Determination of Lithium Cation Basicity from Molecular Structure

Jesús Jover, Ramón Bosque, and Joaquim Sales*

Departament de Química Inorgànica, Universitat de Barcelona, Martí i Franquès, 1, 08028-Barcelona, Spain

Received May 19, 2004

A quantitative structure−property relationship (QSPR) is developed to calculate the Lithium Cationic Basicity (LCB) of a large set of 229 compounds, of very different chemical nature. The proposed models derived from multiple linear regression analysis (MLRA) and computational neural networks (CNN) contain seven descriptors calculated solely from the molecular structure of compounds. The models were validated by an external prediction set. Good results were obtained from both methodologies, being the best those from CNN, that give a rms error of 6.54 ($R^2 = 0.954$) and an average error of 3.57% for the training set; for the prediction set the rms error is 8.61 ($R^2 = 0.914$) and the average error 4.39%. The models derived from the two approaches contain descriptors that belong to the same classes, constitutional and electrostatic. The comparison with the results obtained from high level theoretical methods shows that the values obtained from the QSPR approach are very similar and even better, especially when the sets compared are large and contain compounds of different chemical structure. These good results shows that, despite the complexity of Li^+-base interactions, the proposed models contain descriptors which encode properly the characteristics of the molecules directly related to their gas-phase basicity against the Li cations.

INTRODUCTION

The acid−base interactions are, undoubtedly, one of the most important processes in chemistry and biochemistry. Quantitative studies in the gas phase provide the intrinsic acidities and basicities of chemical substances, free of interference of solvent molecules and counterions. The most widespread studies concern different gas-phase proton-transfer equilibrium. Obviously, the proton affinity (PA) and the gas basicity (GB) have been the most studied processes, and values for these magnitudes are known for many inorganic and organic compounds.[1] These magnitudes are defined in terms of the hypothetical gas-phase reaction

$$B_{(g)} + H^+_{(g)} \rightarrow B - H^+_{(g)} \qquad (1)$$

being PA and GB, the negative of the enthalpy change and of the Gibbs free energy, respectively, of the reaction 1. These properties of the isolated molecules can throw some light on the bulk behavior of the chemical compounds, giving also information on their acid−base characteristics in solution. Thus, several relationships have been established between the acid−base properties in both gas phase and solution of different kinds of chemical substances.[2] Very recently, the pK_a's of various organic acids in DMSO have been predicted from the theoretically derived values of their gas-phase acidity.[3]

Besides the proton, the lithium ion, Li^+, was the first metal cation to be studied in the gas phase. Analogously, the gas-phase lithium cation basicity (LCB) is defined as the negative of the Gibbs free energy associated with the reaction 2

$$B_{(g)} + Li^+_{(g)} \rightarrow B - Li^+_{(g)} \qquad (2)$$

The coordination properties of Li^+ are very different of H^+. Proton adds to the base, giving a polar covalent σ bond with a very extensive charge transfer, while the bonds formed by Li^+ are largely due to ion−dipole (electrostatic) interaction. As a result, the LCBs are much smaller than GBs and cover a narrower range in the energy scale. On the other hand, the Li cation can form chelate complexes in which the metal bridges two or more basic centers of the base. The different bonding types in H^+ and Li^+ adducts lead to widely varying basicity orders.[4]

Most available data on gas-phase basicities and acidities have been obtained from experiments in which the equilibrium constants of cation transfer reactions are determined. The LCB values are obtained by means of different experimental techniques, mass spectrometry being a frequently used one. Fourier transform ion cyclotron resonance (FT-ICR),[5] equilibrium constant determination by high-pressure mass spectrometry (HPMS),[6] unimolecular dissociation (the Cook's kinetic method),[7] and energy-resolved collision-induced dissociation (CID)[8] are also widely used. On the other hand, several theoretical methods at different levels, ab initio (G2 and G2MP2) and density functional theory (B3LYP/6-311+G**), have been used for the calculation of the LCB.[4] The results obtained are well correlated with the experimental values.

The Quantitative Structure−Property Relationship approach (QSPR) has become very useful in the prediction and interpretation of several physical and chemical properties. The basis of such relationships is the assumption that the variation of behavior of the compounds, as expressed by any measured physical or chemical properties, can be correlated with changes in molecular features of the compounds termed descriptors. While the traditional approach often needs some intuitive vision to derive the relevant mathematical relationship, QSPR methods are based on statistically determined linear or nonlinear functional forms that relate the property

* Corresponding author fax: +34934907725; e-mail: joaquim.sales@qi.ub.es.

10.1021/ci0498362 CCC: $27.50
Published on Web 08/25/2004

of interest with descriptors. Descriptors are numerical values used to describe different characteristics about certain structure in order to yield information about the property being studied. The QPSR approach has been successfully applied to the correlation of many diverse physicochemical properties of chemical compounds. Thus, its application to the estimation of technologically relevant physical properties has been reviewed.[9] Other properties as dissociation energies,[10] reaction rates,[11] chromatographic retention parameters,[12,13] and NMR chemical shifts have been also studied.[14,15]

In relation to the prediction of acid−base constants, several approaches similar to QSPR, such as QSAR, LFER, and CoMFA, have been applied to the estimation of aqueous pK_a values of different kind of organic compounds (carboxylic acids, phenols, amines, etc.). These methods estimate the pK_a values from several molecular descriptors which are either experimental or theoretically derived.[16] Theoretical linear solvation energy relationships approach (TLSER), using theoretically determined parameters, has also been used to the calculation of gas-phase acidities for different sets of organic compounds.[17]

The gas-phase lithium cation basicity can be an appropriate property to be studied by QSPR methodology, because its numerical value depends on the molecular structure; moreover, there are not solvent effects that must be taken into account. Therefore, good LCB predictions can be expected from the molecular descriptors, despite the very complex character of the Li^+-base interactions.

In this paper we study linear (multiple linear regression) and nonlinear (computational neural networks) QSPR models for LCB determination for a large set of compounds, to calculate these values from the molecular structure only. A good knowledge of the influence of the effects of molecular structure on the different types of interactions observed in the Li^+-base systems can provide a better understanding of the basicity in gas phase and also in solution. The set studied contains 229 compounds of very different chemical nature including organic compounds with different functional groups, fluoro derivatives, fused rings, amino acids, and several inorganic compounds such as water, ammonia, SO_2, and $POCl_3$.

DATA AND COMPUTATIONAL METHODS

Data Set. The data set was comprised of 229 molecules whose experimental LCBs are known. The values have been taken mainly from the paper of Burk et al.[4] containing more than 200 compounds with different functional groups. Thus, the set contains O-bases: alcohols, ethers, carbonyl groups, S═O (sulfoxides and sulfones) and P═O compounds; N-bases: cyano, amines, amides; S-bases: thiols, thioethers; P-bases: fosfines, fosfine oxides; and inorganic compounds: water, ammonia, SO_2, $POCl_3$. The set contains also compounds whose LCB have been determined very recently. These compounds are three alkyl-benzenes (ethyl-, *n*-butyl, and *n*-heptyl),[18] methyl benzoate and the three isomeric dimethyl phthalates,[19] fifteen amino acids,[20] and four aromatic bases with a different number of fused rings (naphthalene, azulene, anthracene, and phenantrene).[21] Table 1 contains the experimental and the calculated LCB for all the compounds studied. The LCB ranged from 74.80 to 221.33 kJ/mol, with a mean value of 145.60 kJ/mol.

Structural Descriptors. The generation of the descriptors was performed with the CODESSA program.[22] The structures of the compounds were drawn with HyperChem Lite (Hypercube, Inc), and the geometries were fully optimized, without symmetry restrictions, using the semiempirical method AM1[23] implemented in the MOPAC 6.0 program.[24] In all cases frequency calculations have been performed in order to ensure that all the calculated geometries correspond to true minima. The MOPAC output files were used by the CODESSA program to calculate about 600 descriptors, which can be classified in five classes: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.); topological (Wiener index, Randic indices, Kier-Hall shape indices, etc.); geometrical (moments of inertia, molecular volume, molecular surface area, etc.); electrostatic (minimum and maximum partial charges, polarity parameter, etc.); and quantum (reactivity indices, dipole moment, HOMO and LUMO energies, etc.). In the calculation of the electrostatic descriptors the program uses partial charges derived from the empirical approach proposed by Zefirov,[25] based on the Sanderson electronegativity. Many of these electrostatic descriptors are also calculated using the charges derived from the quantum-chemical methods. There is a specific type of electrostatic descriptors, called Charge Partial Surface Area descriptors (CPSA), proposed by Jurs et al.[26] These descriptors are based on the surface area of the whole molecule and on the charge distribution in the molecule, so they combine shape and electronic information to characterize the molecule and therefore they encode features responsible for polar interactions between molecules. This set of CPSA descriptors was further developed to account for any particular type of polar interaction such as hydrogen bonding interactions.[27]

Linear and nonlinear QSPR models were built using multiple linear regression analysis (MLRA) and computational Neural Networks (CNN). CNNs have the ability to generate nonlinear methods that produce predicted values comparable to experimental ones.

Multiple Linear Regression Models. The data set was split randomly into a 185 member training set (tset) and an external prediction set (pset) of 44 compounds. To find the best correlation models, the heuristic multilinear regression procedures available in the framework of the CODESSA program were used to find the best correlation models. These procedures provide collinearity control (i.e., any two descriptors intercorrelated above 0.8 are never involved in the same model) and implement heuristic algorithms for the rapid selection of the best correlation, without testing all possible combinations of the available descriptors. After the heuristic reduction the pool of descriptors was reduced to 180.

The goodness of the correlation has been tested by the coefficient regression (R^2), the F-test, and the standard deviation (SD). The root-mean-square (rms) error and the average error, calculated as the mean of the absolute value of 100[(calc-exptl)/exptl], are also reported for the training and prediction sets. The stability of the correlations was tested against the cross-validated coefficient, R^2_{cv}, which describes the stability of a regression model obtained by focusing on the sensitivity of the model to the elimination of any single data point. The *t*-test of each coefficient as well as the standardized regression coefficients (beta) are also reported. To further validate the model other tests were

Table 1. Experimental and Calculated LCB (kJ/mol)

no.	compound	experimental	MLRA	error[e]	Method I	error[e]	Method II	error[e]
1	$(C_6H_5O)_3PO$[a,c]	189.12	198.95	5.20	187.72	0.74	191.28	1.14
2	$(CF_3)_2CHOMe$[a,c]	109.62	110.99	1.25	110.34	0.66	114.63	4.57
3	$(CF_3)_2CO$[a,c]	79.91	76.50	4.27	80.69	0.96	83.32	4.26
4	$(CF_3)_3CNH_2$[a,c]	99.58	109.75	10.22	99.66	0.08	100.56	0.98
5	$(CF_3)_3COH$[a,c]	84.94	95.30	12.21	88.37	4.04	93.75	10.38
6	$(CF_3CH_2)_2O$[a,c]	122.17	109.25	10.58	109.40	10.46	114.38	6.38
7	$(CF_3CO)_2CH_2$[a,c]	114.22	129.85	13.68	122.43	7.18	120.32	5.34
8	$(CF_3CO)_2NH$[a,c]	147.28	140.31	4.73	138.91	5.68	142.93	2.95
9	$(CF_3S)_2$[a,c]	80.33	45.21	43.72	72.07	10.28	82.71	2.95
10	$(CH_2)_3OSO_2$[a,c]	153.55	151.45	1.37	155.81	1.47	136.11	11.36
11	$(CH_2)_4S$[a,c]	107.95	117.58	8.92	105.07	2.66	104.85	2.87
12	$(CH_2)_5S$[a,c]	108.37	124.42	14.81	113.92	5.13	111.58	2.97
13	$(CHF_2)_2CO$[a,c]	102.93	106.82	3.79	104.83	1.85	99.59	3.24
14	$(EtO)_3PO$[a,c]	188.70	186.55	1.14	183.50	2.75	185.87	1.50
15	$(i\text{-}PrO)_2PO(H)$[a,c]	184.51	180.59	2.13	180.60	2.12	192.08	4.10
16	$(Me_2N)_3PO$[a,c]	198.74	203.09	2.19	193.19	2.79	199.09	0.18
17	$(MeO)_2PO(H)$[a,c]	177.82	188.32	5.90	185.30	4.21	178.25	0.24
18	$(MeO)_2PO(Me)$[a,c]	184.10	180.60	1.90	187.49	1.84	171.47	6.86
19	$(MeO)_2SO_2$[a,c]	142.26	136.73	3.88	135.02	5.09	147.47	3.66
20	$(MeO)_3PO$[a,c]	182.84	166.10	9.15	166.80	8.77	171.67	6.11
21	$(MeO)MePO(Ph)$[a,c]	188.70	203.30	7.74	194.10	2.86	178.89	5.20
22	$(MeS)SO_2(Me)$[a,c]	152.30	148.32	2.61	150.17	1.40	157.75	3.58
23	$[(CF_3)_2CF]_2CO$[a,c]	91.63	90.10	1.67	89.86	1.93	86.02	6.12
24	1,2,4-triazole[a,c]	136.82	140.16	2.44	136.36	0.33	135.05	1.29
25	1,2-dimethylimidazole[a,c]	174.89	162.84	6.89	166.59	4.74	174.37	0.30
26	1,3,5-trimethylpyrazole[a,c]	160.25	161.09	0.53	163.13	1.80	164.03	2.36
27	1,4-dimethylpyrazole[a,c]	154.81	154.53	0.18	156.73	1.24	158.81	2.59
28	1,5-dimethylpyrazole[a,c]	157.32	154.76	1.63	157.02	0.19	162.28	3.15
29	1-adamantyl cyanide[a,c]	159.41	154.31	3.20	155.76	2.29	161.87	1.54
30	1-methylimidazole[a,c]	168.20	155.19	7.73	159.51	5.16	165.87	1.38
31	1-methylpyrazole[a,c]	143.51	148.84	3.71	152.27	6.10	153.88	7.22
32	2,4,5-trimethylimidazole[a,c]	178.24	174.11	2.32	177.54	0.39	173.11	2.88
33	2,5-dimethyl tetrahydrofuran[a,c]	146.44	139.64	4.64	143.30	2.14	145.72	0.49
34	2,6-difluoropyridine[a,c]	138.91	128.76	7.30	123.30	11.24	135.76	2.27
35	2-fluoropyridine[a,c]	146.86	135.72	7.58	133.45	9.13	144.69	1.48
36	3(5)-methylpyrazole[a,c]	146.86	150.35	2.38	152.05	3.54	150.89	2.75
37	3,4,5-trimethylpyrazole[a,c]	161.92	169.64	4.77	160.03	1.17	165.61	2.28
38	3-chloropyridine[a,c]	132.21	160.83	21.65	149.53	13.10	136.25	3.05
39	4-(trifluoromethyl)pyridine[a,c]	123.43	144.64	17.18	136.71	10.76	122.25	0.95
40	4-(dimethylamino)pyridine[a,c]	175.73	142.29	19.03	173.50	1.27	173.46	1.29
41	$4\text{-}CF_3PhOPO(Ph)_2$[a,c]	185.35	222.31	19.94	195.29	5.36	190.20	2.61
42	$4\text{-}MeC_6H_4SO_2(Me)$[a,c]	168.20	171.21	1.79	177.65	5.62	182.08	8.25
43	$4\text{-}NO_2C_2H_4SO_2(Me)$[a,c]	150.62	174.10	15.58	180.11	19.58	155.48	3.22
44	alanine[a,c]	180.75	176.16	2.54	177.16	1.98	185.40	2.57
45	anthracene[a,c]	141.50	132.13	6.62	134.04	5.27	140.20	0.92
46	aspartic acid[a,c]	215.48	202.71	5.93	213.35	0.98	218.76	1.52
47	azulene[a,c]	145.10	126.58	12.76	130.22	10.26	133.61	7.92
48	BrCN[a,c]	123.01	102.97	16.29	122.02	0.80	117.28	4.65
49	$C_5H_{11}CHO$[a,c]	143.93	154.78	7.54	149.71	4.02	151.15	5.01
50	$C_6H_{13}CHO$[a,c]	144.77	159.51	10.18	154.77	6.91	150.93	4.26
51	$C_6H_5OPO(Ph)_2$[a,c]	196.23	205.69	4.82	193.77	1.25	198.27	1.04
52	C_6H_6[a,c]	112.55	118.27	5.09	125.65	11.64	114.66	1.87
53	$c\text{-}C_6H_{11}CH_2OH$[a,c]	143.51	157.19	9.53	153.94	7.27	144.10	0.41
54	CCl_3CH_2OH[a,c]	127.19	145.49	14.38	143.24	12.62	122.08	4.02
55	CCl_3CHO[a,c]	113.80	136.64	20.07	133.69	17.47	109.04	4.19
56	CF_2O[a,c]	76.99	79.56	3.35	76.82	0.21	84.81	10.16
57	$CF_3(CH_2)_3NH_2$[a,c]	155.23	149.97	3.39	140.91	9.22	147.93	4.70
58	$CF_3CH_2OCH{=}CH_2$[a,c]	114.64	129.46	12.93	125.27	9.27	128.58	12.15
59	CF_3CH_2OMe[a,c]	123.85	118.09	4.65	116.92	5.60	129.55	4.60
60	CF_3CN[a,c]	89.12	80.50	9.67	85.65	3.89	88.63	0.55
61	CF_3CO_2Et[a,c]	128.03	135.04	5.48	129.53	1.17	135.92	6.16
62	CF_3CO_2Me[a,c]	120.92	122.60	1.39	120.57	0.28	124.28	2.78
63	CF_3COCH_2COMe[a,c]	147.70	157.08	6.36	149.18	1.01	146.56	0.77
64	CF_3CONH_2[a,c]	141.84	164.01	15.64	167.04	17.77	158.44	11.70
65	CF_3CONMe_2[a,c]	166.10	164.05	1.23	166.23	0.07	176.81	6.44
66	CF_3COSMe[a,c]	125.10	130.37	4.21	128.00	2.32	119.91	4.15
67	$CH_2(CN)_2$[a,c]	110.04	111.42	1.26	125.26	13.83	118.79	7.95
68	CH_2ClCN[a,c]	123.01	122.93	0.07	130.73	6.28	121.16	1.50
69	$CH_2ClPO(OEt)_2$[a,c]	184.51	193.88	5.08	190.67	3.33	93.26	4.74
70	CH_2FCN[a,c]	109.62	108.03	1.45	113.37	3.42	110.06	0.40
71	$CHCl_2CN$[a,c]	115.90	125.45	8.24	131.50	13.47	111.15	4.09
72	$POCl_3$[a,c]	145.18	115.18	20.67	121.06	16.62	145.95	0.53
73	$c\text{-}Pr_2CO$[a,c]	160.67	159.23	0.89	154.96	3.55	145.69	9.32
74	$c\text{-}PrCOMe$[a,c]	156.48	148.70	4.98	143.37	8.38	149.24	4.63

Table 1 (Continued)

no.	compound	experimental	MLRA	error[e]	Method I	error[e]	Method II	error[e]
75	dimethylacetamide[a,c]	179.08	163.84	8.51	166.12	7.23	176.69	1.33
76	dimethylisophthalate[a,c]	157.11	169.09	7.63	165.55	5.37	165.90	5.60
77	dimethylphthalate[a,c]	196.98	170.10	13.65	166.97	15.24	179.27	8.99
78	Et_2CO[a,c]	153.55	151.55	1.31	147.47	3.96	150.23	2.17
79	Et_2O[a,c]	139.33	131.84	5.37	135.83	2.51	135.76	2.56
80	Et_2S[a,c]	110.46	123.82	12.09	112.35	1.71	107.58	2.60
81	Et_3PO[a,c]	195.39	186.15	4.73	190.38	2.56	202.22	3.49
82	EtCHO[a,c]	137.24	139.21	1.44	133.39	2.80	136.33	0.66
83	EtCN[a,c]	147.70	124.55	15.67	130.97	11.32	145.82	1.27
84	ethylbenzene[a,c]	130.21	137.57	5.66	138.34	6.25	128.43	1.37
85	EtOH[a,c]	127.19	135.66	6.65	129.48	1.80	131.39	3.30
86	EtSH[a,c]	89.54	105.20	17.49	91.39	2.07	91.13	1.78
87	glutamic acid[a,c]	221.33	212.46	4.01	214.79	2.96	218.78	1.16
88	glycol sulfate[a,c]	138.07	130.09	5.78	129.27	6.37	135.21	2.07
89	glycol sulfite[a,c]	148.95	144.49	2.99	147.97	0.66	150.39	0.96
90	H_2CO[a,c]	106.27	118.95	11.92	114.08	7.35	104.56	1.61
91	H_2O[a,c]	103.34	120.25	16.36	107.43	3.95	104.83	1.43
92	HCN[a,c]	108.37	108.80	0.40	120.56	11.25	106.79	1.46
93	HCO_2-n-Pr[a,c]	143.51	157.58	9.80	154.76	7.84	153.66	7.07
94	$HCONH_2$[a,c]	157.32	154.24	1.96	148.48	5.62	154.64	1.71
95	HCONHMe[a,c]	165.69	157.07	5.20	160.89	2.89	154.05	7.02
96	*i*-BuOH[a,c]	135.98	145.57	7.05	141.52	4.07	135.20	0.58
97	*i*-BuSH[a,c]	99.16	117.81	18.81	105.33	6.22	100.00	0.85
98	*i*-BuSMe[a,c]	114.64	129.91	13.32	121.05	5.59	111.53	2.72
99	imidazole[a,c]	159.41	150.47	5.61	155.92	2.19	150.88	5.35
100	*i*-Pr_2CO[a,c]	156.90	159.15	1.44	154.91	1.27	149.99	4.40
101	*i*-Pr_2S[a,c]	120.92	131.39	8.66	125.86	4.09	119.74	0.98
102	*i*-PrOH[a,c]	135.14	141.83	4.95	136.57	1.05	133.21	1.43
103	*i*-PrSH[a,c]	93.72	112.12	19.63	97.28	3.80	99.25	5.89
104	isoleucine[a,c]	189.54	190.64	0.59	189.82	0.15	190.71	0.62
105	leucine[a,c]	189.12	191.56	1.29	192.11	1.58	189.11	0.00
106	Me_2NCN[a,c]	163.18	141.28	13.42	144.37	11.52	163.60	0.26
107	Me_2NH[a,c]	134.31	137.50	2.38	137.61	2.46	124.21	7.52
108	Me_2O[a,c]	123.43	122.44	0.80	127.16	3.02	135.57	9.84
109	Me_2S[a,c]	97.91	112.86	15.27	95.98	1.97	100.25	2.39
110	Me_2SO_2[a,c]	155.23	145.90	6.01	150.88	2.80	157.13	1.23
111	Me_3SiOMe[a,c]	144.77	157.62	8.88	166.99	15.35	150.68	4.08
112	MeC(OH)═CHCOMe[a,c]	180.33	179.06	0.71	177.48	1.58	180.35	0.01
113	MeCHO[a,c]	133.05	132.82	0.18	126.83	4.68	120.44	9.48
114	MeCN[a,c]	142.26	118.91	16.41	126.41	11.14	133.64	6.06
115	$MeCO_2Me$[a,c]	147.28	137.85	6.40	138.64	5.86	144.83	1.66
116	MeCOEt[a,c]	150.62	146.69	2.61	141.74	5.90	149.19	0.95
117	$MeCONH_2$[a,c]	166.94	164.43	1.50	168.30	0.81	158.39	5.12
118	MeCONHMe[a,c]	173.64	166.22	4.27	167.02	3.81	165.66	4.59
119	MeCOSMe[a,c]	141.42	137.97	2.44	128.36	9.23	137.85	2.52
120	$MeOCH_2CN$[a,c]	137.24	134.65	1.88	142.86	4.10	138.59	0.99
121	$MeOCONMe_2$[a,c]	166.94	158.47	5.08	157.13	5.88	160.33	3.96
122	$MeSCH_2CN$[a,c]	143.51	127.57	11.11	119.18	16.95	147.13	2.52
123	MeSH[a,c]	84.94	94.81	11.63	85.75	0.96	82.88	2.42
124	methionine[a,c]	210.87	175.39	16.83	206.05	2.29	208.04	1.35
125	methyl benzoate[a,c]	154.60	149.08	3.57	147.78	4.41	151.95	1.72
126	naphthalene[a,c]	127.70	125.47	1.74	129.43	1.36	128.20	0.40
127	*n*-Bu_2O[a,c]	152.72	148.89	2.50	152.40	0.21	152.97	0.17
128	*n*-Bu_2S[a,c]	128.03	142.71	11.47	143.27	11.91	126.25	1.39
129	*n*-BuCHO[a,c]	141.42	149.83	5.95	144.50	2.18	150.63	6.51
130	*n*-BuOH[a,c]	137.24	146.28	6.59	141.16	2.86	135.52	1.25
131	*n*-BuSH[a,c]	100.42	116.26	15.78	103.91	3.48	98.37	2.04
132	*n*-butyl benzene[a,c]	136.40	149.63	9.70	148.72	9.04	140.06	2.68
133	$NCCONMe_2$[a,c]	144.35	165.90	14.93	158.99	10.15	156.00	8.08
134	neo-C_5H_9OH[a,c]	138.49	148.87	7.50	146.75	5.97	133.78	3.40
135	NH_3[a,c]	126.36	129.04	2.12	117.74	6.82	127.62	1.00
136	*n*-heptylbenzene[a,c]	150.08	164.73	9.76	162.44	8.23	153.73	2.43
137	*n*-octylcyanide[a,c]	156.90	153.81	1.97	156.76	0.09	156.08	0.52
138	*n*-Pr_2O[a,c]	145.60	140.86	3.26	144.64	0.66	144.00	1.10
139	*n*-Pr_2S[a,c]	120.92	133.68	10.55	128.53	6.30	115.20	4.73
140	*n*-PrCN[a,c]	148.11	130.46	11.92	135.70	8.38	154.33	4.20
141	*n*-PrOH[a,c]	131.38	141.39	7.62	135.46	3.11	133.89	1.91
142	Ph_2SO[a,c]	183.68	160.58	12.58	172.56	6.05	178.02	3.08
143	Ph_3PO[a,c]	198.74	203.62	2.45	194.31	2.23	196.38	1.19
144	$PhCH_2CN$[a,c]	146.86	146.47	0.27	149.17	1.57	155.53	5.91
145	$PhCH_2OH$[a,c]	149.79	155.48	3.80	153.20	2.28	150.64	0.57
146	PhSO(Me)[a,c]	179.49	160.22	10.74	172.29	4.01	177.14	1.31
147	$PhSO_2$(Me)[a,c]	164.43	161.54	1.76	168.85	2.69	168.23	2.31
148	*p*-MeC_6H_4COMe[a,c]	159.41	163.35	2.47	160.90	0.93	167.49	5.07

Table 1 (Continued)

no.	compound	experimental	MLRA	error[e]	Method I	error[e]	Method II	error[e]
149	proline[a,c]	198.74	176.60	11.14	178.00	10.43	188.14	5.33
150	pyrazine(1,4)[a,c]	119.66	124.69	4.20	124.15	3.75	132.41	10.65
151	pyrazole[a,c]	140.58	137.98	1.85	138.97	1.14	142.11	1.09
152	pyridazine(1,2)[a,c]	173.22	140.70	18.77	159.84	7.72	154.56	10.77
153	pyridine[a,c]	146.44	138.08	5.71	143.02	2.33	143.06	2.31
154	serine[a,c]	203.34	185.80	8.63	206.73	1.67	207.08	1.84
155	SO_2[a,c]	76.15	107.29	40.89	86.62	13.75	100.02	31.35
156	*t*-Bu_2S[a,c]	128.03	138.50	8.18	137.46	7.37	134.30	4.89
157	*t*-BuCN[a,c]	152.30	135.40	11.09	140.04	8.05	156.36	2.66
158	*t*-BuOEt[a,c]	148.11	142.38	3.87	146.83	0.87	143.63	3.02
159	*t*-BuOH[a,c]	139.33	147.31	5.73	143.06	2.68	132.17	5.13
160	*t*-BuOMe[a,c]	143.09	137.44	3.95	141.74	0.94	140.88	1.55
161	tetramethylene sulfone[a,c]	163.18	151.85	6.94	157.25	3.63	170.56	4.52
162	tetramethylene sulfoxide[a,c]	180.33	149.22	17.25	159.31	11.65	175.37	2.75
163	tetramethylguanidine[a,c]	177.40	178.85	0.82	177.50	0.05	168.51	5.01
164	tetrazole[a,c]	139.33	136.90	1.74	150.59	8.08	138.32	0.73
165	threonine[a,c]	208.78	200.29	4.07	206.00	1.33	205.16	1.74
166	valine[a,c]	188.28	185.85	1.29	187.24	0.55	190.31	1.08
167	$(CF_3)_3CCO_2Et$[a,d]	144.35	135.89	5.86	128.16	11.21	129.62	10.20
168	$(MeO)_2CO$[a,d]	154.81	143.59	7.25	144.36	6.75	129.70	16.22
169	1,8-naphthyridine[a,d]	181.59	160.02	11.88	173.95	4.20	157.31	13.37
170	3-(dimethylamino)pyridine[a,d]	169.66	161.50	2.05	171.02	0.80	171.21	0.91
171	$C_6H_5SO_2(Me)$[a,d]	157.74	161.55	2.42	168.89	7.07	168.24	6.66
172	CF_3CCH[a,d]	74.89	88.93	18.74	92.79	23.90	71.04	5.15
173	$CF_3CO_2CH_2CF_3$[a,d]	107.53	112.72	4.83	112.53	4.65	113.55	5.60
174	$ClCO_2Me$[a,d]	120.92	141.24	16.81	142.91	18.19	126.82	4.88
175	dimethyl sulfoxide[a,d]	174.89	143.93	17.70	152.38	12.87	163.08	6.76
176	HCO_2-n-Bu[a,d]	143.51	162.88	13.50	160.02	11.50	162.85	13.47
177	isophorone[a,d]	173.64	170.83	1.62	167.39	3.60	177.08	1.99
178	Me_3PO[a,d]	191.21	180.18	5.77	190.69	0.27	180.89	5.39
179	MeOH[a,d]	119.24	128.65	7.89	120.78	1.29	126.62	6.19
180	*n*-heptyl cyanide[a,d]	153.97	149.57	2.86	152.88	0.71	154.51	0.35
181	phenylalanine[a,d]	202.51	199.62	1.42	202.69	0.09	196.11	3.16
182	*s*-BuOH[a,d]	139.33	146.42	5.09	142.96	2.61	134.02	3.81
183	*t*-$BuCO_2Et$[a,d]	162.76	156.31	3.96	155.51	4.45	163.17	0.25
184	*t*-BuSH[a,d]	99.58	117.94	18.44	104.08	4.52	104.28	4.72
185	valeronitrile[a,d]	149.79	135.17	9.76	139.78	6.68	154.80	3.35
186	$(4\text{-}FC_6H_4O)PO(Ph)_2$[b]	190.79	208.70	9.39	193.87	1.62	192.63	0.96
187	$(CF_3)_2CHOH$[b]	99.58	105.72	6.17	97.56	2.03	102.15	2.59
188	$(CH_2)_3OSO$[b]	164.01	151.97	7.34	162.48	0.93	152.85	6.81
189	$(EtO)_2PO(Me)$[b]	188.28	193.53	2.79	191.78	1.86	193.65	2.85
190	1,2,3-triazole[b]	134.31	140.90	4.91	152.23	13.35	143.57	6.89
191	1,3,4,5,-tetramethylpyrazole[b]	163.18	164.54	0.84	165.60	1.49	166.84	2.25
192	1,4-dioxane[b]	126.78	130.73	3.12	132.83	4.77	135.40	6.80
193	2-methyltetrahydrofuran[b]	143.51	135.21	5.78	139.01	3.14	144.91	0.97
194	3-methylpyridine[b]	152.72	148.62	2.68	152.99	0.18	147.42	3.47
195	4-methylpyrazole[b]	149.37	148.75	0.42	148.56	0.54	151.49	1.42
196	C_5F_5N[b]	93.30	98.46	5.53	91.96	1.44	94.44	1.22
197	CCl_3CN[b]	112.13	120.59	7.55	125.80	12.19	116.07	3.51
198	CF_3CH_2OH[b]	110.88	116.60	5.17	108.21	2.40	109.09	1.61
199	CF_3CHO[b]	91.21	99.08	8.63	95.36	4.55	90.28	1.02
200	CF_3COMe[b]	112.97	121.33	7.40	115.64	2.37	104.58	7.42
201	cysteine[b]	189.12	172.31	8.89	181.50	4.03	203.77	7.75
202	dimethylformamide[b]	173.64	158.16	8.91	159.49	8.14	165.16	4.88
203	dimethylterephthalate[b]	152.00	168.36	10.76	164.79	8.41	176.19	15.91
204	$EtCO_2Me$[b]	151.88	143.34	5.62	143.77	5.34	151.95	0.05
205	EtSMe[b]	104.60	118.89	13.67	104.38	0.21	104.98	0.36
206	Glycine[b]	174.05	171.98	1.19	171.56	1.43	181.12	4.06
207	HCO_2Et[b]	141.84	141.97	0.09	145.14	2.33	147.73	4.16
208	HCO_2Me[b]	135.56	136.44	0.65	139.78	3.11	135.15	0.30
209	*i*-Pr_2O[b]	148.53	142.18	4.28	146.56	1.33	145.33	2.16
210	*i*-PrCN[b]	149.37	130.31	12.76	135.64	9.19	152.64	2.19
211	isooxazole[b]	137.65	130.65	5.09	139.66	1.46	135.38	1.65
212	Me_2CO[b]	147.70	140.52	4.86	134.78	8.75	138.01	6.56
213	Me_3N[b]	133.89	138.07	3.12	140.09	4.63	125.71	6.11
214	$MeCO_2Et$[b]	150.62	142.55	5.36	142.54	5.37	154.68	2.70
215	MeCOOH[b]	136.82	146.23	6.88	141.68	3.55	131.68	3.76
216	$MeNH_2$[b]	130.96	136.19	3.99	132.24	0.98	130.12	0.64
217	$MeOSO_2(Me)$[b]	151.88	147.52	2.87	149.78	1.38	153.95	1.37
218	*n*-PrCHO[b]	139.33	144.77	3.91	139.12	0.15	146.12	4.87
219	*n*-PrSH[b]	94.14	111.30	18.23	97.59	3.67	95.09	1.01
220	Ph_2SO_2[b]	169.87	162.05	4.60	169.03	0.49	178.79	5.25
221	PhCHO[b]	157.74	142.77	9.49	137.90	12.57	149.77	5.05
222	PhCN[b]	148.53	133.24	10.30	137.73	7.27	153.42	3.29

Table 1 (Continued)

no.	compound	experimental	MLRA	error[e]	Method I	error[e]	Method II	error[e]
223	phenantrene[b]	141.08	131.79	6.59	133.72	5.22	144.24	2.23
224	PhOMe[b]	126.36	142.26	12.59	144.22	14.14	143.95	13.92
225	pyrimidine(1,3)[b]	124.68	142.44	14.24	141.92	13.82	143.88	15.40
226	tetrahydrofuran[b]	136.82	130.50	4.62	134.46	1.72	144.95	5.95
227	thiazole[b]	139.75	119.66	14.37	104.51	25.21	146.66	4.95
228	tryptophan[b]	218.82	218.37	0.21	213.03	2.65	195.40	10.70
229	tyrosine[b]	205.02	209.81	2.34	210.40	2.63	217.91	6.29

[a] Training set for MLRA. [b] Prediction set for MLRA and Methods I and II. [c] Training set for Methods I and II. [d] Cross-validation set for Methods I and II. [e] Absolute value of 100[(calcd-exptl)/exptl].

performed for the descriptors, the pairwise correlations and the variance inflation factors (VIF). The VIF values, defined as $(1\text{-}R^2)^{-1}$, were calculated to identify whether excessively high multicollinear coefficients existed among the descriptors; a VIF greater than 10 is indicative of multicollinearity. The statistics of the models have been done with the SPSS program.

The model which passed the statistical diagnosis with the smallest number of descriptors was chosen. When adding another descriptor did not improve significantly the statistics of a model, it was determined that the optimum subset size had been achieved. The optimum model size in this study was seven descriptors. Validation of the model was performed on the external prediction set of compounds withheld from working set.

Computational Neural Network Models. The computations were performed with the ADAPT (Automated Data Analysis and Pattern recognition Toolkit) program,[28,29] including feature selection routines (genetic algorithm[30] and simulated annealing[31]) and CNN procedures.[32]

The use of CNNs requires a cross-validation set (cvset) to determine when to stop the training of the neural network, to prevent their overtraining, and to be sure that the network would have good and general predictive ability. From the initial training set of 185 compounds, a subset of 19 compounds has been selected to form the cvset. Thus, with the CNNs we have worked with a training set of 166 compounds, a cross-validation set of 19 compounds, and a prediction set of 44 compounds.

Two types of CNNs studies have been done. The first, Method I, is a linear/nonlinear hybrid model because it is based on the set of descriptors chosen by the multiple linear regression, but a nonlinear CNN model is developed from these descriptors. In the second method, Method II, the reduced descriptor pool is analyzed by genetic algorithms.

Method I: Nonlinear Models Using Best MLRA Descriptors. Descriptors from the best MLRA model were passed to a CNN. The CNNs used for this analysis are three-layer, fully connected, feed-forward networks, and they have been described in detail by Jurs et al.[32,33] The number of neurons of the input layer corresponds to the number of descriptors in the model. The number of neurons in the hidden layer was considered to be optimized when addition of another neuron did not decrease the tset error significantly. The output layer contains one neuron representing the predicted LCB value. The number of observations in the data restricts the number of neurons in the hidden layer. The ratio of training set observations to adjustable parameters should be kept above 2.0 to avoid overtraining.[34] The number of adjustable parameters (AP) is computed as $AP = [(IL+1) \times HL] + [(HL+1) \times OL]$, where IL, HL, and OL denote the number of neurons in the input layer, hidden layer, and output layer, respectively. A quasi-Newton method BFGS (Broyden-Flectcher-Golfarb-Shanno[33]) was used to train the network. The training is monitored by watching the rms error of the cross-validation set compounds. When this error is minimized, training was stopped because it is believed that the network is beginning to overtrain at this point. CNN results were very dependent upon the starting weights and biases. To find a good set of these values, after initial CNN runs, a simulated annealing algorithm is run on the CNN, and the weights and biases determined are used as the starting values for the standard CNN algorithm. The final network had a 7−5−1 architecture.

Method II: Nonlinear Feature Selection and Nonlinear Modeling. This method combines nonlinear feature selection, utilizing a genetic algorithm, with a CNN having the same architecture as the Method I and the same number of descriptors of the MLRA model. This approach creates a nonlinear model, as a CNN is used to as the fitness evaluator. Just as in Method I model development, the training of a Method II model utilizes a cvset to determine when to stop training. The fitness of descriptors subsets was calculated as $COST = tset + 0.4\ |tset\text{-}cvset|$, where tset and cvset denote rms errors for the training and cross-validation sets, respectively. That is CNNs that produce training and cross-validation that are low and similar in magnitude tend to perform well in predicting properties of interest for compounds not used in the training process.

The set of 180 descriptors selected by CODESSA were imported to the ADAPT program and were subjected to the objective feature selection routines of this program, and a reduced pool of 102 descriptors was obtained and used in the nonlinear feature selection. Once the best subsets of descriptors were found, they were trained by the same procedures outlined above. After testing several models, the best one was evaluated by the external prediction set compounds.

RESULTS AND DISCUSSION

Multiple Linear Regression Models. The QSPR analysis of the lithium cation basicity values for the 185 compounds of the training set resulted in the model containing seven descriptors given in Table 2. The obtained correlation is good, $R^2 = 0.825$; $F = 119.5$; $R^2_{cv} = 0.806$, with a standard deviation of 13.06, which represents a relative SD of 8.97%; the rms error is 13.21 and the average error is 7.53%. The level of significance associated with the *t*-student coefficient shows that there is a linear correlation between LCB and

Table 2. Descriptors of MLRA Model[a]

descriptor (class)	coefficient	SD	beta	*t*-test	VIF
intercept	1.0152e+02	4.2994e+00		23.612	
number of nitrogen atoms (constitutional)	1.0527e+01	1.6557e+00	0.27	6.358	1.83
relative number of sulfur atoms (constitutional)	−1.1385e+02	2.4790e+01	−0.16	−4.592	1.18
relative number of fluorine atoms (constitutional)	−7.4552e+01	7.9737e+00	−0.34	−9.350	1.36
structural information content order(0) (topological)	4.9946e+00	6.8612e-01	0.32	7.279	1.96
FPSA-2 (electrostatic-CPSA)	1.3258e+01	1.8595e+00	0.34	7.130	2.30
HASA-2/TMSA (electrostatic-CPSA)	−4.9069e+02	1.0108e+02	−0.29	−4.854	3.54
HACA-2 (electrostatic-CPSA)	1.0671e+01	1.1392e+00	0.42	9.368	2.06

[a] $R^2 = 0.825$; $F = 119.5$; $SD = 13.06$; $n = 185$; $R^2_{cv} = 0.806$.

the descriptors. The pairwise correlations between the seven descriptors ranged from 0.016 to 0.635 with an average value of 0.22 and a mean VIF value of 2.03.

The model contains three constitutional descriptors (number of nitrogen atoms, relative number of sulfur atoms, and relative number of fluorine atoms), one topological descriptor (structural information content of order zero), and three electrostatic (FPSA-2, HASA-2/TMSA, and HACA-2) descriptors.

The topological descriptors describe the atomic connectivity in the molecule; one specific kind of these descriptors are the molecular complexity indices, which are based on the Shannon information theory,[35] the structural information content of order zero, belongs to this type. The three electrostatic descriptors belong to the CPSA type, and the partial charges have been calculated by quantum methods and have imported directly from the MOPAC program. The fractional charged partial positive surface area descriptor, FPSA-2, is the total charge weighted positive surface area (PPSA-2) divided by the total molecular surface area (TMSA). The other two descriptors are related to the hydrogen-bonding interactions. The descriptors HACA-2 and HASA-2/TMSA refer to the area-weighted surface charge of hydrogen bonding donor atoms. According to the standardized regression beta coefficients (Table 2) the seven descriptors have similar significances, being HACA-2 the most important one.

Good results are also obtained with the prediction set of 44 molecules, showing the high prediction capacity of the model. The statistical parameters are $R^2 = 0.878$, the rms is 10.45, and the average error is 6.32%. Figure 1 shows a plot of the calculated versus observed values for all the compounds studied, the training and the prediction sets.

The bonds formed by Li cation and the bases are largely due to electrostatic (ion-dipole) interactions, and the Li^+ retains 0.8−0.9 units of the positive charge in the complex. Besides this electrostatic interaction, the Li cation presents other specific interactions with the substrates, mainly the formation of chelates, which can modify widely the LCB values. These chelates can be classical, in the sense that the cation is coordinated to oxygen and/or nitrogen atoms, or nonclassical chelates with the Li^+ involving fluorine or chlorine atoms in ring formation.[4] Other types of interactions have been proposed depending on the structure of the base involved, for example, the so-called "scorpion effect" with alkylbenzenes,[18] and the formation of metal/cation π-complexes with bases containing fused rings.[21]

Classical chelates through oxygen and/or nitrogen atoms have been established for many of the compounds of the studied set. They are also present in the adducts with the

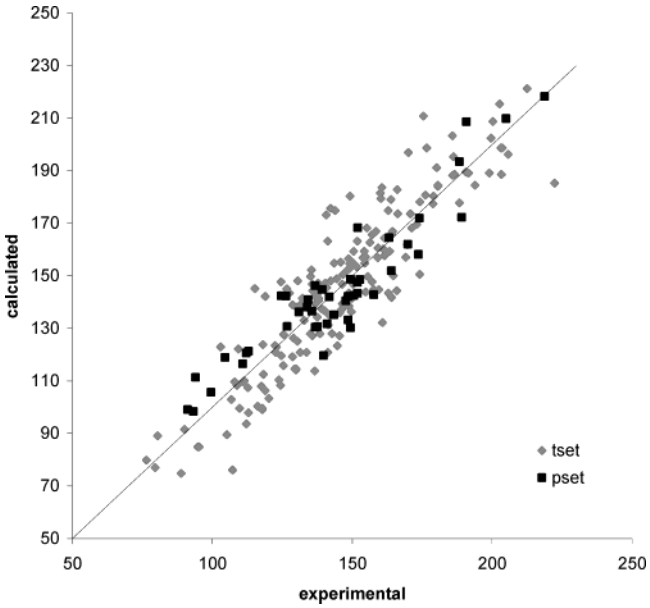

Figure 1. Plot of calculated vs experimental LCB (kJ/mol) values, using MLRA, for the tset and pset.

methyl phthalates, thus, the dimethyl phthalate is the strongest base of the three isomeric derivatives due to the chelation effect of the two carboxyl groups in the *ortho* position.[19] Regarding the amino acids, the highest LCB values are found in compounds with oxygen-bearing functional groups in the side chain, alcohols and carboxylic acids. Aspartic acid has a strong binding energy likely as a result of the formation of a very stable, six-membered ring chelation structure involving the side chain carbonyl group and the nitrogen atom. A less stable, seven-membered ring structure would result with glutamic acid, and smaller binding enhancements are also observed for compounds with aromatic side chains (phenylalanine). Conversely, glycine and alanine without coordinating groups in the side chain have the lowest lithium cation binding energies. On the other hand, the lithium complexes do not adopt zwitterionic forms.[20] The interactions of alkylbenzenes with Li^+ follow clearly the "scorpion effect", and the most stable adducts correspond to π-complexes in which the alkyl chain coils toward the aromatic ring to favor its interaction with the metal cation. The formation of these complexes leads to significant enhancement of the Li^+ binding energies (≈20−30 kJ/mol) higher than those estimated for alkylbenzene π-complexes in which an uncoiled chain remains distant from the cation.[18] The aromatic bases containing a different number of fused rings form metal-cation/π complexes with Li^+, and theoretical studies have found that the strength of the binding to an aromatic cycle decreases as the number of cycles directly

Table 3. Statistics and Descriptors of the Models for the Families Studied[a−f]

family	n	R^2	F	SD	R^2_{cv}	av error	descriptors (class)
oxygen	77	0.919	112.2	6.4	0.893	3.6	Relative number of F atoms (a); Information Content (order 0) (b); Atomic charge-weighted partial positive surface area, PPSA-3 (d); Area-weighted surface charge of a hydrogen bonding donor atom, HACA-2 (d); count of H-donor sites (c); Average Bonding Information Content (order 1) (b); Min e−n attraction for a C atom (e).
nitrogen	70	0.902	81.1	7.3	0.872	3.9	Partial positive surface area, PPSA-1 (e); Total Dipole of the molecule (e); Number of triple bonds (a); Fractional partial negative surface area, FNSA-3 (d); Max atomic state energy for a C atom (e); Min electrophilic reactivity index for a C atom (e); Average bonding Information content (order 0) (b).
sulfur	39	0.981	230.4	4.8	0.960	2.8	Total charge-weighted partial positively charged surface area, PPSA-2 (d); Max $\sigma-\pi$ bond order (e); Hydrogen bonding acceptor ability, HASA-1 (d); Relative number of S atoms (a); Relative negative charged surface area, RNCS (d); Surface-weighted charged partial positively charged surface area, WPSA-3 (d); Information Content (order 1) (b).
fluorine	35	0.960	93.6	6.6	0.920	4.0	Relative number of F atoms (a); Information content (order 0) (b); Number of N atoms (a); Average nucleophilic reactivity index for a C atom (e); Principal Moment of Inertia C (f); Total dipole of the molecule (e); Number of double bonds (a).

[a] Constitutional. [b] Topological. [c] Electrostatic. [d] Electrostatic-CPSA. [e] Quantum. [f] Geometrical.

fused to it increases. Hence, the stability of the outer π-complexes, in which Li^+ is attached to the peripheral rings, is systematically greater than that of the complexes in which the metal is attached to the inner rings.[21] All these features show the high complexity of the Li^+-base interactions and the fact that small or subtle modifications in the chemical structure of the base can modify significantly the LCB values. Despite these reasons, the obtained MLRA results are good, and they are similar to those derived from high level quantum methods (see below). As indicated before, the proposed MLRA model (Table 2) contains three constitutional descriptors (number of nitrogen atoms, relative number of fluorine atoms, and relative number of sulfur atoms) that are related to the basic characteristics of the substrates since they refer to the number of basic heteroatoms. As more nitrogen atoms are present in the molecule the more basic it is. Although the negative sign of the relative number of F and S atoms descriptors is unexpected, it is interesting to note that the QSPR studies on particular families of compounds containing fluorine or sulfur atoms (see below) give models that also contain these descriptors with negative sign. The electrostatic descriptors of the model are of the CPSA type, they have been related to intermolecular interactions, and they reflect the ion−dipole interactions between the lithium cation and the neutral molecule.

To improve the QSPR study, we have selected four subsets from the general set of 229 compounds, containing at least one atom of oxygen (77 compounds), nitrogen (70), sulfur (39) or fluorine (35); obviously, the compounds of these families can contain other heteroatoms besides their own characteristic one. The fluorine derivatives have been also included in this study because more than 15% of the bases contain this halogen, and, as it has been said before, in some cases the bases are coordinated to the Li^+ through the fluorine atoms. Table 3 shows the statistical parameters and the descriptors contained in the MLRA models for each family. Significant improvements in the factor correlation, the standard deviation, and the average error are observed for all the families, in comparison to the results of the set containing all the compounds. The models contain descriptors very similar to those found in the model for the complete set. Thus, the models for the O-, N-, and S-families contain constitutional, electrostatic, and topological descriptors, analogously to the training set; however, in the case of the F-family no electrostatic descriptors are included in the model. In each class, the descriptors belong to the same type, thus encoding very similar information. The constitutional descriptors refer to the number of atoms of oxygen, nitrogen, or sulfur elements, and it is worth noting that the relative number of fluorine and sulfur atoms also have a negative sign in the models of their respective families, as in the model of the training set. The electrostatic descriptors are of the CPSA type, and the topological descriptors are in all the cases based on the Shannon information theory.[35] The similarity of the descriptors contained in the models for the set with all the compounds and for each family shows that these descriptors encode very well the chemical characteristics of the molecules explaining their basic properties in gas phase.

Table 4. Statistics of the Methods Used for LCB Estimation

method	set	n	R^2	av error	RMS
MLRA	tset	185	0.825	7.53	13.21
MLRA	pset	44	0.878	6.32	10.45
Method I	tset	166	0.905	5.02	9.42
Method I	pset	44	0.885	4.82	9.80
Method I	cvset	19	0.880	6.60	11.53
Method II	tset	166	0.954	3.57	6.54
Method II	pset	44	0.914	4.39	8.61
Method II	cvset	19	0.877	5.92	11.47

Computational Neural Networks Models: Method I. The seven descriptors of the model derived by MLRA were imported to the ADAPT program and to CNNs routines. A 7−5−1 architecture gave good results. This neural network contains 46 adjustable parameters, corresponding to a ratio of 3.6 for training set observations (166) to adjustable parameters, above the minimum acceptable rate of 2.

The regression coefficients, R^2, the rms, and the average errors (%) obtained are 0.905, 9.42, 5.02 for the tset; 0.885, 9.80, 4.82 for the pset; and 0.880, 11.53, 6.60, for the cvset (Table 4). These results are better than those derived by MLRA, mainly for the training set. Table 1 gives the calculated values for LCB.

Computational Neural Networks Models: Method II. A genetic algorithm routine using a CNN fitness evaluator was applied to a 7−5−1 architecture for the data set. An

Table 5. Seven Descriptors Forming the Method II Model

descriptor	class
number of hydrogen atoms	constitutional
number of oxygen atoms	constitutional
number of nitrogen atoms	constitutional
HOMO-1 energy	quantum
total dipole of the molecule	quantum
total molecular electrostatic interaction/ number of atoms	quantum
area-weighted surface charge of hydrogen bonding donor atom, HDCA-2	electrostatic (CPSA)

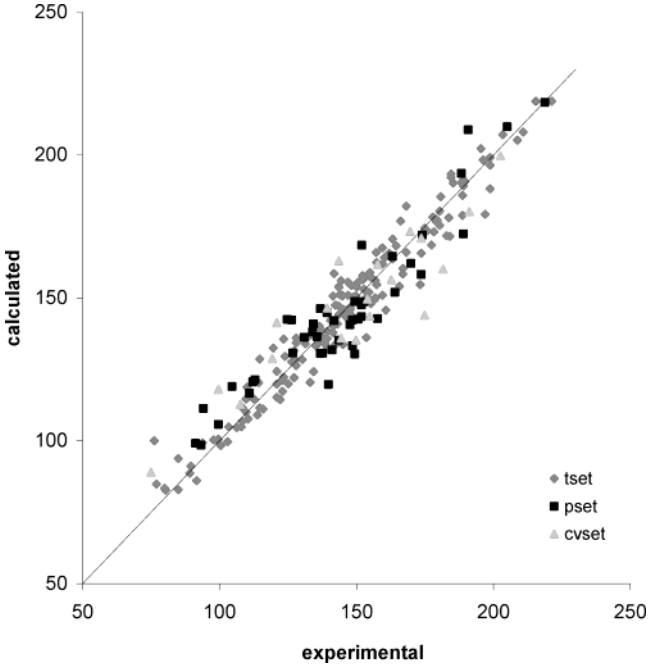

Figure 2. Plot of calculated vs experimental LCB (kJ/mol) values, using Method II, for the tset, pset, and cvset.

extensive search of the reduced descriptor pool yielded several seven-descriptor models. These fully nonlinear models were then trained and tested as above. The seven descriptors forming the best Method II model are shown in Table 5. This model contains three constitutional, one electrostatic of the type CPSA, and three quantum descriptors. Analogously to the MLRA model the numbers of atoms of different elements, hydrogen, oxygen, and nitrogen, are important in deciding the LCB values. The HDCA-2, an area-weighted surface charge of hydrogen bonding donor atoms, is also similar to CPSA descriptors of the MLRA model; this descriptor and the total dipole moment describe, mainly, the ion−dipole interaction between the base and Li^+. With this Method II model, the rms error and the average error of the training set are reduced to 6.54 and 3.57%, respectively, with $R^2 = 0.954$. For the prediction set these errors are 8.61 (rms) and 4.39% and ($R^2 = 0.914$) (Table 4). For the training set the errors have improved about 50% and for the prediction set about 30%. The values obtained for LCB are shown in Table 1 and plotted in Figure 2.

As it was mentioned before, some theoretical calculations have been done for the LCB determination. LCB values have been calculated for a set of 36 compounds[4] using standard ab initio at G2 and G2MP2 levels and DFT methods. Density functional theory calculations at the B3LYP/6-311+G** level have also been used in the calculation of LCB for another set of 66 compounds.[4,18,21] To test the suitability of the QSPR approach proposed, we have compared our results with those calculated by the high level theoretical methods. Table 6 shows the results obtained, with the small set of 36 compounds; the best results are those obtained from *ab initio* methods, but QSPR gives also good results, mainly using Method II. With the larger set of 66 compounds, the rms and average errors are higher than for the small set, since this large set contains compounds with more different chemical structures. For this set, the QSPR results from both approaches are better than those derived from DFT.

Table 6. Statistics of Different Methods for LCB Estimation

method	n	R^2	av error	rms
G2	36	0.974	4.69	5.14
G2MP2	36	0.974	4.57	5.21
DFT	36	0.975	14.78	6.53
MLRA	36	0.832	5.77	9.32
Method I	36	0.906	4.34	7.14
Method II	36	0.935	3.31	5.57
DFT	66	0.922	13.21	11.43
MLRA	66	0.843	6.78	10.09
Method I	66	0.875	5.71	9.15
Method II	66	0.942	3.68	6.20

CONCLUSIONS

Multiple linear regression and computational neural networks have been used in order to develop useful models for the estimation of the LCB of a set of 229 compounds of very different chemical nature. The models contain seven descriptors, which are calculated solely from the chemical structure. The obtained results with this QSPR approach are very good, taking into account the complexity of Li^+-base interactions and the wide variety of chemical structures of the compounds studied. Although both methods gave good results, the best ones have been obtained with the CNN approach. The descriptors contained in the MLRA and CNN's models belong to the same class of descriptors, so both models contain constitutional and electrostatic descriptors. The constitutional ones refer to the number of atoms of some of the elements present in the molecules, and the electrostatic descriptors are of the CPSA type.

When families of compounds containing at least one N-, O-, S-, or F-atoms are analyzed, the derived models include the same type of descriptors as before. These results show that, although the Li^+-base interactions can be very complex, the descriptors encode properly the molecular characteristics that explain the gas-phase basicity.

The comparison with the results obtained from high level theoretical methods, *ab initio* and DFT, with sets containing the same compounds, shows that the values obtained from the QSPR approach are very similar and even better, especially when the sets compared contain more and more different compounds. Consequently, the QSPR approach constitutes the most rapid method with appropriate accuracy to the LCB prediction.

ACKNOWLEDGMENT

The authors thank Prof. Peter C. Jurs (Pennsylvania State University) for giving us access to the ADAPT program. Financial support from the Catalan Government (Grant 2001 SGR 00052) is also gratefully acknowledged.

REFERENCES AND NOTES

(1) Hunter, E. P. L.; Lias, S. G. Evaluated Gas-Phase Basicities and Proton Affinities of Molecules: An Update. *J. Phys. Chem. Ref. Data* **1998**, *27*, 413−656.

(2) Taft, R. W.; Bordwell, F. G. Structural and Solvent Effects Evaluated from Acidities Measured in Dimethyl Sulfoxide and in the Gas Phase. *Acc. Chem. Res*. **1988**, *21*, 463−469.

(3) Fu, Y.; Liu, L.; Li, R.-O.; Liu, R.; Guo, Q.-X. First-Principle Predictions of Absolute pk_a's of Organic Acids in Dimethyl Sulfoxide Solution. *J. Am. Chem. Soc.* **2004**, *126*, 814−822.

(4) 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. Revised an Expanded Scale Gas-Phase Lithium Cation Basicities. An Experimental and Theoretical Study. *J. Phys. Chem. A* **2000**, *104*, 2824−2833.

(5) Marshall, A. G. Ion Cyclotron Resonance and Nuclear Magnetic Resonance Spectroscopies: Magnetic Partners for Elucidation of Molecular Structure and Reactivity. *Acc. Chem. Res*. **1996**, *29*, 307−316.

(6) Castleman, A. W.; Keesee, R. G. Ionic Clusters. *Chem. Rev*. **1986**, *86*, 589−618.

(7) Cook, R. G.; Patrick, J. S.; Kotiaho, T.; McLukey, S. A. Thermochemical determinations by the Kinetic Method. *Mass Spectrom. Rev*. **1994**, *13*, 287−339.

(8) Rodgers, M. T.; Armentout, P. B. Noncovalent Metal−Ligand Bond Energies as Studied by Threshold Collision-Induced Dissociation. *Mass Spectrom. Rev*. **2000**, *19*, 215−247.

(9) Katritzky, A. R.; Maran. U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.

(10) Bosque, R.; Sales, J. A QSPR Study of O−H Bond Dissociation Energy in Phenols. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 637−642.

(11) Bakken, G.; Jurs, P. C. Predictions of Hydroxyl Radical rate Constants from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1064−1075.

(12) Baczek, T.; Kaliszan, R. J. Predictive approaches to gradient retention based on analyte structural descriptors from calculation chemistry. *J. Chromatogr. A* **2003**, *987*, 29−37.

(13) Bosque, R.; Sales, J.; Bosch, E.; Rosés, M.; García-Alvarez-Coque, M. C.; Torres-Lapasió, J. R. A QSPR Study of the *p* Solute Polarity to estimate Retention in HPLC. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1240−1247.

(14) Clouser, D. L.; Jurs, P. C. The Simulation of ^{13}C nuclear magnetic resonance spectra of dibenzofurans using multiple regression analysis and neural networks. *Anal. Chim. Acta* **1996**, *321*, 127−135.

(15) Bosque, R.; Sales, J. A QSPR Study of the ^{31}P NMR Chemical Shifts of Phosphines. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 225−232.

(16) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Gancia, E.; Manallack, D. T. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. *Quant. Struct.-Act. Relat.* **2002**, *21*, 473−485.

(17) Famini, G. R.; Marquez, B. C.; Wilson, L. Y. Using Theoretical Descriptors in Quantitative-Activity Relationships: Gas-Phase Acidity. *J. Chem. Soc., Perkin Trans. 2* **1993**, 773−782.

(18) Mo, O.; Yañez, M.; Gal, J.-F.; Maria, P.-C.; Decouzon, M. Enhanced Li^+ Binding Energies in Alkylbenzene Derivatives: The Scorpion Effect. *Chem. Eur. J.* **2003**, *9*, 4330−4338.

(19) Gal, J.-F.; Maria, P.-C.; Decouzon, M. Adduct formation between phathalate esters and Li^+ in the gas phase: a thermochemical study by FT-ICR mass spectrometry. *Int. J. Mass Spectrom.* **2002**, *217*, 75−79.

(20) Feng, W. Y.; Gronert, S.; Lebrilla, C. The Lithium Binding Energy of Gaseous Amino Acids. *J. Phys. Chem. A* **2003**, *107*, 405−410.

(21) Gal, J.-F.; Maria, P.-C.; Decouzon, M.; Mo, O.; Yañez, M.; Abboud, J.-L. Lithium-Cation/ π Complexes of Aromatic Systems. The Effect of Increasing the Number of Fused Rings. *J. Am. Chem. Soc.* **2003**, *125*, 10394−10401.

(22) Katritzky, A. R.; Lovanov, V. S.; Karelson, M. *CODESSA, Reference Manual V 2.13*, Semichem and the University of Florida, 1997.

(23) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. P. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(24) Stewart, J. P. P. *MOPAC 6.0 Quantum Chemistry Program Exchange;* QCPE, No 455, Indiana University, Bloomington, IN, 1989.

(25) Zefirov, N. S.; Kirpichenok, M. A.; Izmailov, F. F.; Trofimov, M. I. Calculation Schemes for atomic electronegativities in molecular graphs within the framework of Sanderson Principle. *Dokl. Akad. SSSR* **1987**, *296*, 883−887.

(26) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.

(27) Stanton, D. T.; Egolf, L. M.; Jurs, P. C. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306−316.

(28) Jurs, P. C.; Chow, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, E. C., Christorffersen, R. E., Eds.; The American Chemical Society: Washington, DC, 1979; pp 103−129.

(29) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function;* Wiley: New York, 1979.

(30) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.

(31) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.

(32) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure−Activity Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.

(33) Wessel, M. D.; Jurs, P. C. Prediction of Reduced Ion Mobility Constants from Structural Information Using Multiple Linear Regression Analysis and Computational Neural Networks. *Anal. Chem.* **1994**, *66*, 2480−2487.

(34) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.

(35) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structure*; Wiley-Interscience: New York, 1983.

CI0498362