42.5	39.2	41.5	41.9
212	2,4,5-trimethylpyrazole	a	B	42.6	40.1	41.0	39.3
213	2-methoxyethanol	a	C	42.7	35.5	36.1	37.2
214	*N*,*N*-dimethylacetamide	a	A	42.8	38.7	40.3	39.8
215	methyl phenyl sulfoxide	a	B	42.9	37.9	39.9	39.8
216	tetrahydrothiofene 1-oxide	a	C	43.1	34.0	37.9	37.3
217	acetylacetone	a	A	43.1	38.5	40.2	39.0
218	alanine	c	B	43.2	41.2	41.8	42.2
219	1,8-naphthyridine	a	C	43.4	38.8	39.4	38.3
220	trimethyl phosphate	a	A	43.7	42.4	44.0	44.3
221	diphenyl sulfoxide	a	B	43.9	37.0	40.7	40.5
222	dimethyl methylphosphonate	a	C	44.0	43.4	39.6	40.9
223	1,7-diphenylheptane	f	A	44.0	43.1	44.1	44.6
224	diisopropyl phosphonate	a	B	44.1	46.3	45.2	45.0
225	diethyl (chloromethyl)phosphonate	a	C	44.1	46.5	45.4	45.2
226	1,3-diphenylpropane	f	A	44.2	38.6	37.2	38.2
227	acetophenone	h	B	44.2	38.5	39.5	39.7
228	4-(trifluoromethyl)phenyl diphenylphosphinate	a	C	44.3	52.7	43.5	44.7
229	1-(3-methylphenyl)ethan-1-one	h	A	44.9	40.8	41.9	43.7
230	diethyl methylphosphonate	a	B	45.0	45.4	44.9	44.7
231	valine	c	C	45.0	43.6	44.8	44.8
232	dimethyl phenylphosphinite	a	A	45.1	49.2	44.4	44.9
233	triethyl phosphate	a	B	45.1	45.4	44.9	44.9
234	triphenyl phosphate	a	C	45.2	48.2	44.6	44.7
235	leucine	c	A	45.2	45.0	46.3	46.0
236	cysteine	c	B	45.2	40.5	46.5	44.6
237	isoleucine	c	C	45.3	45.0	46.3	46.2
238	4-fluorophenyl diphenylphosphinate	a	A	45.6	51.1	45.9	46.8
239	trimethylphosphine oxide	a	B	45.7	40.7	43.4	45.1
240	3-chloro-4-methoxyacetophenone	h	C	45.8	44.4	46.1	45.6

TABLE I
(*Continued*)

No.	Compounds	Ref.	Subset	Exp.	Predicted		
					BMLR[i]	SVM[j]	PPR[j]
241	4-methylphenylethan-1-one	h	A	46.3	40.8	42.0	43.8
242	triethylphosphine oxide	a	B	46.7	44.2	46.1	46.5
243	phenyl diphenylphosphinate	a	C	46.9	49.1	45.8	46.3
244	dimethyl phthalate	b	A	47.1	41.1	37.9	40.7
245	2-(pyridin-2-yl)ethanamine	g	B	47.4	45.2	48.5	47.0
246	hexamethylphosphoramide	a	C	47.5	52.8	49.3	48.4
247	triphenylphosphine oxide	a	A	47.5	48.2	46.7	47.1
248	1-[4-(methylthio)phenyl]ethan-1-one	h	B	47.5	41.7	41.9	42.2
249	proline	c	C	47.5	40.7	43.1	45.4
250	1-(4-methoxyphenyl)ethanone	h	A	48.3	43.0	45.0	45.5
251	phenylalanine	c	B	48.4	47.6	48.0	47.5
252	serine	c	C	48.6	45.9	48.0	47.7
253	tyrosine	c	A	49.0	52.2	50.6	52.2
254	threonine	c	B	49.9	47.2	48.9	48.8
255	methionine	c	C	50.4	42.7	48.2	47.0
256	aspartic acid	c	A	51.5	46.5	48.7	47.7
257	tryptophan	c	B	52.3	54.6	50.2	53.3
258	glutamic acid	c	C	52.9	48.2	50.1	49.4
259	1-[4-(dimethylamino)phenyl]ethan-1-one	h	A	54.7	46.5	50.3	48.1

[a] Ref.[24]; [b] ref.[17]; [c] ref.[21]; [d] ref.[26]; [e] ref.[27]; [f] ref.[25]; [g] ref.[12]; [h] ref.[30]; [i] according to the general model; [j] averaged values from the submodels.

od AM1 (Austin Model 1) with no symmetry constraints imposed. All calculations were carried out applying a gradient norm limit of 0.01 kcal/mol as a stopping criterion. The optimized geometries were then loaded into CODESSA package[32]. Overall, 883 descriptors classified as (i) constitutional, (ii) topological, (iii) geometrical, (iv) electrostatic and (v) quantum chemical were calculated. These descriptors encode information about the connections between atoms, shape, branching, symmetry, distribution of charge, and quantum-chemical properties of the molecule.

Best multilinear regression. The best multilinear regression (BMLR) method[33] implemented in the CODESSA package was used for systematic development of multi-linear QSPR equations: preselection of descriptors by eliminating those which are not available for every structure, have a small variation or having *F*-test or *t*-values less than the predetermined ones. BMLR implements the strategy described in the supplementary material.

Support vector machine for regression (SVR). SVM algorithm was proposed in 1995 by Vapnik[34]. The SVM methods are designed around the computation of an optimal separating hyperplane which provides minimum expected generalisation error in a multidimensional space called “feature space”. In this m-dimensional space each compound is represented by a point which may be thought of as vector of m numbers (descriptors). The support vector machine can actually locate the hyperplane without ever representing the feature space explicitly, simply by defining a function, called a kernel function.

The main advantages of SVM are: (i) stable, reproducible results independent of the optimization algorithm; (ii) the optimum solution (global minima) is guaranteed; (iii) only a few parameters have to be adjusted: the regularization parameter (C), the nature and the parameters of the kernel function.

Basic theory of projection pursuit regression (PPR). For many practical problems, the data is usually high dimensional. Thus, we should project the original high-dimensional data into a lower-dimensional space, line or a plane, etc., to try to find the intrinsic structure for visual inspection. Given a dataset (X_1, ..., X_n), $X \in IR^k$ results in a k-dimensional matrix ($k \times n$), where k is the number of observed variables and n is the number of units. The matrix Z with a dimension ($m \times n$) is constructed by multiplying the m-dimensional orthonormal matrix A ($m \times k$) to X ($k \times n$) and represents the coordinates of the projection data onto the m-dimensional ($m < k$) space spanned by rows of A. As there are an infinite number of projections from a higher dimension to a lower dimension, it is important to have a technique to pursue a finite sequence of projections that can reveal the most interesting structures of the data. Projection pursuit (PP) is powerful tool that combines ideas of both projection and pursuit[35].

The nonlinear nature of the SVM and PPR methods provides higher flexibility than BMLR for describing complex phenomena difficult to treat by the standard linear Hansch approach. Both SVM and PPR algorithms were written using the R-programming language as implemented in the R statistical package[36].

QSAR PROCEDURE

In this work a modified QSPR approach, aiming to combine the advantages of the two modeling procedures most frequently used was applied, i.e., (i) using all available data points to build the model and to apply as a sole validation the standard internal crossvalidation procedure or (ii) to use only a part of the available data for building the model, keeping the remaining data points for external validation. Our recommended procedure to build a reliable QSPR model is as follows:

1. All data points of the full dataset were ordered in a descending order of their LCB values.

2. The initial set was separated into three subsets (conditionally denoted as A, B and C) by selection of every third point from the original dataset in order to obtain a similar distribution of the investigated property values for each subset, A, B, C.

3. Three new datasets were constructed using the three binary sums combinations: A+B, A+C and B+C.

4. The standard QSAR modeling procedure including best multiple linear regression method (BMLR) was applied to those three datasets obtained in step 3.

5. The complementary parts to each of these three datasets (C, B and A, respectively) were used as external validation datasets by considering their consistency.

6. All the descriptors that appeared in the obtained models of step 4 were tested to obtain a general model including the full dataset of compounds.

7. The general model was again validated using classical internal crossvalidation and scrambling procedures.

For evaluation of the model performance we utilized: (i) R^2, to measure the model's fit performance and (ii) RMSE, as defined in Eq. (*2*), to evaluate the prediction performance:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_{ke} - y_{kp})^2}{n}} \tag{2}$$

where k represents the k-th molecule, y_{ke} is the experimental property, y_{kp} is the predicted property, and n is the number of compounds in the analyzed set.

Using the models developed, the predicted LCB values of compounds in the training set were compared with the observed values. Compounds with deviations larger than 3 times that of the standard deviation were judged to be statistical outliers and removed from the training set. The model fitting process was then repeated using the remaining data. In addition, due to the nonlinear and/or nonparametric nature of the modeling methods used in this study, mathematical expressions of the resulting nonlinear models are not available.

RESULTS AND DISCUSSION

An estimation of the predictive power of the previous models. We estimated the predictive power of the models previously reported[1,2] using: (i) the six parameter multilinear model published in Table 2 of ref.[1] which gave R^2 = 0.658 and RMSE = 4.502 when applied to the current dataset and (ii) the seven parameter multilinear model published in Table 2 of ref.[2] which performed slightly better (R^2 = 0.728 and RMSE = 4.579). However, these results are of moderate quality and do not describe satisfactory LCBs of the current dataset which has considerably increased diversity.

BML results. For successful QSPR modeling, the data investigated should posses a normal distribution; furthermore the statistical parameters (mean standard deviation and skewness) for the general population and for the samples should have similar magnitudes. As can be seen from the histograms shown in Fig. 1 for the general population and Fig. 2 for the samples, the data does indeed possess a normal distribution in terms of their mean standard deviation and skewness.

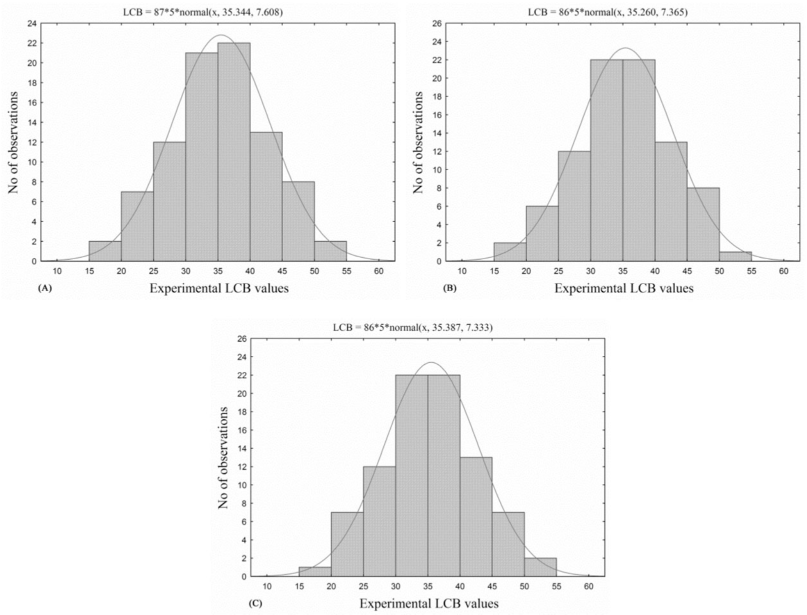

FIG. 2
Data distribution for the A, B and C subsets

Best multilinear regression models including up to seven descriptors were generated. To avoid “over-parametrization” of the models, $\Delta R^2 = R^2{}_n - R^2{}_{n-1} \leq 0.02$ was chosen as a stopping criterion[37]. Thus, six- and seven-parameter submodels (Table II) were chosen as optimal. In Tables II and III, X denotes the regression coefficients, ΔX their errors, t-test is the Student criterion, F represents the Fisher criterion, $R^2{}_{\mathrm{cv}}$ denotes the square of the leave-one-out cross-validated correlation coefficient and s^2 the standard deviation of the regression.

The descriptors appearing in Table II for the submodels of datasets A+B, A+C and B+C are quite similar, with small differences due to the procedure applied for the descriptor selection in the BLMR method (see the intercorrelation matrix shown in Table IV). Depending on the dataset, different (but physically similar and highly intercorrelated) descriptors may appear in the different models. Namely, only one of a pair or a set of highly intercorrelated descriptors is used in the further model development.

In the next stage of the modeling process, we built a general QSPR model based only on the descriptors proven to be effective for the submodels (see Table II). This subset of 14 unique descriptors was further treated by the BLMR procedure and a general model for all 259 compounds was derived. The statistical parameters of the general six parameter model obtained are shown in Table III and Fig. 3.

Using the model of Table III, the LCBs for all compounds were predicted and the results are shown in Table I (column 6).

To examine the sensitivity of the proposed QSAR model to chance correlations a scrambling procedure was applied, i.e., the model was fitted to randomly reordered activity values and then compared with the one obtained for the actual activities[38]. Ten randomizations (Table V), resulting in average $R^2 = 0.025$ were performed. The substantial difference between the R^2 of the general model of Table III and the averaged R^2 from the scrambling procedure proves the stability of the model.

Despite the satisfactory quality of the general MLR model presented in Table III, none of the descriptors of Table III showed individual strong linear relationships with the LCB. In order to investigate possible nonlinearities and to avoid the limitations imposed by the multilinear method we shifted the focus of our research to the application of SVM and PPR for QSPR modeling. These descriptors were used as inputs to develop nonlinear models by SVM and PPR. Again, datasets constructed by the three binary sums combinations: A+B, A+C and B+C were used as training subsets and C, B and A, respectively were used as external validation datasets by considering their consistency. By averaging the predicted results, the final predicted

TABLE II
Best six- and seven-parameters BMLR submodels

ID	X	ΔX	t-Test	Descriptor
$R^2_{\mathrm{trAB}} = 0.857$; $R^2_{\mathrm{cv}} = 0.842$; $F = 140.87$; $s^2 = 8.329$; $R^2_{\mathrm{testC}} = 0.669$				
0	21.236	0.929	22.870	Intercept
1	1.993	0.140	14.262	Structural Information content (order 0)
2	0.031	0.002	14.660	Tot molecular 1-center E-N attraction/No. of atoms
3	–10.195	0.969	–10.522	Min net atomic charge
4	–4.875	0.587	–8.302	Number of S atoms
5	3.838	0.504	7.614	HACA-2 [Zefirov's PC]
6	125.530	20.036	6.265	Max SIGMA-PI bond order
7	0.020	0.005	4.484	(1/2)X BETA polarizability (DIP)
$R^2_{\mathrm{trAC}} = 0.805$; $R^2_{\mathrm{cv}} = 0.782$; $F = 97.39$; $s^2 = 11.276$; $R^2_{\mathrm{testB}} = 0.699$				
0	12.217	2.838	4.305	Intercept
1	1.980	0.163	12.169	Structural Information content (order 0)
2	–20.001	2.202	–9.084	Relative number of F atoms
3	4.656	0.575	8.101	HACA-2 [Zefirov's PC]
4	–10.189	1.340	–7.606	Min net atomic charge
5	–59.481	9.006	–6.605	Relative number of S atoms
6	–6.136	2.070	–2.963	RPCG Relative positive charge (QMPOS/QTPLUS) [Zefirov's PC]
7	18.848	4.691	4.018	Molecular volume/XYZ Box
$R^2_{\mathrm{trBC}} = 0.783$; $R^2_{\mathrm{cv}} = 0.756$; $F = 99.12$; $s^2 = 12.085$; $R^2_{\mathrm{testA}} = 0.773$				
0	23.470	1.142	20.548	Intercept
1	1.088	0.103	10.532	Bonding Information content (order 1)
2	–12.620	1.222	–10.330	Min net atomic charge
3	–1.246	0.152	–8.200	Number of F atoms
4	2.014	0.339	5.942	Count of H-acceptor sites [Zefirov's PC]
5	–4.122	0.703	–5.864	Number of S atoms
6	–10.070	1.955	–5.151	RPCG Relative positive charge (QMPOS/QTPLUS) [Zefirov's PC]

results for each compound were calculated and used to generate the regression plots between the predicted and experimental LCB values for both SVM and PPR models.

SVM modeling results. The quality of SVM depends on a good choice of the following parameters: kernel type *K* and its corresponding parameters γ, capacity parameter *C*, and ε-insensitive loss function. The most important parameter is the kernel type *K* because it, together with its corresponding

TABLE III
Best six-parameter linear QSPR model

ID	*X*	Δ*X*	*t*-Test	Descriptor
0	23.68	0.935	25.34	Intercept
1	–13.05	1.749	–13.40	Min net atomic charge
2	1.011	0.9734	12.71	Bonding Information content (order 1)
3	–16.12	0.08012	–9.217	Relative number of F atoms
4	2.333	0.2631	8.861	Count of H-acceptor sites [Zefirov's PC]
5	–5.013	0.5864	–8.552	Number of S atoms
6	–9.799	1.772	–5.531	RPCG Relative positive charge (QMPOS/QTPLUS) [Zefirov's PC]

$N = 259$; $R^2 = 0.786$; $R^2_{cv} = 0.769$; RMSE = 3.424, $F = 153.8$; $s^2 = 12.05$

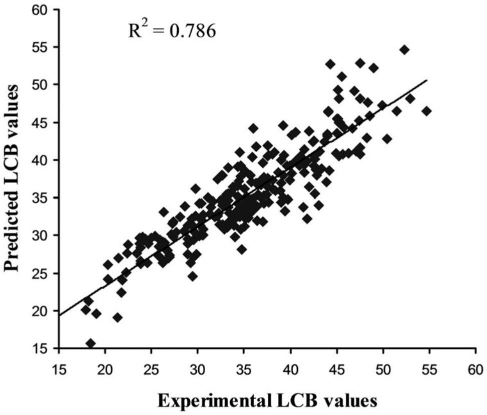

FIG. 3
Predicted versus experimental LCBs by general MLR model

TABLE IV
Correlation matrix

ID	Descriptor	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Number of F atoms	1.00	0.93	–0.08	–0.05	–0.02	0.11	–0.12	–0.06	0.00	0.20	0.09	0.80	–0.04	0.13
2	Relative number of F atoms		1.00	–0.06	–0.03	–0.11	0.02	–0.10	–0.05	0.02	0.34	0.11	0.88	–0.04	0.16
3	Number of S atoms			1.00	0.87	–0.14	–0.10	0.03	–0.14	–0.07	–0.01	0.11	0.04	–0.07	–0.29
4	Relative number of S atoms				1.00	–0.22	–0.20	0.05	–0.12	–0.07	0.11	0.12	–0.07	–0.09	–0.25
5	Bonding Information content (order 1)					1.00	0.89	–0.29	0.15	0.12	–0.37	–0.03	0.15	0.29	–0.14
6	Structural Information content (order 0)						1.00	–0.44	0.01	0.06	–0.30	0.06	–0.01	0.20	–0.35
7	Molecular volume/XYZ Box							1.00	0.00	0.06	–0.16	–0.04	0.15	0.13	0.45
8	HACA-2 [Zefirov's PC]								1.00	0.83	0.05	0.04	0.04	0.16	0.29
9	Count of H-acceptor sites [Zefirov's PC]									1.00	0.03	0.01	–0.01	0.15	0.30
10	RPCG Relative positive charge [Zefirov's PC]										1.00	–0.02	–0.56	–0.16	–0.11
11	Max SIGMA-PI bond order											1.00	–0.15	0.04	–0.04
12	Tot molecular 1-center E-N attraction/No. of atoms												1.00	0.05	0.03
13	(1/2)X BETA polarizability (DIP)													1.00	0.13
14	Min net atomic charge														1.00

parameters γ, defines the distribution of the training set examples in the high-dimensional feature space (mapping space) plus the linear model constructed in this space and therefore, controls the generalization ability of SVM. Factor γ greatly affects the number of support vectors (SVs) used to construct the regression function: too many SVs can produce overfitting and make the training time longer. Gaussian radial basis function (RBF) was preferred in this study because of its effectiveness and speed in the training process. *C* is the regularization, which controls the trade-off between maximizing the margin and minimizing the training error: a too small *C* leads to insufficient fitting on the training data while a too large *C* results in overfitting on the training data. Factor ε depends on the quality of the noise present in the data, which is usually unknown. The value of ε can also affect the number of SVs: the larger ε, the fewer SVs are selected but with a risk to distort the data.

As the three parameters (γ, ε and *C*) influence each other, a systematic grid search (GS) method was utilized to determine the best combination. The optimal model setting parameters are included in Table VI along with the statistical parameters for the training, test and the whole dataset, respectively. As expected, the results for the whole dataset and those based on the training and test subsets are very close to each other, thus demon-

TABLE V
Scrambling procedure statistical results

#	R^2	*s*	*F*
1	0.014	55.43	0.577
2	0.029	54.57	1.244
3	0.007	55.77	0.309
4	0.028	54.60	1.218
5	0.020	55.09	0.835
6	0.030	54.51	1.297
7	0.016	55.28	0.684
8	0.009	55.67	0.387
9	0.062	52.71	2.779
10	0.030	54.49	1.308
Average	0.025	54.80	1.064

TABLE VI
Statistical results and corresponding model parameters for three models by SVM and PPR

Method	Model parameters	Train set			Test set			Total set	
			R^2	RMSE		R^2	RMSE	R^2	RMSE
SVM	C = 100, γ = 0.01, ε = 0.2	B+C	0.869	2.661	A	0.851	2.946	0.863	2.759
	C = 140, γ = 0.05, ε = 0.04	A+C	0.901	2.342	B	0.821	3.164	0.873	2.644
	C = 78, γ = 0.09, ε = 0.15	A+B	0.909	2.260	C	0.802	3.258	0.873	2.633
Average			0.893	2.426		0.822	3.125	0.869	2.679
PPR	nterms = 6, opt = 3, span = 0.21	B+C	0.909	2.208	A	0.833	3.103	0.882	2.544
	nterms = 6, opt = 3, span = 0.27	A+C	0.888	2.485	B	0.853	2.812	0.877	2.598
	nterms = 6, opt = 3, span = 0.32	A+B	0.897	2.397	C	0.817	3.179	0.869	2.683
Average			0.898	2.366		0.833	3.036	0.876	2.602

strating the consistence of the QSPR modeling procedure applied. By averaging the predicted from the ABC scheme results, we obtained the final LCB values shown in column 7 of Table I. Compared to the general linear model the SVM results showed significant improvement (R^2 = 0.889 and RMSE = 2.473). A plot of the predicted versus experimental LCB values is shown on Fig. 4.

PPR modeling results. Three parameters "nterms", "optlevel" and "span"[33] had to be determined. The parameter "nterms" controls the number of variables to be entered in the model, "optlevel" means the levels of optimization which differ in how thoroughly the models are refitted during this process, and "span" defines the fraction of the observations in the span of the running lines smoother. The algorithm proposed by Friedman was used where values of g_i are found by smoothing operation that entails a backfitting[35].

The optimal model setting parameters and the corresponding statistical parameters for the training, test and the whole datasets are shown in Table VI. The averaged predicted LCB values are shown in column 8 of Table I. For the whole dataset, the calculated R^2 and RMSE were 0.896 and 2.384, respectively. A graphical presentation of the relationship between the experimental and the average predicted LCBs by the PPR model is shown on Fig. 5.

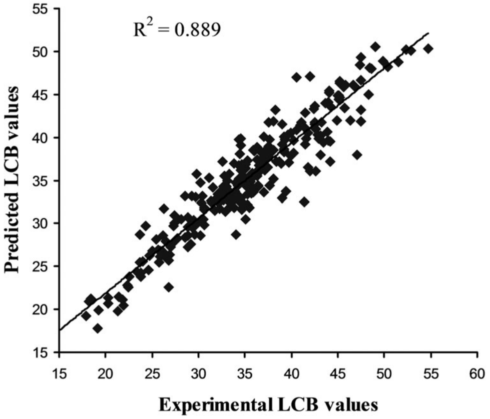

FIG. 4
Plot of the predicted versus experimental LCBs for SVM

Closer examination of the MLR, SVM and PPR residuals reveals that, generally, the datapoints from the nonlinear models have smaller deviations from the regression line than the linear models. The linear predictions for several compounds (IDs 5 (trifluoromethyl disulfide), 6 (perfluoro-*tert*-butyl alcohol), 7 (methanethiol), 9 (ethanethiol), 13 (2-propanethiol) and 228 (4-(trifluoromethyl)phenyl diphenylphosphinate)), resulted in deviations larger than 5 kcal/mol while the nonlinear models produced deviations less than 1 kcal/mol. However, for some compounds, the linear model generated much better results than the nonlinear, i.e., compounds with IDs 37 (trichloroacetonitrile), 201 (isophorone) and 209 (dimethyl phosphite). In addition, for the following compounds, the deviations were larger than 5 kcal/mol by all models: IDs 47 (pyrazine), 60 (pyrimidine), 117 (*N*,*N*-dimethylcyanoformamide), 204 (dimethyl sulfoxide), 213 (2-methoxyethanol), 216 (tetrahydrothiophene 1-oxide), 226 (1,3-diphenylpropane) and 248 (proline).

When comparing the performance of all QSPR models developed in terms of their statistical parameters and predictive power, it appears that the results obtained by the nonlinear approaches compare favorably with those obtained by the linear modeling, possibly indicating the nonlinearity exhibited in the given dataset. However, the application of both, PPR and SVM modeling procedures resulted in similar statistical parameters, demonstrating the extent of their equivalency.

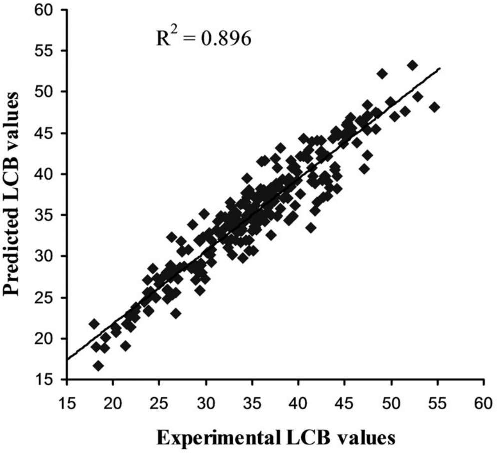

FIG. 5
Plot of the predicted versus experimental LCBs for PPR

As can be seen from Table VI for the whole dataset the nonlinear methods produced very similar results, which outperform significantly those obtained by the general MLR approach in the following order: PPR > SVM > MLR.

The above results show that the SVM and PPR approaches, although they seem rather different in their concepts, provide quite similar statistical parameters, thus showing comparable performance. Therefore, the choice of SVM or PPR as modeling techniques in treating LCB becomes a personal preference of the researcher.

Comparison of the performance of the MLR, SVM, PPR, MLR-CNN and GA-CNN methods. The performance of the now utilized SVM and PPR methods was compared directly to the performance of the models in refs[1,2] using the corresponding descriptor and datasets. The results are summarized in Table VII. As can be seen the performance of the models decreases in the following order: GA-CNN > PPR > SVM > MLR-CNN > MLR. However, it was not a surprise that the GA-CNN produced better results than any of the other nonlinear techniques – its nonlinear feature (descriptor) selection mechanism is superior to all linear future selection procedures employed by the other methods. Again, the SVM and PPR approaches generated comparable results, with PPR being slightly better in terms of R^2 and RMSE. As a linear method MLR was found inferior to the other techniques, producing results characterized by significantly lower statistical parameters.

Interpretation of the descriptors in the model. The QSPR model developed should provide an accurate prediction for the studied property but also helps to understand the underlying physical phenomenon and identification of the key physical variables. By interpreting the descriptors in the MLR model, it is possible to gain some insight into structural features that

TABLE VII
A comparison of the statistical parameters for the MLR, SVM, PPR, MLR-CNN and GA-CNN models

Statistical parameters	Ref.[1], N = 205			Ref.[2], N = 229				
	MLR	SVM	PPR	MLR	MLR-CNN	GA-CNN	SVM	PPR
R^2	0.751	0.876	0.881	0.793	0.899	0.939	0.900	0.902
RMSE	3.358	2.343	2.302	3.296	2.300	1.792	2.279	2.264

affect the lithium cation basicities. Among the six descriptors, two are constitutional (*RN_F*, relative number of fluorine atoms, *N_S*, number of sulfur atoms), one is topological (*^1BIC*, first order of bonding information content), two are electrostatic (*CHA*, count of H-acceptor sites, *RPCG*, relative positive charge (QMPOS/QTPLUS) [Zefirov's PC]) and the remaining one is of quantum chemical origin (*Q_{min}*, minimum net atomic charge). These descriptors encode different information affecting lithium cation basicity.

As the descriptors are not normalized, the value of the coefficients cannot be treated as an indicator of the importance of the descriptor in an equation or a model. Instead, the *t*-test values for each of the descriptors have been used for this purpose. Descriptors with larger absolute *t*-test values are considered statistically more significant for the description of the studied property (LCB).

The most significant descriptor, *Q_{min}*, contributes to the intensity of the electrostatic, in particular the Coulombic, interactions. Logically, it is the most important descriptor as Li^+ forms highly ionic bonds with bases. This descriptor could be also related to the hydrogen-bonding (HB) formation phenomena, because a high positive value on the hydrogen atoms implies good HB donor propensity, whereas a high negative value on heteroatoms (N, O, F, S and P) implies good acceptor ability. In our previous treatment[1], this descriptor was also selected as the most significant one. The second most important descriptor *^1BIC* is defined on the basis of the Shannon information theory. It reflects the branching of the molecule and its "informational richness". The "informational richness" describes how many different types of atoms build the molecule and how diverse the branching of these atoms is at zero to second valence level. Thus, it may describe the difference of the steric properties of the molecules and undoubtedly affects the formation of Li^+-base bond. Following are two constitutional descriptors, *RN_F*, which is defined as the ratio of the number of fluorine atoms to the total number of atoms in the base, and *N_S*, number of sulfur atoms in the base. Both descriptors are related to the basic characteristics of the substrates, since they refer to the number of basic heteroatoms. Their importance in the models can be rationalized as a measure of the local polarizability, as the polarizability of base should influence significantly the strength of the Li^+-base bond. As in our previous work[1] and that reported by Jover et al.[2], both descriptors are characterized by negative regression coefficients. *CHA* and *RPCG* belong to charged partial surface area (*CPSA*) descriptors invented by Jurs et al.[39]. As a hydrogen-bond-acceptor descriptor, *CHA* reflects hydrogen bond basicity. Acceptor groups include any functional group possessing sufficient electron density to participate in a hydrogen bond. *RPCG* is the charge on the most positive atom divided by the total charge summed over all positive atoms and was developed and used to account for the effects of polar intermolecular interactions[40]. It also encodes indirect information on the size of the molecule via the sum of the partial positive charges. The above two descriptors show the importance of the electrostatic interactions between the molecules and suggest that they are polar in nature. Combined together, these factors reflect the electrostatic nature of the Li^+-base interaction. They also affect the Li^+-base bond in a complex way and this relationship can be correlated more accurately in a nonlinear manner using nonlinear surface methodology such as SVM and PPR.

CONCLUSIONS

QSPR models were developed for 259 highly diverse compounds to predict their lithium cation basicity (LCB) and study the relationships between LCB and structural characteristic features, which were represented by molecular descriptors calculated by CODESSA software. Seven descriptors selected by the best multilinear regression (BMLR) method implemented in CODESSA were used as input vectors of two nonlinear modeling methods, i.e., support vector machine (SVM) and projection pursuit regression (PPR). A comparison of the results by these models demonstrated the superiority of nonlinear methods in predicting LCB. In addition, the analysis of the six descriptors indicates that the LCB of the studied compounds depends mainly on electrostatic features.

REFERENCES

1. Tämm K., Fara D. C., Katritzky A. R., Burk P., Karelson M.: *J. Phys. Chem. A* **2004**, *108*, 4812.
2. Jover J., Bosque R., Sales J.: *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1727.
3. *NIST Chemistry Web Book Standard Reference Database No. 69*. July 2001 Release.
4. Smith B. J., Radom L.: *J. Am. Chem. Soc.* **1993**, *115*, 4885.
5. Freiser B. S.: *Organometallic Ion Chemistry*. Kluwer, Dordrecht 1996.
6. Staley R. H., Beauchamp J. L.: *J. Am. Chem. Soc.* **1975**, *97*, 5920.
7. Woodin R. L., Beauchamp J. L.: *J. Am. Chem. Soc.* **1978**, *100*, 501.
8. Keesee R. G., Castleman A. W.: *J. Phys. Chem. Ref. Data* **1986**, *15*, 1011.
9. Taft R. W., Anvia F., Gal J.-F., Walsh S., Capon M., Holmes M. C., Hosn K., Oloumi G., Vasanwala R., Yazdani S.: *Pure Appl. Chem.* **1990**, *62*, 17.
10. Speers P., Laidig K. E.: *J. Chem. Soc., Perkin Trans. 2* **1994**, 799.
11. Fujii T.: *Mass Spectrom. Rev.* **2000**, *19*, 111.
12. Maeda H., Irie M., Than S., Kikukawa K., Mishima M.: *Bull. Chem. Soc. Jpn.* **2007**, *80*, 195.

13. Tissandier M. D., Cowen K. A., Feng W. Y., Gundlach E., Cohen M. H., Earhart A. D., Coe J. V., Tuttle T. R., Jr.: *J. Phys. Chem. A* **1998**, *102*, 7787.
14. Wieting R. D., Staley R. H., Beauchamp J. L.: *J. Am. Chem. Soc.* **1975**, *97*, 924.
15. Alcami M., Mó O., Yáñez M., Anvia F., Taft R. W.: *J. Phys. Chem.* **1990**, *94*, 4796.
16. Buncel E., Decouzon M., Formento A., Gal J. F., Herreros M., Li L., Maria P. C.: *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 262.
17. Gal J. F., Maria P. C., Decouzon M.: *Int. J. Mass Spectrom.* **2002**, *217*, 75.
18. Dzidic I., Kebarle P.: *J. Phys. Chem.* **1970**, *74*, 1466.
19. Rodgers M. T., Armentrout P. B.: *J. Phys. Chem. A* **1997**, *101*, 2614.
20. Lin C. Y., Dunbar R. C.: *Organometallics* **1997**, *16*, 2691.
21. Feng W. Y., Gronert S., Lebrilla C.: *J. Phys. Chem. A* **2003**, *107*, 405.
22. García-Muruais A., Cabaleiro-Lago E. M., Hermida-Ramón J. M., Ríos M. A.: *Chem. Phys.* **2000**, *254*, 109.
23. Remko M., Liedl K. R., Rode B. M.: *J. Phys. Chem. A* **1998**, *102*, 771.
24. Burk P., Koppel I. A., Koppel I., Kurg R., Gal J. F., Maria P. C., Herreros M., Notario R., Abboud J. L. M., Anvia F., Taft R. W.: *J. Phys. Chem. A* **2000**, *104*, 2824.
25. Gal J. F., Maria P. C., Mó O., Yáñez M., Kuck D.: *Chem. Eur. J.* **2006**, *12*, 7676.
26. Gal J. F., Maria P. C., Decouzon M., Mó O., Yáñez M.: *Int. J. Mass Spectrom.* **2002**, *219*, 445.
27. Gal J. F., Maria P. C., Decouzon M., Mó O., Yáñez M., Abboud J. L. M.: *J. Am. Chem. Soc.* **2003**, *125*, 10394.
28. Becke A. D.: *J. Chem. Phys.* **1993**, *98*, 5648.
29. Burk P., Sults, M.-L., Tammiku-Taul J.: *Proc. Estonian Acad. Sci. Chem.* **2007**, *56*, 107.
30. Hallmann M., Raczynska E. D., Gal J. F., Maria P. C.: *Int. J. Mass Spectrom.* **2007**, *267*, 315.
31. www.hyper.com
32. www.codessa-pro.com
33. Karelson M.: *Molecular Descriptors in QSAR/QSPR*. Wiley-Interscience, New York 2000.
34. Vapnik V. in: *Advanced in Kernel Methods: Support Vector Learning* (B. Scholkopf, B. Burges and A. Smola, Eds). The MIT Press, Cambridge, MA 1999.
35. Friedman J. H.: *J. Am. Stat. Assoc.* **1987**, *82*, 249.
36. www.r-project.org
37. Katritzky A. R., Kuanar M., Fara D. C., Karelson M., Acree W. E.: *J. Bioorg. Med. Chem.* **2004**, *12*, 4735.
38. Eriksson L., Jaworska J., Worth A. P., Cronin M. T. D., McDowell R. M., Gramatica P.: *Environ. Health Perspect.* **2003**, *111*, 1361.
39. Stanton D. T., Jurs P. C.: *Anal. Chem.* **1990**, *62*, 2323.
40. Sannigrahi A. B.: *Adv. Quantum Chem.* **1992**, *23*, 301.