Available online at www.sciencedirect.com
SCIENCE DIRECT
Bioorganic & Medicinal Chemistry 14 (2006) 4987–5002
Bioorganic & Medicinal Chemistry

QSAR modeling of the inhibition of Glycogen Synthase Kinase-3

Alan R.Katritzky,[a,*] Liliana M. Pacureanu,[a] Dimitar A. Dobchev,[a,d] Dan C. Fara,[a] Pablo R. Duchowicz[b] and Mati Karelson[c,d]

[a]*Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, FL 32611, USA*
[b]*Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), Universidad Nacional de La Plata, La Plata 1900, Argentina*
[c]*Institute of Chemistry, Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia*
[d]*Department of Chemistry, University of Tartu, 2 Jakobi Street, Tartu 51014, Estonia*

Received 16 January 2006; revised 7 March 2006; accepted 7 March 2006
Available online 2 May 2006

Abstract—Quantitative structure–activity relationship (QSAR) models of the biological activity (pIC_{50}) of 277 inhibitors of Glycogen Synthase Kinase-3 (GSK-3) are developed using geometrical, topological, quantum mechanical, and electronic descriptors calculated by CODESSA PRO. The linear (multilinear regression) and nonlinear (artificial neural network) models obtained link the structures to their reported activity pIC_{50}. The results are discussed in the light of the main factors that influence the inhibitory activity of the GSK-3 enzyme.

1. Introduction

Approximately 500 protein kinases encoded in the human genome play essential roles in virtually all-known cellular processes and in almost all known diseases. These kinases catalyze the phosphorylation of various proteins involved in the mechanisms that regulate the metabolism and functioning of cells. The abnormal phosphorylation mediated by these kinases causes several illnesses. This has led to extensive screening for pharmacological inhibitors, that have been thoroughly reviewed.[1,2,3a–d]

Glycogen Synthase Kinase-3 (GSK-3) is a multifunctional serine/threonine kinase ubiquitously expressed in mammalian tissues. It is involved in multiple physiological processes including the Wnt and Hedgehog signaling pathways, cell cycle regulation, response to DNA damage, insulin action on glycogen synthesis, HIV-1 Tat-mediated neurotoxicity, hyperphosphorylation of tau (one of the diagnostic features of Alzheimer's disease) circadian rhythm, and others.[4,5] This set of events is related to cancers, type-2 diabetes, neurodegenerative disorders (Alzheimer, bipolar disorders), proliferation of protozoan parasites, and viral infections (HIV, cytomegalo virus, herpes virus).[6] Two human genes related to GSK-3, GSK-3α and GSK-3β, share a high homology in their binding site, but as demonstrated in previous reports[7,8] they are not functionally interchangeable.

Structurally diverse classes of GSK-3β inhibitor candidates, identified mainly in vitro, include lithium chloride,[9] indirubins,[10] paullones,[11] maleimides,[12] aloisines,[13] hymenialdisine,[14] etc. Seven compounds have recently been co-crystallized with GSK-3β and are localized within the ATP-binding pocket of the enzyme.[15] Kinetics experiments have shown that most inhibitors compete reversibly with ATP for binding to the kinase, with the exception of a few lithium or thiadiazolidinone compounds.[16,17]

A typical challenge during the synthesis of new kinase inhibitors relies on their degree of selectivity toward different types of kinases, that is to say, independent inhibitory activity. For instance, lithium chloride is not selective for GSK-3 since it also inhibits, for example, Casein Kinase 2. A high potency is also required, expressed as a low micromolar-range inhibitory activity. The synthesis of pharmacological-friendly drugs is

Keywords: Kinase; Inhibitory activity IC_{50}; Stepwise regression; Artificial neural network; CODESSA PRO.
* Corresponding author. Tel.: +1 352 392 0554; fax: +1 352 392 9199; e-mail: katritzky@chem.ufl.edu

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.03.009

also affected by different factors such as an increased solubility, optimal cell permeability, most favorable tissue, and intracellular distribution.

A large number of qualitative structure–activity relationships (SAR) have been reported[18–28] but only a few quantitative models based on molecular modeling and a three-dimensional QSAR approach[11,18,29] were developed. The purpose of the present work was to establish a QSPR model for the inhibition of Glycogen Synthase Kinase-3 that could serve as a guide for the rational design of further potent and selective inhibitors.

Comparative molecular similarity indices analysis (CoMSIA), on a training set consisting of 52 paullone derivatives and a test set of 23 compounds, provided a series of 3D-QSAR models.[11] The property modeled was pIC_{50} for GSK-3β (obtained from IC_{50} in μM units), and the descriptors employed were steric, electrostatic, hydrophobic, and hydrogen bond donor- and acceptor-fields. A partial least squares treatment achieved cross-validated results corresponding to five principal components: $R^2_{CV} = 0.554$, PRESS = 0.853 (0.140 μM). The authors also reported the conventional $R^2 = 0.871$ and $S = 0.458$ (0.348 μM).

Another interesting quantitative model[18] correlated the IC_{50} activities of indirubins with calculated interaction energies derived from molecular mechanics docking-scoring calculations, utilizing recent co-crystal structures of indirubin analogues with GSK-3. The method involved two main steps: (a) correlation-coupled receptor minimization and (b) unconstrained ligand relaxation/Monte Carlo search.[18] A mixture of both isoforms GSK-3α/β was analyzed. Based on the results for the main model in Ref. 18, a small set of new molecules were predicted and experimentally assessed as inhibitors. The authors concluded that the affinity of indirubins for GSK-3α/β depends mainly on the hydrophobic van der Waals energy term, which accounts for 66–92% of the sum of the three energy terms (VDW, electrostatics, and H-bonding).

Zeng et al.[29] investigated the inhibition of GSK-3 by aloisines. Two template conformations—the lowest energy and that extracted from the co-crystal structure—were used to check the influence of the spatial arrangement of the compounds in an approach that involved Comparative Molecular Similarity Index Analysis (CoMSIA) and Comparative Molecular Field Analysis (CoMFA) techniques. CoMSIA provided the best QSAR model for the higher energy conformation with $R^2 = 0.938$ and $q^2 = 0.673$, involving steric, electrostatic, and hydrophobic descriptors. They[29] concluded that the biologically active conformation did not necessarily have the lowest energy because of the confinement of kinase residues in the binding pocket. In order to provide some insight into the structure–activity relationship, they superimposed GSK-3 into the co-crystal structure of aloisine-CDK2 and concluded that the same factors influence the inhibitory activity.

A recent 3D-QSAR (CoMFA method) investigation[30] of 3-anilino-4-arylmaleimides and the available co-crystal structure with GSK-3β allowed a comparison between 3D-QSAR results and experimental intermolecular interactions. The 3D-QSAR results led to the characterization of the active site and gave insight into the essential features of the ligand–receptor interactions. The statistical results provided by a six-component regression equation are: $R^2 = 0.891$, $q^2 = 0.805$, $s = 0.146$.[30]

Methodology for a general QSAR/QSPR approach has been developed and coded as the CODESSA PRO software package. CODESSA PRO enables the calculation of numerous quantitative descriptors solely on the basis of molecular structural information (Hansch-type approach).[31,32] Research using CODESSA PRO has successfully correlated and predicted various physical properties[33] including gas chromatographic properties,[34a] melting and boiling points,[34b] solvent scales, and refractive indexes.[34c] Recent examples include QSPR treatments of (i) the binding energies for 1:1 complexation systems between various organic guest molecules and β-cyclodextrin,[35] (ii) the in vitro minimum inhibitory concentration (MIC) of 3-aryloxazolidin-2-one antibacterials to inhibit growth of *Staphylococcus aureus*,[36] (iii) partition coefficients of drugs between human breast milk and plasma,[37] and (iv) investigations of platelet-derived growth factor inhibition.[38]

Present work attempts to provide comprehensive QSAR models for the GSK-3 inhibitory activity of compounds. First, we define the data set employed and the basic calculation strategy. Second, we discuss the results obtained. Finally, a summary of the conclusions and suggestions for future extensions of the current treatment are provided.

2. Data set

The whole data set consists of 277 experimental IC_{50} values for GSK-3α, GSK-3β, and mixtures of both isoforms GSK-3α/β, determined from dose–response curves.[9–28,39] The data points were converted into molar units of pIC_{50} values and were then used instead of IC_{50}, in order to improve the normal distribution of the experimental data points.

All the data are collected in Table 1, including the following information: (i) the CAS number for the 277 compounds (second column), (ii) IC_{50} values taken from the original references and converted into molar units (third column), (iii) experimental pIC_{50} (fourth column), pIC_{50} values calculated from the multilinear models (fifth column), and predicted pIC_{50} values from the neural network model (sixth column).

The experimental data points were divided by classes of compounds into four subsets as follows:

Table 1. Experimental values of IC_{50}[a] and calculated pIC_{50}[b] for the full data set of inhibitors of GSK-3

Compound	CAS number	IC_{50}[a] $\times 10^{-9}$ (M)	pIC_{50}[b]		
			Experimental	Predicted	
				Multilinear models	ANN model
Class I					
1	101291-07-0	529	6.277	6.133	6.422[d]
2	264207-26-3	301	6.521	6.357	6.628[d]
3	264214-83-7	704	6.152	6.029	6.352[d]
4	264214-79-1	149	6.827	6.821	7.153[d]
5	264214-82-6	291	6.536	6.707	6.669
6	264214-81-5	143	6.845	6.707	6.669
7	264206-86-2	404	6.394	6.417	6.150
8	264208-87-9	2613	5.583	5.579	5.595
9	264213-20-9	216	6.666	6.673	6.590
10	264210-58-4	195	6.710	6.764	6.8120[d]
11	264214-76-8	374	6.427	6.372	6.828
12	264216-52-6	152	6.818	6.851	6.865
13	264214-72-4	93	7.032	7.208	6.826
14	264214-75-7	136	6.866	6.864	6.842
15	264214-74-6	74	7.131	7.139	7.469[d]
16	264210-60-8	161	6.793	6.598	6.623[d]
17	264222-25-5	337	6.472	6.569	6.182[d]
18	264213-24-3	216	6.666	6.683	6.633
19	264210-49-3	114	6.943	6.759	6.777
20	264213-50-5	259	6.587	6.378	6.749
21	264213-63-0	139	6.857	6.936	6.635
22	264213-72-1	82	7.086	7.174	6.837
23	264208-22-2	110	6.959	6.956	7.765[d]
24	264222-23-3	187	6.728	6.574	6.834
25	264214-12-2	104	6.983	7.050	6.751
26	264218-33-9	251	6.600	6.701	6.838
27	264218-23-7	104	6.983	7.108	7.254[d]
28	264216-07-1	52	7.284	7.416	7.230[d]
29	264217-67-6	28	7.553	7.558	7.455[d]
30	264222-45-9	131	6.883	6.764	6.656
31	264217-40-5	1478	5.830	6.254	6.825
32	264217-28-9	94	7.027	6.939	7.152[d]
33	264215-79-4	58	7.237	7.159	6.886
34	264215-80-7	134	6.873	6.797	7.221[d]
35	264217-24-5	76	7.119	7.060	6.858
36	264210-27-7	532	6.274	6.455	6.461
37	264222-36-8	460	6.337	6.135	6.434
38	264210-48-2	257	6.590	6.475	6.719
39	264215-20-5	472	6.326	6.156	6.791
40	264215-16-9	142	6.848	6.833	6.820
41	264215-19-2	195	6.710	6.722	6.890[d]
42	264215-18-1	85	7.071	6.963	7.175[d]
43	264207-93-4	203	6.693	6.666	6.791
44	264213-31-2	141	6.851	6.864	6.439
45	264209-34-9	70	7.155	6.908	6.634
46	264209-39-4	236	6.627	6.604	6.949[d]
47	264213-38-9	123	6.910	6.890	6.848
48	264213-25-4	59	7.229	7.138	6.860
49	264211-21-4	20	7.699	7.441	7.485[d]
50	264213-05-0	79	7.102	7.161	7.372[d]
51	264214-59-7	26	7.585	7.455	6.866
52	264209-30-5	152	6.818	6.794	6.859[d]
53	264208-99-3	1398	5.854	5.838	5.621
54	264222-21-1	161	6.793	6.641	6.698
55	264207-11-6	514	6.289	6.309	6.401
56	264207-01-4	447	6.350	6.565	6.533
57	264208-84-6	407	6.390	6.208	6.814
58	264209-73-6	317	6.499	6.492	6.832
59	264216-61-7	173	6.762	6.729	6.839
60	264209-67-8	91	7.041	7.029	7.409[d]
61	264211-44-1	186	6.730	6.853	6.729
62	264211-48-5	109	6.963	7.133	6.846

(*continued on next page*)

Table 1 (*continued*)

Compound	CAS number	IC_{50}[a] $\times 10^{-9}$ (M)	pIC_{50}[b]		
			Experimental	Predicted	
				Multilinear models	ANN model
63	264207-08-1	529	6.277	6.449	6.486
64	264208-93-7	2285	5.641	5.848	6.473
65	264222-01-7	1412	5.850	6.179	6.164[d]
66	264206-99-7	390	6.409	6.409	6.603
67	264207-22-9	156	6.807	6.812	6.967[d]
68	264215-15-8	481	6.318	6.245	6.783
69	264215-12-5	83	7.081	6.962	6.771
70	264215-14-7	214	6.670	6.912	6.674
71	264206-97-5	243	6.614	6.706	6.782
72	264222-11-9	694	6.159	6.280	6.358[d]
73	264215-97-6	71	7.149	7.273	6.887
74	264211-18-9	392	6.407	6.539	6.543
	Range I	20-2613	5.583–7.699	5.579–7.558	5.595–7.765
Class II					
75	551919-61-0	19	7.721	7.567	7.468
76	748142-08-7	6000	5.222	7.148	6.005
77	551920-54-8	10	8.000	7.480	7.925
78	748142-06-5	31	7.509	6.857	7.016
79	681432-33-7	125	6.903	7.012	7.166
80	681432-36-0	50	7.301	6.821	7.019
81	748142-09-8	12	7.921	7.995	8.024
82	681432-47-3	10	8.000	7.882	7.763
83	681432-38-2	10	8.000	7.618	7.530[d]
84	681432-32-6	12	7.921	7.211	7.796
85	748142-07-6	16	7.796	8.321	7.960
86	748141-84-6	50	7.301	7.151	7.431[d]
87	748141-85-7	199	6.701	6.062	6.664
88	748141-88-0	794	6.100	6.177	6.294
89	748142-10-1	100	7.000	7.295	7.769
90	748142-11-2	50	7.301	7.360	7.448[d]
91	748142-12-3	50	7.301	7.565	7.383[d]
92	748142-13-4	79	7.102	7.151	7.653
93	748142-14-5	125	6.903	6.948	6.372
94	748142-15-6	199	6.701	6.692	6.549
95	748142-16-7	3162	5.500	6.028	6.779[d]
96	748142-17-8	630	6.201	5.565	6.426
97	748142-18-9	39	7.409	7.747	7.420[d]
98	748142-19-0	50	7.301	7.912	6.853
99	748142-20-3	10	8.000	6.964	7.836[d]
100	748142-21-4	10	8.000	8.437	7.803
101	748141-95-9	158	6.801	6.356	6.675
102	748142-22-5	125	6.903	6.640	7.051
103	681432-49-5	10	8.000	7.438	6.686
104	681432-52-0	12	7.921	7.348	6.722
105	681432-56-4	10	8.000	7.659	6.727[d]
106	681432-55-3	11	7.959	7.566	7.879
107	681432-53-1	20	7.699	7.476	6.598
108	681432-57-5	20	7.699	6.932	7.802
109	681432-61-1	20	7.699	7.646	7.425
110	681432-63-3	25	7.602	6.963	7.597
111	681432-62-2	40	7.398	6.661	7.458
112	681432-60-0	10	8.000	7.086	7.749[d]
113	681432-02-0	5010	5.300	5.856	5.463
114	681432-05-3	31600	4.500	5.267	4.539
115	681432-06-4	7940	5.100	5.335	4.818
116	681432-10-0	31000	4.509	5.230	4.705
117	681432-07-5	5010	5.300	5.874	5.037
118	681432-65-5	158	6.801	6.794	6.605
119	681432-70-2	316	6.500	6.429	6.174
120	681432-67-7	50	7.301	6.820	6.689
121	681432-69-9	1000	6.000	6.293	6.020[d]
122	681432-20-2	1250	5.903	7.141	6.618

Table 1 (*continued*)

Compound	CAS number	IC_{50}[a] $\times 10^{-9}$ (M)	pIC_{50}[b]		
			Experimental	Predicted	
				Multilinear models	ANN model
123	681432-23-5	2510	5.600	6.522	6.546
124	681432-24-6	1000	6.000	6.519	6.046
125	681432-26-8	3980	5.400	6.142	5.922[d]
126	681432-25-7	796	6.099	6.826	6.349
127	681432-71-3	50	7.301	8.152	7.449[d]
128	681432-75-7	40	7.398	7.093	7.248[d]
129	681432-77-9	31	7.509	6.661	7.122
130	681432-76-8	100	7.000	7.065	7.093
131	681432-74-6	32	7.495	7.904	7.684
132	681432-78-0	794	6.100	6.184	6.733
133	681432-83-7	1580	5.801	5.981	5.881
134	681432-82-6	630.	6.201	6.135	6.239[d]
135	681432-84-8	158	6.801	6.888	7.215[d]
136	681432-85-9	25	7.602	7.182	7.599[d]
137	681432-89-3	25	7.602	6.900	8.329[d]
138	681432-91-7	16	7.796	7.737	8.220
139	681432-88-2	16	7.796	7.605	7.380
140	681432-90-6	50	7.301	7.149	7.743
141	681432-92-8	19000	4.721	4.626	4.576
142	681432-96-2	23000	4.638	4.703	6.860[d]
143	748142-01-0	80	7.097	7.550	7.186
144	548797-12-2	99	7.004	6.538	6.369
145	405222-59-5	4	8.398	8.015	7.411[d]
146	439290-41-2	7	8.155	7.365	6.482
147	557113-38-9	2697	5.569	6.873	5.441
148	557113-39-0	691	6.161	7.240	6.840
149	405222-60-8	22	7.658	7.425	7.750
150	405223-00-9	11	7.959	8.884	7.627
151	405223-04-3	7	8.155	7.742	7.892
152	405222-61-9	5	8.301	8.880	7.859
153	405222-94-8	9	8.046	7.089	8.262
154	405224-05-7	5	8.301	8.198	8.326[d]
155	405222-72-2	5	8.301	7.727	7.944
	Range II	4-31600	4.500–8.398	4.626–8.880	4.539–8.326
Class III					
156	583038-29-3	75	7.125	7.336	6.678
157	583038-60-2	0.80	9.097	8.803	8.770[d]
158	583038-34-0	8	8.097	7.759	8.518
159	583038-56-6	5	8.301	8.487	8.532
160	583038-54-4	7	8.155	8.649	8.124[d]
161	583038-58-8	24	7.620	7.746	7.580
162	583038-71-5	4	8.398	8.178	8.243[d]
163	583038-40-8	12	7.921	7.816	7.526
164	548797-19-9	498	6.303	6.435	6.571[d]
165	548797-18-8	15	7.824	7.997	7.961[d]
166	548797-15-5	42	7.377	6.364	6.932
167	548797-37-1	481	6.318	6.311	6.550[d]
168	548797-26-8	828	6.082	6.898	6.247
169	548797-33-7	320	6.495	6.396	6.825
170	548797-32-6	50	7.301	6.650	7.519
171	548797-34-8	35	7.456	6.896	6.847
172	548797-14-4	215	6.668	6.446	6.902
173	548797-17-7	329	6.483	6.657	6.943
174	583038-96-4	39	7.409	7.626	7.589
175	583039-51-4	7	8.155	7.791	7.853
176	583039-55-8	141	6.851	6.825	6.916
177	583039-27-4	7	8.155	7.732	8.051
178	583039-39-8	99	7.004	7.353	7.117
179	583039-44-5	16	7.796	7.631	8.201[d]
180	583039-25-2	18	7.745	7.210	7.777
181	583039-36-5	14	6.602	6.451	6.658

(*continued on next page*)

Table 1 (*continued*)

Compound	CAS number	IC_{50}[a] $\times 10^{-9}$ (M)	pIC_{50}[b]		
			Experimental	Predicted	
				Multilinear models	ANN model
182	107042-54-6	250	6.276	6.459	6.520
183	405224-27-3	530	6.367	5.828	6.122[d]
184	405224-21-7	430	5.900	6.243	6.016[d]
185	439290-93-4	1260	6.536	6.809	7.310[d]
186	405221-12-7	291	7.367	6.839	7.280[d]
187	405221-13-8	43	7.252	7.088	7.168[d]
188	405221-08-1	56	7.721	7.246	7.598[d]
189	405221-32-1	19	8.301	7.342	7.372[d]
190	405221-87-6	5	5.551	5.870	7.329[d]
191	557112-45-5	2810	5.447	5.446	5.973[d]
192	557112-46-6	3572	6.449	6.743	6.794
193	557112-47-7	356	5.630	6.996	6.107
194	557112-48-8	2343	7.745	7.417	7.311[d]
195	405221-39-8	18	7.699	7.631	7.488[d]
196	405221-48-9	20	8.155	7.858	6.780
197	405221-61-6	7	7.569	7.725	7.318[d]
198	405221-38-7	27	7.959	7.318	7.836
199	405221-09-2	11	6.354	7.333	6.399
200	405221-67-2	443	6.070	6.854	7.220[d]
201	405221-69-4	851	6.772	7.242	7.455[d]
202	557112-49-9	169	6.618	6.919	6.666
203	405221-70-7	241	6.372	6.099	6.541
204	583038-30-6	425	8.097	8.345	8.092
205	583038-28-2	8	6.903	6.689	6.649
206	583038-36-2	125	7.444	7.684	7.663
207	583038-42-0	36	5.798	6.218	5.943
208	583038-38-4	1593	9.000	8.336	8.346
209	583038-63-5	1	8.222	7.964	8.150
210	583038-51-1	6	6.382	6.248	6.953[d]
211	583038-33-9	415	6.630	6.906	6.659
212	583038-93-1	234	7.060	7.269	7.457
213	583039-16-1	87	6.417	7.244	6.728
214	583038-84-0	383	7.921	8.094	8.515
215	583038-46-4	12	9.000	9.142	8.547
216	583038-76-0	1	7.678	7.514	8.269[d]
217	583038-48-6	21	7.125	7.336	6.678
	Range III	1-3572	5.447–9.097	5.446–9.142	5.943–8.770
Class IV[c]					
218	319490-29-4		7.000	6.539	7.438
219	710947-39-0		5.500	6.713	5.795
220	710947-40-3		5.600	6.072	6.434[d]
221	650837-82-4		7.000	6.811	7.173[d]
222	710947-42-5		6.800	6.224	6.675
223	710947-43-6		6.800	6.173	6.959
224	650626-35-0		5.600	6.305	5.718
225	650626-94-1		8.200	7.742	6.827
226	650626-57-6		6.000	7.030	7.061
227	650626-54-3		7.000	6.407	7.609
228	650626-56-5		6.500	6.636	6.347
229	650626-52-1		5.400	5.953	5.911[d]
230	650626-51-0		5.700	6.514	5.950
231	710947-45-8		6.800	6.567	6.749
232	710947-46-9		6.200	7.101	6.264[d]
233	650626-79-2		6.500	7.200	5.693
234	650626-27-0		6.600	7.212	6.791
235	650637-59-5		5.900	6.859	6.632[d]
236	710947-48-1		7.500	7.291	6.609[d]
237	650637-83-5		8.500	7.525	8.443
238	650637-35-7		7.400	6.785	7.334
239	650637-81-3		7.400	6.515	7.330
240	650627-96-6		8.000	7.178	7.940
241	650627-94-4		7.600	7.573	7.548[d]

Table 1 (*continued*)

Compound	CAS number	IC_{50}[a] $\times 10^{-9}$ (M)	pIC_{50}[b]		
			Experimental	Predicted	
				Multilinear models	ANN model
242	650627-93-3		7.500	7.640	6.945[d]
243	650627-92-2		6.900	7.679	7.876
244	650627-66-0		8.100	7.737	7.921
245	650627-68-2		8.100	7.863	7.900
246	650627-67-1		7.800	7.881	7.629
247	650627-43-3		8.000	7.939	7.571
248	650627-36-4		8.600	7.936	7.770
249	650627-89-7		8.400	8.652	8.415
250	650627-42-2		8.600	8.347	8.271[d]
251	650627-79-5		7.900	7.685	7.535
252	650627-77-3		6.800	7.098	6.716
253	650627-80-8		5.600	6.093	5.389
254	650627-65-9		8.200	7.787	8.419
255	650627-54-6		5.600	7.552	5.622
256	650627-27-3		8.300	8.314	8.110[d]
257	710947-51-6		8.400	7.520	8.230
258	710947-52-7		7.900	7.499	8.248[d]
259	710947-53-8		7.800	7.592	7.693
260	710947-54-9		8.300	8.414	8.091[d]
261	650627-81-9		7.800	7.674	8.026[d]
262	706809-13-4		7.500	7.918	7.173
263	706809-12-3		7.800	8.442	7.747
264	710947-56-1		8.100	8.470	8.235
265	710947-57-2		8.000	7.524	8.261[d]
266	710947-58-3		8.000	7.064	7.960[d]
267	650628-06-1		7.900	7.051	8.067[d]
268	650628-03-8		8.000	7.001	8.112[d]
269	650627-98-8		7.500	6.640	7.666
270	710947-59-4		7.200	7.121	8.096[d]
271	710947-60-7		5.200	7.265	5.842
272	650840-91-8		6.200	7.015	6.350[d]
273	706809-14-5		7.700	7.017	7.804
274	650637-85-7		8.400	7.692	8.469
275	650637-86-8		8.200	7.877	7.902[d]
276	706809-15-6		8.800	8.317	8.698
277	650637-87-9		8.300	8.871	8.474[d]
278	706809-16-7		8.300	8.089	8.349[d]
	Range IV		5.200–8.800	5.953–8.871	5.389–8.698
	Range FS	1-31.600	4.500–9.097	4.626–9.142	4.539–8.770

[a] Taken from the original reference as follows: *Class I* from 12; *Class II* from 20 (compounds **75**–**143**) and 26 (compounds **144**–**155**]; *Class III* from 23 (compounds **156**–**181**), 24 (compounds **182**–**203**), and 25 (compounds **204**–**217**); and *Class IV* from 22 (compounds **218**–**271**) and 21 (compounds **272**–**278**).

[b] Calculated with $pIC_{50} = -\log(IC_{50})$.

[c] The original authors reported the pIC_{50} values and the IC_{50} values were not calculated.

[d] Validation set.

(i) *Class I*: $n = 74$, 3-anilino-4-aryl-maleimide derivatives.
(ii) *Class II*: $n = 81$, 5-aryl-pyrazolo[3,4-*b*]pyridazine and *N*-phenyl-4-pyrazolo[1,5-*b*]pyridazin-3-yl-pyrimidin-2-amine derivatives.
(iii) *Class III*: $n = 62$, 5(6)-aryl-pyrazolo[3,4-*b*]pyridine and 6-heteroaryl-pyrazolo[3,4-*b*]pyridine derivatives, and
(iv) *Class IV*: $n = 61$, [1-(1*H*-benzimidazol-7-yl)-1*H*-pyrazolo[3,4-*d*]pyrimidin-4-yl]arylhydrazone and [1-aryl-1*H*-pyrazolo[3,4-*d*]pyrimidin-4-yl]arylhydrazone derivatives (see Table 1). The templates of the four classes of GSK-3 inhibitors are shown in Figure 1.

3. Results and discussions

In the current study, we present Quantitative Structure–Activity Relationship (QSAR) models for pIC_{50} involving theoretical descriptors, which have been calculated solely from molecular structure. The results and

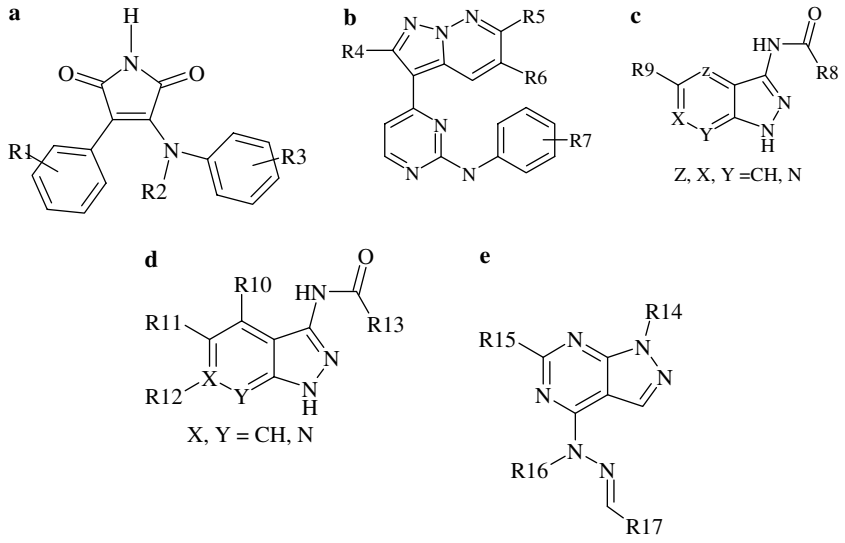

Figure 1. The templates of GSK-3 inhibitors: (a) *Class I*: 3-anilino-4-aryl-maleimide derivatives; (b and c) *Class II*: 5-aryl-pyrazolo[3,4-*b*]pyridazines and *N*-phenyl-4-pyrazolo[1,5-*b*]pyridazin-3-yl-pyrimidin-2-amine derivatives; (d) *Class III*: 5(6)-aryl-pyrazolo[3,4-*b*]pyridines and 6-heteroaryl-pyrazolo[3,4-*b*]pyridine derivatives; (e) *Class IV*: [1-(1*H*-benzimidazol-7-yl)-1*H*-pyrazolo[3,4-*d*]pyrimidin-4-yl]arylhydrazones and [1-aryl-1*H*-pyrazolo[3,4-*d*]pyrimidin-4-yl]arylhydrazones.

discussion from both approaches (i) multilinear and (ii) nonlinear are presented in the next two subsections.

3.1. Multilinear QSAR modeling

3.1.1. Class I. A preliminary selection of the molecules in *Class I* included 74 data points for 3-anilino-4-aryl-maleimides. There are three skeletons (templates) that can be used as general pattern for the compounds of *Class I* (maleimides). All the substitution patterns connected to these three skeletons are presented in Supplementary material SM-1.

The BMLR[40] procedure was used to obtain the best multilinear QSAR model for the inhibition of GSK-3 by 3-anilino-4-arylmaleimides.

To find the optimum number of descriptors describing pIC_{50} for the current set of organics, we analyzed multiparameter correlations containing up to 10 descriptors. Figure 3 shows the relationships of R^2 with the number of descriptors. As it can be seen from Figure 3, R^2 rises steeply as the number of parameters increases from two to ten and the breakpoint occurs at the sixth descriptor. Therefore, we used the best correlation equation with six descriptors, shown in Table 2, for the basic analysis.

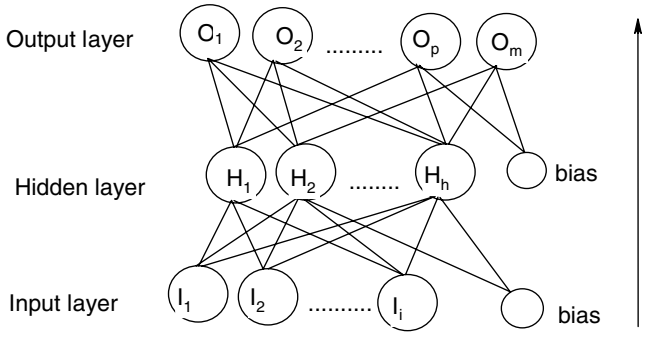

Figure 2. Three-layer back propagation neural network.

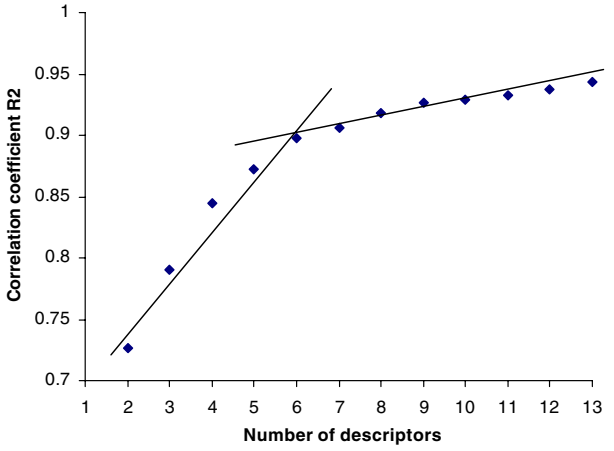

Figure 3. Correlation coefficient versus number of descriptors for pIC_{50}.

The R^2 for the six- and seven-parameter models are 0.896 and 0.904, respectively; $\Delta R^2 = 0.008 < 0.02$ (Fig. 3).

The QSAR equation for *Class I* is characterized by the following statistical parameters shown in Table 2: $N = 74$, $n = 6$, $R^2 = 0.896$, $R^2_{cv} = 0.874$, $F = 96.683$, $s^2 = 0.019$, where N is the number of data points; n is the number of descriptors; R^2 is the squared correlation coefficient; R^2_{cv} is the squared cross-validated correlation coefficient; F is the Fisher's criterion; and s^2 is the squared standard error. In Table 2 the notation is as follows: b is the regression coefficient of the linear model, Δb is the standard error for each regression coefficient,

Table 2. Selected molecular descriptors and statistical characteristics provided by the best QSAR models for *Classes I–IV*

Descriptor name	Symbol	*b*	Δ*b*	*t*	R^2	R^2_{cv}	s^2
Intercept	—	76.064	4.689	16.221	—	—	—
Randic index (order 1)	D_1	0.539	0.034	15.839	0.440	0.410	0.099
Average bond order for atom H	D_2	−89.158	6.410	−13.907	0.655	0.624	0.062
LUMO energy	D_3	1.803	0.156	11.510	0.755	0.727	0.044
HACA-2 (MOPAC PC)	D_4	−0.315	0.035	−8.863	0.822	0.795	0.032
Charged surface area (MOPAC PC) for atom N	D_5	0.703	0.105	6.681	0.871	0.845	0.024
Max coulombic interaction for bond C–C	D_6	1.755	0.436	4.024	0.896	0.874	0.019
Class II							
Intercept	—	44.352	23.914	1.854	—	—	—
Shadow plane ZX	D_7	−0.055	0.006	−8.145	0.170	0.129	0.890
Max e–n attraction for bond H–C	D_8	−1.906	0.254	−7.499	0.180	0.118	0.890
Max electroph. react. index for atom C	D_9	197.519	27.334	7.225	0.374	0.311	0.689
Max SIGMA–PI bond order	D_{10}	−20.261	3.243	−6.246	0.496	0.436	0.561
count of H-donor sites (MOPAC PC)	D_{11}	0.154	0.030	5.089	0.560	0.498	0.496
XY Shadow/XY Rectangle	D_{12}	−8.689	2.370	−3.665	0.635	0.567	0.417
Max exchange energy for bond H–C	D_{13}	18.463	5.996	3.078	0.677	0.593	0.374
Class III							
Intercept	—	95.451	27.757	3.438	—	—	—
HASA-1 (MOPAC PC) (all)	D_{14}	0.013	0.002	6.334	0.254	0.204	0.612
Max partial charge (Zefirov) for atom H	D_{15}	44.101	7.332	6.014	0.505	0.466	0.413
Min e–e repulsion for bond C–C	D_{16}	−0.129	0.022	−5.700	0.622	0.570	0.321
Min 1-electron react. index for atom C	D_{17}	−47.361	12.643	−3.745	0.707	0.660	0.253
Max atomic state energy for atom C	D_{18}	−0.759	0.264	−2.867	0.745	0.703	0.224
Class IV							
Intercept	—	120.565	28.880	4.174	—	—	—
Max exchange energy for bond H–C	D_{13}	−22.740	5.497	−4.136	0.126	0.070	0.867
Shadow plane YZ	D_{19}	−0.089	0.024	−3.736	0.135	0.025	0.873
Structural Information content (order 2)	D_{20}	0.105	0.034	−3.018	0.328	0.211	0.690
Tot hybridization comp. of the molecular dipole	D_{21}	0.433	0.151	2.868	0.439	0.322	0.586
Highest normal mode vib transition dipole	D_{22}	−0.109	0.047	−2.273	0.475	0.343	0.558
Max e–e repulsion for bond H–C	D_{23}	0.216	0.117	1.846	0.507	0.370	0.535

and *t* is the *t*-test values for each coefficient. Also, the R^2, R^2_{cv}, and s^2 values for each individual model consisting of the given descriptors and those listed above appear in Table 2.

The linear plot between observed versus predicted pIC_{50} is given in Figure 4 and illustrates the fit of the predicted values resulting from the best QSAR model.

For this model, two compounds were registered as outliers: entries 31 and 65 from Table 1. These compounds displayed lower values for the predicted pIC_{50} values. Moreover, indoline 65 is supposed to have a

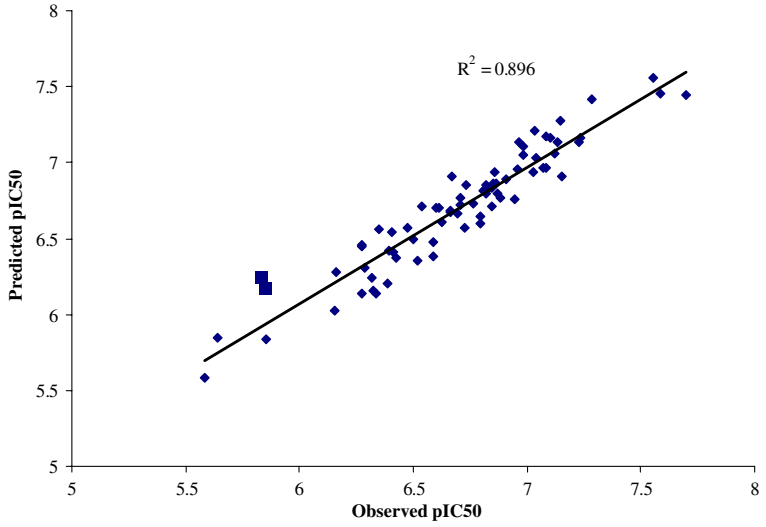

Figure 4. Plot of predicted versus observed pIC_{50} values for *Class I* of compounds according to the model in Table 3.

different binding mode with respect to *N*-methyl-3-anilino-4-arylmaleimides and 3-anilino-4-arylmaleimides.[12]

3.1.2. Class II. Figure 5 *Class II* consists of 81 5-arylpyrazolo[3,4-*b*]pyridazine and *N*-phenyl-4-pyrazolo[1,5-*b*]pyridazin-3-yl-pyrimidin-2-amine derivatives. The resulting QSAR model (Table 2) obtained by CODESSA PRO for this class had the following statistical characteristics: $R^2 = 0.677$, $R^2_{cv} = 0.593$, $F = 21.900$, $s^2 = 0.374$, $n = 7$, $N = 81$. The linear fit between the predicted and experimental pIC_{50} values according to the model in Table 2 is presented graphically in Figure 6.

The number of significant descriptors was defined by using the break point in Figure 5, which shows the number of descriptors versus R^2 for the equations obtained by the BMLR procedure.

As can be seen from Table 2, the multilinear model for *Class I* is statistically better than the model for *Class II*. However, the number of compounds involved in model of *Class II* is larger than in *Class I*. The poorer statistical characteristics of the *Class II* model indicate that these compounds are more difficult to model.

3.1.3. Class III. The best QSAR for *Class III* of compounds, which consists of 61 5(6)-aryl-pyrazolo[3,4-*b*]-pyridine and 6-heteroarylpyrazolo[3,4-*b*]pyridine derivatives (one chiral compound was excluded from this data set), involves a five-parameter model (Table 2) with $N = 61$, $n = 5$, $R^2 = 0.745$, $R^2_{cv} = 0.703$, $F = 32.240$. $s^2 = 0.224$. The linear fit between the experimental and predicted pIC_{50} is shown in Figure 8. The optimal number of descriptors for the equation in Table 2 was again defined by the break point rule in Figure 7.

A comparison between the models of *Classes II* and *III* shows that the latter is somewhat better in terms of the statistical parameters. The number of compounds used for these models is different as well as the number of descriptors involved.

3.1.4. Class IV. *Class IV* is represented by 61 [1-(1*H*-benzimidazol-7-yl)-1*H*-pyrazolo[3,4-*d*]pyrimidin-4-yl]-arylhydrazone and [1-aryl-1*H*-pyrazolo[3,4-*d*]pyrimidin-4-yl]arylhydrazone derivatives. The best QSAR model obtained for *Class IV* is characterized by: $N = 6$, $n = 61$, $R^2 = 0.507$, $R^2_{cv} = 0.370$, $F = 9.245$, $s^2 = 0.535$. Further, we defined the number of descriptors for this

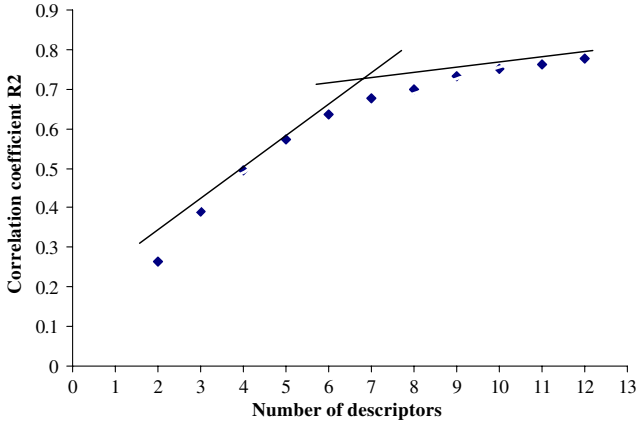

Figure 5. Plot of correlation coefficient R^2 versus number of descriptors for *Class II* of compounds.

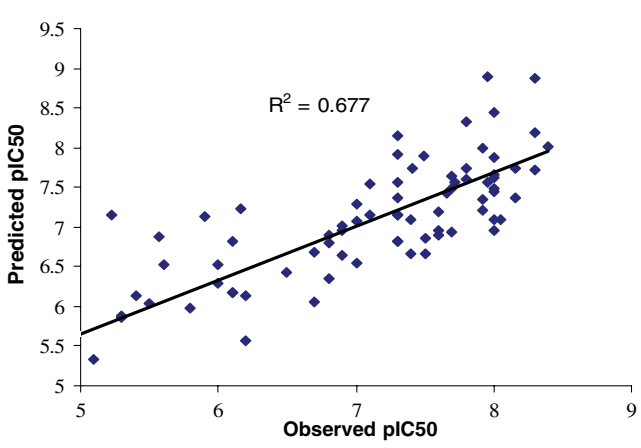

Figure 6. Plot of predicted versus observed pIC_{50} values for *Class II* of compounds using the seven-parameter model.

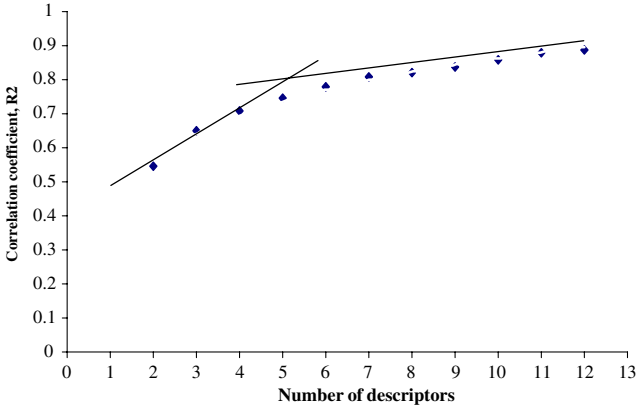

Figure 7. Plot of correlation coefficient versus number of descriptors for *Class III*.

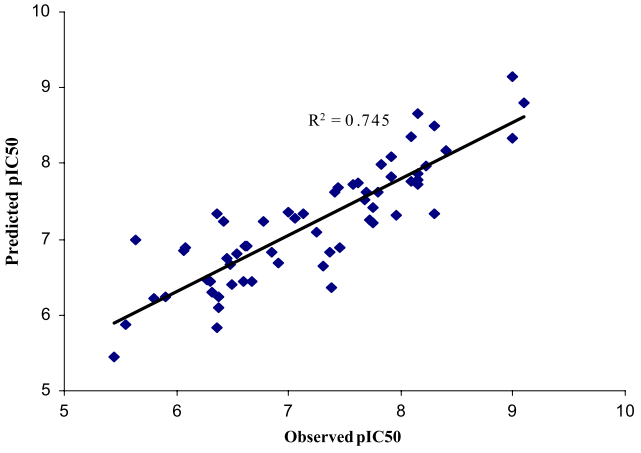

Figure 8. Plot of predicted versus observed pIC_{50} values for *Class III* according to model in Table 4.

model by the break point, showing no significant improvement of R^2.

The model for *Class IV* (Table 2) is poorer according to its statistical criteria. Obviously, modeling the 61 [1-(1*H*-benzimidazol-7-yl)-1*H*-pyrazolo[3,4-*d*]pyrimidin-4-yl]-arylhydrazone and [1-aryl-1*H*-pyrazolo[3,4-*d*]pyrimidin-4-yl]arylhydrazone derivatives is not an easy task. The authors present this model for completeness and with the purpose of discussing its descriptors, since they seem to be similar to the ones appearing in the previous classes.

3.1.5. Validation of the multilinear QSAR models

3.1.5.1. Leave-one-out. The first technique applied for the validation of the proposed multilinear QSAR models was based on the leave-one-out algorithm. The corresponding squared cross-validated correlation coefficient (R^2_{cv}) for all selected models, which is calculated automatically by the validation module implemented in CODESSA PRO[41] package, is listed in Table 3. For a better comparison, the squared correlation coefficient of the model is also given.

3.1.5.2. Internal validation. As mentioned in Section 5, our internal validation predicts the property values for each one-third of the compounds with the model fitted for the remaining two-third of the compounds. This procedure was applied for *Classes I–IV* of compounds by using the corresponding QSAR models (see Table 4).

The general algorithm of internal validation involves the following steps:

(i) division of the data set to be analyzed into three sets: C_{iA} (the 1st, 4th, 7th, etc. entries), C_{iB} (the 2nd, 5th, 8th, etc., entries) and C_{iC} (the 3rd, 6th, 9th, etc., entries);
(ii) in each of three combinations, two of the sets are combined into one and the correlation equation with the same descriptors, as in the QSAR model to be validated, is derived;
(iii) the equation developed in step (ii) is used to predict the property values for the remaining set;
(iv) a comparison of the average of squared correlation coefficients for the fitted and predicted sets is made at the end.

The results of the internal validation applied to our data are listed in Table 4.

3.2. Nonlinear QSAR modeling

In this study, we used the ANN methodology for prediction of the pIC_{50} values. Thus, it was possible to build a general nonlinear QSAR model based on all the experimental data. To do this, all the experimental data (277 data points) for pIC_{50} were divided into training (187) and validation subsets (90). In the first stage, before the neural network treatment started, both experimental pIC_{50} and descriptor values were normalized to a range 0–0.9 (see Eq. 4). The next stage of the ANN modeling is the selection of the most significant descriptors from the large descriptor pool. Thus, this descriptor pool (consisting of 961 descriptors) was reduced by the following procedures:

(i) descriptors with both high intercorrelations ($R^2 > 0.6$) and at probability level $p < 0.05$, thus 722 descriptors were excluded.
(ii) descriptors with small variance ratio $\sigma/|d_{max} - d_{min}| < 1e{-}006$ were excluded (1 2 3);
(iii) descriptors for which no values were available for all structures were excluded (83). Thus, the descriptor pool was reduced to 33 descriptors;
(iv) from this reduced pool of descriptors, we rejected another 21 descriptors since they showed random variations by exploring the scatter plots between the property and the corresponding descriptor.

Thus, the final descriptor pool was reduced to 12 descriptors for which sensitivity-stepwise analysis was performed by building the ANN models with simple 1-1-1 architecture for each relevant descriptor. Those descriptors that showed the lowest prediction error at the ANN output were chosen for building the optimum ANN model. Finally, six descriptors were found to be significant for building the ANN model.

During the training stage the weights were adjusted according to the output prediction error by using the

Table 3. Leave-one-out validation of the proposed QSAR models

Class	R^2	R^2_{cv}	$\Delta R^2 = R^2 - R^2_{cv}$
I	0.896	0.876	0.020
II	0.677	0.593	0.084
III	0.745	0.703	0.042
IV	0.507	0.370	0.137

Table 4. Internal validation of the models—statistical characteristics

Set to fit	R^2_{fit}	s^2_{fit}	Set to predict	$R^2_{predict}$	$s^2_{predicted}$
Class I					
$C_{1A}+C_{1B}$	0.901	0.020	C_{1C}	0.881	0.023
$C_{1A}+C_{1C}$	0.887	0.021	C_{1B}	0.886	0.024
$C_{1B}+C_{1C}$	0.915	0.016	C_{1A}	0.853	0.029
Average	0.901	0.019		0.873	0.025
Class II					
$C_{2A}+C_{2B}$	0.674	0.415	C_{2C}	0.659	0.362
$C_{2A}+C_{2C}$	0.694	0.364	C_{2B}	0.584	0.573
$C_{2B}+C_{2C}$	0.736	0.318	C_{2A}	0.537	0.627
Average	0.701	0.365		0.593	0.520
Class III					
$C_{3A}+C_{3B}$	0.691	0.295	C_{3C}	0.858	0.116
$C_{3A}+C_{3C}$	0.763	0.191	C_{3B}	0.721	0.324
$C_{3B}+C_{3C}$	0.791	0.214	C_{3A}	0.593	0.287
Average	0.748	0.233		0.724	0.242
Class IV					
$C_{3A}+C_{3B}$	0.451	0.511	C_{3C}	0.580	0.681
$C_{3A}+C_{3C}$	0.582	0.518	C_{3B}	0.246	0.709
$C_{3B}+C_{3C}$	0.519	0.620	C_{3A}	0.443	0.531
Average	0.517	0.549		0.423	0.640

backpropagation algorithm. The validation set error (and also R^2) was monitored in order to avoid the over-training of the ANN and to stop the training process.

We found that a six-descriptor model (6-6-6-1) was appropriate for the pIC_{50} property. The root-mean-squared (rms) error for the training and validation data is 0.67 and 1.54, respectively. In addition, an exploration of the standard deviations of the neural network models with different numbers of hidden units was performed. The seven-descriptor models (7-6-6-1) did not give significant improvement over the six-descriptor models (rms = 1.11). The same result was found for the 5-6-6-1 models with increased hidden units (rms = 0.95). The predicted values of pIC_{50} obtained are given in Table 1. Graphical presentation as a linear fit for the training set is given in Figure 9.

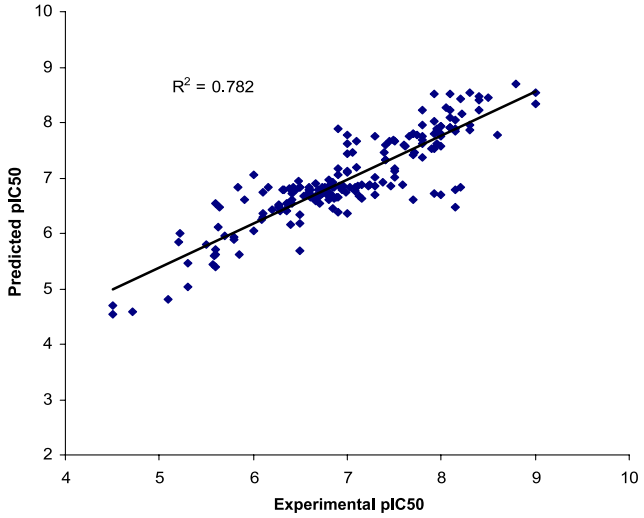

Figure 9. Plot of predicted versus observed pIC_{50} values for the training set.

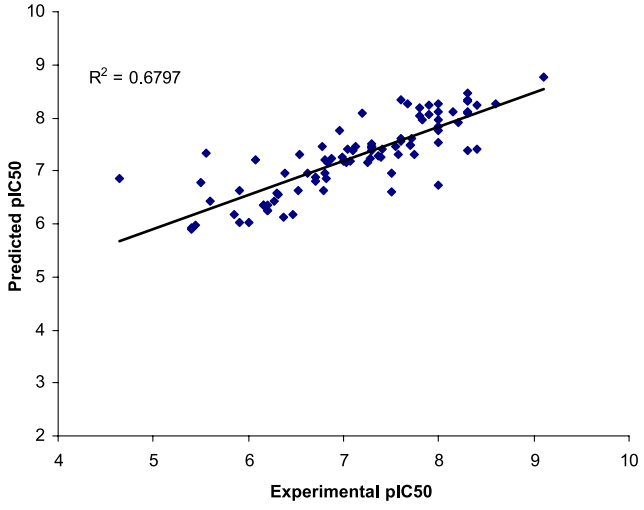

Figure 10. Plot of predicted versus observed pIC_{50} values for the validation set (90).

The maximum squared correlation coefficient for the training set was $R^2 = 0.782$ for 187 experimental data points. The corresponding validation set (which was not used to train the ANN model) had a maximum $R^2 = 0.679$ at which the training of the network was stopped. Figure 10 shows the linear fit between the experimental and predicted pIC_{50} values for the validation set.

As can be seen from Table 1 and Figures 9 and 10, the ANN model had superiority over the multilinear QSAR models developed in the previous section. This conjecture is supported by the statistical (i.e., R^2 and the number of compounds) characteristics of the two types of models.

The ANN model included the following descriptors used as inputs: *count of H-acceptor sites (Zefirov)*, *HA dependent HDSA-1 (Zefirov)*, *Kier and Hall index, Minimum partial charge for all atom types*, *RNCG Relative negative charged SA (SAMNEG/RNCG) (MOPAC PC), and Shadow plane ZX*. Most of these descriptors are charge-related descriptors.

3.3. Discussion of the descriptors

The six types of descriptors involved in equation for *Class I* (Table 2) are: topological (D_1), molecular orbital related (D_2, D_3), quantum chemical (D_6), and electrostatic (D_4, D_5). The *t*-test indicated the following order of significance for the descriptors included in the equation for *Class I* (Table 2): $D_1 > D_2 > D_3 > D_4 > D_5 > D_6$.

The direct interpretation of the descriptors appearing in the model from Table 2, nevertheless, is rather difficult, considering the complex nature of the GSK-3 inhibition processes. However, some indirect links between those descriptors and the physico-chemical phenomena behind the inhibitor-interactions with the GSK-3 enzyme are suggested.

The most statistically significant descriptor is the Randic index (order 1) (D_1) that has a topological origin and is a measure of the compactness of the molecule.[42] In Eq. 1, D_i, D_j are the edges of the adjacent atoms *i* and *j*, and the summation is performed over all pairs of edges *i* and *j* of the molecule. The magnitude of the drug–receptor interaction is directly related to the degree of branching as well as to the molecular size. In Table 2, this descriptor has a large positive influence.

$$^{m}\chi = \sum_{\text{path}} \left(D_i D_j \ldots D_k\right)^{-1/2} \qquad (1)$$

According to the *t*-test values, the second important descriptor is the average bond order for atom H. Bond orders can be related to the bonding nature σ/π. They are defined as the covariance of the electron population of two atoms and are related to the total atomic valences and atomic charges.[43–46] The presence of this descriptor in the QSAR model (Table 2) suggests the hydrogen bonding ability of the inhibitor and emphasizes the electrostatic interactions GSK-3-inhibitor. LUMO energy is

usually approximated with electron affinity and characterizes the susceptibility of the molecule toward the attack by nucleophilic reactants. The maximum coulombic interaction for bond C–C, descriptor (D_6) accounts for the sum of the electronic repulsion energy, electron–nuclear attraction energy, and nuclear repulsion energy between N and H atoms (Eq. 2).[41] In the equation from Table 2, descriptor D_6 has a small positive influence. The additive energy between two atoms can be expressed as:

$$E(\mathrm{AB}) = E_{\mathrm{ee}}(\mathrm{AB}) + E_{\mathrm{ne}}(\mathrm{AB}) + E_{\mathrm{nn}}(\mathrm{AB}), \qquad (2)$$

where $E_{\mathrm{ee}}(\mathrm{AB})$ is the electronic repulsion energy between two atoms, $E_{\mathrm{ne}}(\mathrm{AB})$ is the electron–nuclear attraction energy between two atoms, and $E_{\mathrm{nn}}(\mathrm{AB})$ is the nuclear repulsion energy between two atoms.

Electrostatic descriptors HACA-2 (MOPAC PC) and charged surface area for atom N describe the electrostatic component of the inhibitor—GSK-3 interaction.

$$\mathrm{HACA} = \sum_A \frac{q_A\sqrt{S_A}}{\sqrt{S_{\mathrm{tot}}}} \qquad (3)$$

where S_{A} represents the solvent accessible surface area of hydrogen-bonding acceptor atoms, selected by threshold charge, q_{A} is the partial charge on hydrogen-bonding acceptor atoms selected by threshold charge, and S_{tot} total is the solvent-accessible molecular surface area.

The QSAR models obtained for *Classes II–IV* involve descriptors that can be related to size and shape (shadow plane YZ (D_{19}), shadow plane ZX (D_7), and XY shadow/XY rectangle (D_{12})), structural information content of 2nd order (D_{20}), electrostatic and hydrogen bonding (count of H donor sites (D_{11}), HASA-1 (MOPAC PC) all (D14), maximum partial charge (Zefirov) for atom H (D_{15}), total hybridization component of the molecular dipole (D_{21})). Therefore, it can be concluded that the interaction between inhibitor and GSK-3 occurs by hydrogen bonding. It is in accordance with 3D-QSAR studies on GSK-3 inhibition that reported the involvement of steric, hydrophobic, and electrostatic factors into the calculated interaction energy and emphasize the influence of these factors on the inhibitory activity.[11,18,29] Due to molecular flexibility, the compounds belonging to these classes include a large number of conformers. The lower quality of these QSAR could be explained taking into account the complexity of the modeled property and the errors related to the experimental data.

The ANN was built on six descriptors used as inputs for the network model: *count of H-acceptor sites (Zefirov)*, *HA-dependent HDSA-1 (Zefirov)*, *Kier and Hall index, Minimum partial charge for all atom types*, *RNCG Relative negative-charged SA (SAMNEG/RNCG) (MOPAC PC), and Shadow plane ZX*. Most of these descriptors are charge-related descriptors showing the importance of the electrostatic interactions between the inhibitors and the GSK enzyme. To summarize, the total number of descriptors involved in all models developed in this work can be classified as follows:

(i) charge-related—9
(ii) topological—3
(iii) geometrical—3
(iv) MO and quantum chemical—12
(v) thermodynamic—1

4. Conclusions

A data set involving diverse chemical classes of compounds was investigated to relate pIC_{50} values against Glycogen Synthase Kinase-3 (GSK-3) to the molecular structure. QSAR modeling of the in vitro pIC_{50} inhibitory concentration that reduces 50% of the GSK-3 activity for 3-anilino-4-aryl-maleimides was carried out using the CODESSA PRO technique. A total of 739 descriptors were calculated based on molecular structure—constitutional, geometrical, topological, electrostatic, quantum chemical, and thermodynamic. The correlations obtained clearly show that the inhibitory activity of these compounds can be modeled quite satisfactorily by means of the CODESSA treatment. Each multilinear model was verified by leave-one-out and internal validation methods that confirmed the correct prediction of the inhibitory activity of 3-anilino-4-arylmaleimides.

In this study, a nonlinear treatment of the property under investigation was also carried out. An artificial neural network (ANN) was built for all the data points showing superior prediction over the multilinear models. However, the physico-chemical interpretation of the ANN is rather difficult compared with the multilinear ones. Therefore, the ANN model can be used mostly for prediction of novel inhibitors. In addition, the ANN model was externally validated during the training procedure.

These studies gave an insight into the dominant role played by the electrostatic, bonding, and steric interactions on the modulation of the inhibitory activity. As a result of the current investigations, the nature of GSK-3—inhibitor interaction is found to be electrostatic.

Since little work has been previously reported related to quantitative and predictive models for the inhibition of the GSK-3 kinase, the present article constitutes a pioneer study in this area. Furthermore, these QSPR studies could be applied to other type of kinases, as cyclin-dependent kinases. Research along these lines is continued in our laboratories and final results will be presented elsewhere.

The design of drugs with favorable pharmacology is of interest in modern medicinal chemistry; and the search for them should be assisted and guided by appropriate computational methods. It is hoped that the present work will help to accomplish this objective.

5. Methodology

The 2D structures of compounds were drawn using ISIS/Draw as implemented in the ISIS 2.4 package,

and their geometry preoptimized using the molecular mechanics force field (MM+) included in Hyperchem 7.5.[47] Final refined molecular geometries were obtained using the AM1 (Austin Model-1) semiempirical method[48] applying a gradient norm limit of 0.01 kcal/Å.

All the 960 molecular descriptors, classified as (i) 38 constitutional, (ii) 38 topological, (iii) 14 geometrical, (iv) 367 charge-related, (v) 468 semiempirical, and (vi) 35 thermodynamical, were calculated using the CODESSA-PRO software.[41]

5.1. Linear QSAR models

The best multilinear regression (BMLR) procedure[40] was used to find the best correlation models from the selected noncollinear descriptors. BMLR selects the best two-parameter regression equation, the best three-parameter regression, etc., based on the highest R^2 value in the stepwise regression procedure.[49] During the BMLR procedure the descriptor scales are normalized, centered automatically, and the final result is given in natural scales. This procedure has the best representations of the property in the given descriptor pool.

A major decision in developing successive QSPRs is when to stop adding descriptors to the model during the stepwise regression procedure. The lack of an adequate control leads to over-correlated equations, which contain an excessive number of descriptors and are difficult to analyze in terms of interaction mechanisms. A simple procedure to control the model expansion is the so-called 'break point,' resulting in the plot of the number of descriptors involved in the models versus the corresponding squared correlation coefficient. Moreover, augmentation of the number of descriptors could lead to the inclusion of variables that are highly intercorrelated. During the BMLR technique, the addition of descriptors to the QSAR equations was monitored. Thus, if no-significant improvement of the statistical parameters s, F, and especially R^2 was observed, then the current model with a certain number of descriptors reaching the break point was considered the optimum.

The QSAR models derived herein were validated by (i) the leave-one-out method and (ii) internal correlation whereby each one-third of the compounds is predicted with a model fitted by the remaining two-third of the compounds.

5.2. Nonlinear QSAR models: Artificial Neural Network (ANN)

In this work, a backpropagation ANN[50–52] was developed and used to obtain a nonlinear QSAR model. Topologically, it consists of input, hidden, and output layers of neurons or units connected by weights as shown in Figure 2. Each input layer node corresponds to a single independent variable (molecular descriptor) with the exception the bias node. Similarly, each output layer node corresponds to a different dependent variable (property under investigation).

Associated with each node is an internal state designated by I_i, H_h, and O_m for the input, hidden, output layers, respectively. Each of the input and hidden layer has an additional unit, termed a bias unit, whose internal state is assigned a value of 1. The input layer's I_i values are related to the corresponding independent variables by the scaling equation 4:

$$I_i = \frac{\mathrm{D}_i - \mathrm{D}_{i(min)} + 0.1}{\mathrm{D}_{i(\max)} - \mathrm{D}_{i(\min)} + 0.1}, \tag{4}$$

where D_i is the value of the ith descriptor, $\mathrm{D}_{i(\max)}$, and $\mathrm{D}_{i(\min)}$ are its maximum and minimum values, respectively. The state H_h of each hidden unit is calculated by the squashing (sigmoid, logistic) function:

$$H_h(\varphi_h) = \frac{1}{1 + \mathrm{e}^{-\varphi_h}}, \tag{5a}$$

$$\varphi_h = \sum_i w_{hi} I_i + \theta_h, \tag{5b}$$

where w_{hi} is the weight of the bond that connects hidden unit h with input unit i and θ_h is the weight connecting hidden unit h to the input layer bias unit. The state O_m of output unit m is calculated by,

$$O_m(\varphi_h) = \frac{1}{1 + \mathrm{e}^{-\varphi_m}}, \tag{6a}$$

$$\varphi_m = \sum_h W_{mh} H_h + \theta_m, \tag{6b}$$

where W_{mh} is the bond that connects output unit m to hidden layer bias unit. The network-calculated O_m values are within the range [0,1].

The training of the neural network is achieved by minimizing an error function E with respect to the bond weights $\{w_{hi}, W_{mh}\}$

$$E = \sum_p E_p = \frac{1}{2} \sum_p \sum_m \left(a_{pm} - O_{pm}\right)^2, \tag{7}$$

where E_p is the error of the pth training pattern, defined as the set of descriptors and activity corresponding to the pth data points, or chemical compounds; a_{pm} corresponds to the experimentally measured value of the mth dependent variable, in this case the PIC_{50}. These values were also scaled in the same manner as in Eq. 4.

One of the standard algorithms for minimizing E is the delta rule.[50–52] The algorithm is based on an iterative procedure for updating the weights of the neural network from their initially assigned random values. The equations for updating the weights are given below in Eqs. 8a and 8b:

$$W_{mh}^{n+1} = W_{mh}^{n} - \eta \frac{\partial E}{\partial W_{mh}}, \tag{8a}$$

$$w_{hi}^{n+1} = w_{hi}^{n} - \eta \frac{\partial E}{\partial w_{hi}}. \tag{8b}$$

In Eqs. 8a and 8b the superscript n indicates the consecutive iterations in the minimization procedure and η is the learning rate with values typically less than 1. Similar equations are used for θ_h and θ_m.

Acknowledgments

P.R.D. thanks the American Chemical Society for a travel grant and Dr. Eduardo A. Castro for helpful advice.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2006.03.009.

References and notes

1. Sridhar, R.; Hanson-Painton, O.; Cooper, D. R. *Pharm. Res.* **2000**, *17*, 1345.
2. Dumas, J. *Expert Opin. Ther. Patents* **2001**, *11*, 405.
3. (a) Cohen, P. *Protein Nat. Rev. Drug. Discov.* **2002**, *1*, 309; (b) Dhavale, D. D.; Desai, V. N.; Saha, N. N.; Tilekar, J. N. *ARKIVOC* **2002**, 91; (c) Murano, T.; Takechi, H.; Yuasa, Y.; Yokomatsu, T.; Umesue, I.; Soeda, S.; Shimeno, H.; Shibuya, S. *ARKIVOC* **2003**, 266; (d) Calò, S.; Tondi, D.; Venturelli, A.; Ferrari, S.; Pecorari, P.; Rinaldi, M.; Ghelli, S.; Costi, M. P. *ARKIVOC* **2004**, 382.
4. Grimes, C. A.; Jope, R. S. *Prog. Neurobiol.* **2001**, *65*, 391.
5. Hardwood, A. J. *Cell* **2001**, *105*, 821.
6. Knockaert, M.; Greengard, P.; Meijer, L. *Trends Pharmacol. Sci.* **2002**, *23*, 417.
7. Woodgett, J. R. A. *EMBO J.* **1990**, *9*, 2431.
8. Lau, K. F.; Miller, C. C. J.; Anderton, B. H.; Shaw, P. C. *J. Pept. Res.* **1999**, *54*, 85.
9. Ryves, W. J.; Harwood, A. J. *Biochem. Biophys. Res. Commun.* **2001**, *280*, 720.
10. Leclerc, S.; Garnier, M.; Hoessel, R.; Marko, D.; Bibb, J. A.; Snyder, G. L.; Greengard, P.; Biernat, J.; Wu, Y. Z.; Mandelkow, E. M.; Eisenbrand, G.; Meijer, L. *J. Biol. Chem.* **2001**, *276*, 251.
11. Kunick, C.; Lauenroth, K.; Wieking, K.; Xie, X.; Schultz, C.; Gussio, R.; Zaharevitz, D.; Leost, M.; Meijer, L.; Weber, A.; Jørgensen, F. S.; Lemcke, T. *J. Med. Chem.* **2004**, *47*, 22.
12. Smith, D. G.; Buffet, M.; Fenwick, A. E.; Haigh, D.; Ife, R. J.; Saunders, M.; Slingsby, B. P.; Stacey, R.; Ward, R. W. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 635.
13. Mettey, Y.; Gompel, M.; Thomas, V.; Garnier, M.; Leost, M.; Ceballos-Picot, I.; Noble, M.; Endicott, J.; Vierfond, J.-M.; Meijer, L. *J. Med. Chem.* **2003**, *46*, 222.
14. Meijer, L.; Thunnissen, A. M. W. H.; White, A. W.; Garnier, M.; Nikolic, M.; Tsai, L.-H.; Walter, J.; Cleverly, K. E.; Salinas, P. C.; Wu, Y. Z.; Biernat, J.; Mandelkow, E. M.; Kim, S. H.; Pettit, G. R. *Chem. Biol.* **2000**, *7*, 51.
15. Dajani, R.; Fraser, E.; Roe, S. M.; Young, N.; Good, V.; Dale, T. C.; Pearl, L. H. *Cell* **2001**, *105*, 721.
16. Martinez, A.; Alonso, M.; Castro, A.; Perez, C.; Moreno, F. J. *J. Med. Chem.* **2002**, *45*, 1292.
17. Conde, S.; Perez, D. I.; Martinez, A.; Perez, C.; Moreno, F. J. *J. Med. Chem.* **2003**, *46*, 4631–4633.
18. Polychronopoulos, P.; Magiatis, P.; Skaltsounis, A. L.; Myrianthopoulos, V.; Mikros, E.; Tarricone, A.; Musacchio, A.; Roe, S. M.; Pearl, L.; Leost, M.; Greengard, P.; Meijer, L. *J. Med. Chem.* **2004**, *47*, 935.
19. Gompel, M.; Leost, M.; Bal De Kier Joffe, E.; Puricelli, L.; Hernandez Franco, L.; Palermo, J.; Meijer, L. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 1703–1707.
20. Tavares, F. X.; Boucheron, J. A.; Dickerson, S. H.; Griffin, R. J.; Preugschat, F.; Thomson, S. A.; Wang, T. Y.; Zhou, H. Q. *J. Med. Chem.* **2004**, *47*, 4716.
21. Peat, A.; Garrido, D.; Boucheron, J. A.; Schweiker, S. L.; Dickerson, S. H.; Wilson, J. R.; Wang, T. Y.; Thomson, S. A. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 2127.
22. Peat, A. J.; Boucheron, J. A.; Dickerson, S. H.; Garrido, D.; Mills, W.; Peckham, J.; Preugschat, F.; Smalley, T.; Schweiker, S. L.; Wilson, J. R.; Wang, T. Y.; Zhou, H. Q.; Thomson, S. A. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 2121.
23. Witherington, J.; Bordas, V.; Gaiba, A.; Naylor, A.; Rawlings, A. D.; Slingsby, B. P.; Smith, D. G.; Tackle, A. K.; Ward, R. W. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3059.
24. Witherington, J.; Bordas, V.; Garland, S. L.; Hickey, D. M. B.; Ife, R. J.; Liddle, J.; Saunders, M.; Smith, D. G.; Ward, R. W. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1577.
25. Witherington, J.; Bordas, V.; Gaiba, A.; Garton, N. S.; Naylor, A.; Rawlings, A. D.; Slingsby, B. P.; Smith, D. G.; Tackle, A. K.; Ward, R. W. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3055.
26. Witherington, J.; Bordas, V.; Haigh, D.; Hickey, D. M. B.; Ife, R. J.; Rawlings, A. D.; Slingsby, B. P.; Smith, D. G.; Ward, R. W. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1581.
27. Nærum, L.; Lauritsen, L. N.; Olesen, P. H. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1525.
28. Olesen, P. H.; Sørensen, A. R.; Ursø, B.; Kurtzhals, P.; Bowler, A. N.; Ehrbar, U.; Hansen, B. F. *J. Med. Chem.* **2003**, *46*, 3333.
29. Zeng, M.; Jiang, Y.; Zhang, B.; Kewen, Z.; Zhang, N.; Yu, Q. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 395.
30. Lescot, E.; Bureau, R.; Sopkova-de Oliveira Santos, J.; Rochais, C.; Lisowski, V.; Lancelot, J. C.; Rault, S. *J. Chem. Inf. Model.* **2005**, *45*, 708.
31. Katritzky, A. R.; Karelson, M.; Lobanov, V. *Pure Appl. Chem.* **1997**, *69*, 245.
32. Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, *24*, 279–285.
33. Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1551.
34. (a) Katritzky, A. R.; Ignachenko, E.; Barcock, R.; Lobanov, V.; Karelson, M. *Anal. Chem.* **1994**, *66*, 1799; (b) Katritzky, A. R.; Lobanov, V.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28; (c) Katritzky, A. R.; Fara, D. C.; Hongfang, Y.; Karelson, M. *Chem. Rev.* **2004**, *104*, 175.
35. Katritzky, A. R.; Fara, D. C.; Yang, H.; Karelson, M.; Suzuki, T.; Solov'ev, V. P.; Varnek, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 529.
36. Katritzky, A. R.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2004**, *12*, 3027.
37. Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2005**, *13*, 1623.
38. Katritzky, A.; Dobchev, D.; Fara, D.; Karelson, M. *Bioorg. Med. Chem.* **2005**, *13*, 6598.
39. Kunick, C.; Lauenroth, K.; Leost, M.; Meijer, L.; Lemcke, T. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 413.
40. Katritzky, A. R.; Lobanov, V.; Karelson, M. *J. Phys. Chem.* **1996**, *100*, 10400.
41. www.codessa-pro.com.
42. Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
43. Julg, A.; Julg, P. *Int. J. Quantum Chem.* **1978**, *13*, 483.

44. Ángyan, J. G.; Rosta, E.; Surjan, P. R.. *Chem. Phys. Lett.* **1999**, *299*, 1.
45. Csizmadia, I. G. *Theory and Practice of MO Calculations on Organic Molecules*; Elsevier: Amsterdam, 1976.
46. Sannigrahi, A. B. *Adv. Quantum Chem.* **1992**, *23*, 301.
47. www.hyper.com.
48. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
49. Karelson, M. In *Molecular Descriptors in QSAR/QSPR*; Wiley Interscience: New York, 2000; p 79, 282.
50. Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999.
51. Masters, T. *Practical Neural Network Recipes in C++*; Academic Press: Boston, 1993.
52. Haykin, S. *In Neural Networks*; Pearson Education: Delhi, 2004.