Three-Dimensional Quantitative Structure−Activity Relationships from Molecular Similarity Matrices and Genetic Neural Networks. 2. Applications

Sung-Sau So*,† and Martin Karplus*,†,‡

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, and Laboratoire de Chimie Biophysique, Institut le Bel, Université Louis Pasteur, 4 rue Blaise Pascal, 67000 Strasbourg, France

Received July 28, 1997[⊗]

Validation of a method that uses a genetic neural network with electrostatic and steric similarity matrices (SM/GNN) to obtain quantitative structure−activity relationships (QSARs) is performed with eight data sets. Biological and physicochemical properties from a broad range of chemical classes are correlated and predicted using this technique. Quantitatively the results compare favorably with the benchmarks obtained by a number of well-established QSAR methods; qualitatively the models are consistent with the published descriptions on the relative contribution of steric and electrostatic factors. The results demonstrate the general utility of this method in deriving QSARs. The implication of the importance of molecular alignment and possible methodological improvements are discussed.

I. Introduction

In the companion paper[1] we have described an approach for using a genetic neural network (GNN)[2,3] to construct quantitative structure−activity relationships (QSARs) from molecular similarity matrices (SMs). We obtained highly predictive statistical models for the well-studied corticosteroid-binding globulin (CBG)-binding steroid data set. Since the steroid data set is somewhat limited, the aim of this paper is to provide an extended test with several additional data sets.

We use five biological data sets that have been extensively studied with established QSAR methods so that the results can be compared with published work. In addition, we apply the method to a set of glycogen phosphorylase (GP) inhibitors. GP is a large enzyme that plays a regulatory role in glycogen metabolism. As a potential therapeutic target relating to the treatment of diabetes, there has been considerable interest in designing a more potent GP inhibitor than α-D-glucose, the physiological regulator.[4] Numerous crystallographic studies,[5−9] kinetic binding experiments,[4,9−11] and 3D QSAR investigations[4,12] of GP−ligand complexes have been made. For this study, Johnson and co-workers have gladly made a set of GP−ligand cocrystallized structures available to us.

Also we use the SM/GNN method to examine physicochemical properties, specifically the Hammett constants of substituted benzoic acids and the pK_a values of imidazoles. These two properties are commonly used as standard 2D QSAR parameters describing electrostatic interactions between a drug and its receptor. 3D QSAR methods, like comparative molecular field analysis (CoMFA), generally rely on an electrostatic field to provide such information. A number of studies have been made to examine the ability of the 3D methods to reproduce these physicochemical properties.[13−17] A good fit of 2D parameters using 3D methods would verify the mutual consistency of their electrostatic descriptions.

II. Method

Data Sets. Eight data sets were examined in this study. Six of these were concerned with biological activity or binding data and two with physicochemical properties. The six biological data sets are (a) 73 polyhalogenated aromatic compounds that bind to the cytosolic aromatic hydrocarbon (Ah) receptor (Table 1a);[18,19] (b) 47 1-(substituted benyzl)-imidazole-2(3*H*)-thiones with inhibitory activities on dopamine β-hydroxylase (DβH; Table 1b);[20,21] (c) 43 β-carboline, pyridodiindole, and CGS inverse agonists and antagonists of benzodiazepine receptor (BzR; Table 1c);[15,17,22−24] (d) 60 structurally diverse inhibitors of acetylcholinesterase (AChE) that have recently been studied with a new CoMFA/q^2-GRS approach (Table 1d);[25] (e) 37 bisamidines with potency against *Leishmania mexicana amazonensis*[26] (Table 1e); and (f) 30 GP inhibitors whose bound X-ray coordinates with the enzyme have been determined (Table 1f).[4,9,10,12] The two physicochemical data sets are (g) 72 substituted benzoic acids with Hammett constant data (Table 1g)[14−17] and (h) 16 imidazoles with pK_a data (Table 1h).[13,15,17]

The compounds in data series a, c, g, and h were manually built using the 3D sketcher facility in the Cerius2 program,[27] and the structures were energy minimized using the default force field.[28] The structures of DβH inhibitors in series b were taken from an example file in a Cerius2 release. The coordinates of the AChE inhibitors in series d were provided by Cho et al.[25] and those of the bisamidines in series e by Montanari et al.[26] The GP inhibitors coordinates in series f were extracted from the X-ray cocrystallization structures.[4,10]

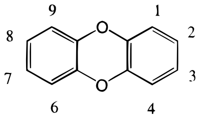
Generation of Similarity Indices. The generation of the molecular fields was done as described in the companion paper.[1] AM1 Mulliken charges were obtained using the MOPAC program[29] (version 6) with the default setting. The electrostatic similarity index between two molecules was computed with the Hodgkin formula[30] (eq 1) using electrostatic potentials that were calculated at regular grid points. A rectilinear grid, whose size had a 6-Å extension beyond all atomic coordinates and a regular grid spacing of 2 Å, was used. A vacuum dielectric ($\epsilon = 1.0$) was used for the computation of Coulombic electrostatic potentials. A truncation cutoff of ±5 kcal/mol was applied to the potentials at all grid points, except for the series d and e where a higher value (±100 kcal/mol) was necessary to make the similarity index more discriminating for molecules that had net positive charges.

$$H_{AB} = \frac{2\sum P_A P_B}{\sum {P_A}^2 + \sum {P_B}^2} \qquad (1)$$

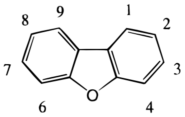
[†] Harvard University.
[‡] Université Louis Pasteur.
[⊗] Abstract published in *Advance ACS Abstracts,* December 1, 1997.

Table 1

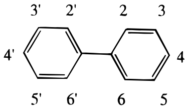
(a) Structure and Affinity Data for the Ah Data Set of Series a



A



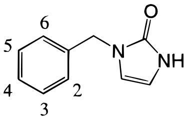
B



C

no.	structure	R	pEC_{50}	no.	structure	R	pEC_{50}
1	A	2,3,7,8-Cl_4	8.000	**38**	B	2,3,4,8-Cl_4	6.699
2	A	1,2,3,7,8-Cl_5	7.102	**39**	B	2,3,7,8-Cl_4	7.387
3	A	2,3,6,7-Cl_4	6.796	**40**	B	1,2,4,8-Cl_4	5.000
4	A	2,3,6-Cl_3	6.658	**41**	B	1,2,4,6,7-Cl_5	7.169
5	A	1,2,3,4,7,8-Cl_6	6.553	**42**	B	1,2,4,7,9-Cl_5	4.699
6	A	1,3,7,8-Cl_4	6.102	**43**	B	1,2,3,4,8-Cl_5	6.921
7	A	1,2,4,7,8-Cl_5	5.959	**44**	B	1,2,3,7,8-Cl_5	7.128
8	A	1,2,3,4-Cl_4	5.886	**45**	B	1,2,4,7,8-Cl_5	5.886
9	A	2,3,7-Cl_3	7.149	**46**	B	2,3,4,7,8-Cl_5	7.824
10	A	2,8-Cl_2	5.495	**47**	B	1,2,3,4,7,8-Cl_6	6.638
11	A	1,2,3,4,7-Cl_5	5.194	**48**	B	1,2,3,6,7,8-Cl_6	6.569
12	A	1,2,4-Cl_3	4.886	**49**	B	1,2,4,6,7,8-Cl_6	5.081
13	A	1,2,3,4,6,7,8,9-Cl_8	5.000	**50**	B	2,3,4,6,7,8-Cl_6	7.328
14	A	1-Cl	4.000	**51**	B	2,3,6,8-Cl_4	6.658
15	A	2,3,7,8-Br_4	8.824	**52**	B	1,2,3,6-Cl_4	6.456
16	A	2,3-Br_2, 7,8-Cl_2	8.830	**53**	B	1,2,3,7-Cl_4	6.959
17	A	2,8-Br_2, 3,7-Cl_2	9.350	**54**	B	1,3,4,7,8-Cl_5	6.699
18	A	2-Br, 3,7,8-Cl_3	7.939	**55**	B	2,3,4,7,9-Cl_5	6.699
19[a]	A	1,3,7,9-Br_4	7.032	**56**	B	1,2,3,7,9-Cl_5	6.398
20	A	1,3,7,8-Br_4	8.699	**57**	B		3.000
21	A	1,2,4,7,8-Br_5	7.770	**58**	B	2,3,4,7-Cl_4	7.602
22	A	1,2,3,7,8-Br_5	8.180	**59**	B	1,2,4,6,8-Cl_5	5.509
23	A	2,3,7-Br_3	8.932	**60**	C	3,3′,4,4′-Cl_4	6.149
24	A	2,7-Br_2	7.810	**61**	C	3,4,4′,5-Cl_4	4.553
25	A	2-Br	6.530	**62**	C	3,3′,4,4′,5-Cl_5	6.886
26	B	2-Cl	3.553	**63**	C	2′,3,4,4′,5-Cl_5	4.854
27	B	3-Cl	4.377	**64**	C	2,3,3′,4,4′-Cl_5	5.367
28	B	4-Cl	3.000	**65**	C	2,3′,4,4′,5-Cl_5	5.041
29	B	2,3-Cl_2	5.326	**66**	C	2,3,4,4′,5-Cl_5	5.387
30	B	2,6-Cl_2	3.609	**67**[b]	C	2,3,3′,4,4′,5-Cl_6	5.149
31	B	2,8-Cl_2	3.590	**68**	C	2,3′,4,4′,5,5′-Cl_6	4.796
32	B	1,3,6-Cl_3	5.357	**69**[b]	C	2,3,3′,4,4′,5′-Cl_6	5.301
33	B	1,3,8-Cl_3	4.071	**70**	C	2,2′,4,4′-Cl_4	3.886
34	B	2,3,4-Cl_3	4.721	**71**	C	2,2′,4,4′,5,5′-Cl_6	4.102
35	B	2,3,8-Cl_3	6.000	**72**	C	2,3,4,5-Cl_4	3.854
36	B	2,6,7-Cl_3	6.347	**73**	C	2,3′,4,4′,5′,6-Cl_6	4.004
37	B	2,3,4,6-Cl_4	6.456				

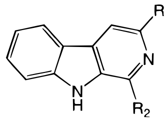
(b) Structure and Affinity Data for the DβH Data Set of Series b



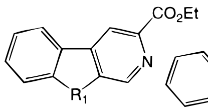
no.	R	pIC_{50}	no.	R	pIC_{50}	no.	R	pIC_{50}
1	2,6-$(CH_3)_2$	3.00	**17**	3-NO_2, 4-OCH_3	3.45	**33**	H	4.48
2	2,6-Cl_2	3.15	**18**	4-OCH_3	3.69	**34**	3-NO_2, 4-OH	4.51
3	2,6-$(OCH_3)_2$	3.30	**19**	3-OCH_3	3.80	**35**	3,4-Cl_2	4.55
4	2-Cl	3.45	**20**	3-OH	3.83	**36**	2,4-Cl_2	4.77
5	2-CH_3	3.47	**21**[c]	3-CF_3, 4-OH	3.92	**37**	3-Br, 4-OH	4.92
6	3,4-$(OCH_3)_2$	3.47	**22**	2,4,6-Cl_3	3.99	**38**	3-Cl	4.92
7	4-CF_3	3.70	**23**	2,5-Cl_2	4.01	**39**	3-F	5.25
8	3-CF_3, 4-OCH_3	3.76	**24**	4-Cl	4.02	**40**	4-OH	5.59
9	2,6-Cl_2, 4-OCH_3	3.81	**25**	2,6-Cl_2, 4-OH	4.12	**41**	3,5-Cl_2	5.62
10	4-CH_3	3.83	**26**	2,3,5,6-F_4, 4-OH	4.21	**42**	3,4-$(OH)_2$	5.66
11	4-Br	3.94	**27**	4-NO_2	4.28	**43**	3-Cl, 4-OH	5.70
12	3-Br, 4-OCH_3	4.08	**28**	2,3-Cl_2	4.28	**44**	3-F, 4-OH	5.82
13	3-F, 4-OCH_3	4.13	**29**	3-CH_3, 4-OH	4.31	**45**	3,5-F_2	5.92
14	2-OCH_3	4.13	**30**	4-F	4.33	**46**	3,5-Cl_2, 4-OH	6.17
15	3-CH_3, 4-OCH_3	4.16	**31**	3,5-Cl_2, 4-OCH_3	4.33	**47**	3,5-F_2, 4-OH	7.13
16	2-OH	3.24	**32**	3,5-F_2, 4-OCH_3	4.44			

Table 1 (Continued)

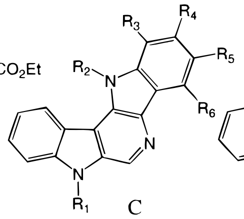
(c) Structure and Affinity Data for the BzR Data Set of Series c[d]



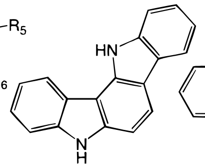
A



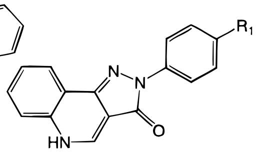
B



C



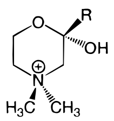
D



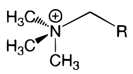
E

no.	structure	R1	R2	R3	R4	R5	R6	pIC50
1	A	CO_2CH_3	H					8.30
2	A	$CO_2CH_2CH_3$	H					8.30
3	A	OCH_2CH_3	H					7.62
4	A	$OCH(CH_3)_2$	H					6.29
5	A	$OCH_2CH_2CH_3CH_3$	H					7.01
6	A	OCH_3	H					6.91
7	A	$OCH_2CH_2CH_3$	H					7.96
8	A	$COCH_2CH_2CH_3$	H					7.64
9	A	$CH_2CH_2CH_2CH_3$	H					6.64
10	A	H	H					5.79
11	A	$CO_2C(CH_3)_3$	H					8.00
12	A	Cl	H					7.35
13	A	NO_2	H					6.90
14	A	$CO_2CH_2C(CH_3)_3$	H					6.12
15	A	CO_2CH_3	CH_2CH_3					5.12
16	A	H	CH_2CH_3					3.60
17	A	H	CH_3					4.91
18	B	C(=O)						4.59
19	B	C(=NOH)						5.24
20	B	O						5.07
21	B	CH_2						6.17
22	B	C(=O)N(H)						5.62
23	B	S						5.77
24	C	H	H	H	H	H	H	8.40
25	C	H	H	CH_3	H	H	H	7.17
26	C	H	H	H	CH_3	H	H	8.10
27	C	H	H	H	H	CH_3	H	6.65
28	C	H	H	H	H	H	CH_3	5.16
29	C	CH_3	H	H	H	H	H	5.93
30	C	H	CH_3	H	H	H	H	6.80
31	C	CH_3	CH_3	H	H	H	H	5.71
32	C	H	H	H	H	H	OCH_3	6.60
33	C	H	H	H	H	H	Cl	6.15
34	D							5.71
35	E	H						9.40
36	E	Cl						9.22
37	E	OCH_3						10.00
38	A	$OCH(CH_3)CH_2CH_3$	H					6.33
39	A	$OCH_2CH(CH_3)_2$	H					7.03
40	A	$OCH_2CH_2CH(CH_3)_2$	H					6.27
41	A	$OCH_2C(CH_3)_3$	H					7.02
42	A	$OCH_2C_6H_5$	H					6.00
43	A	$COC(CH_3)_3$	H					6.45

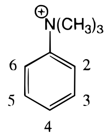
(d) Structure and Affinity Data for the AChE Data Set of Series d



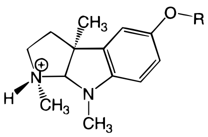
A



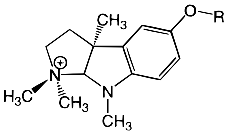
B



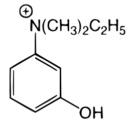
C



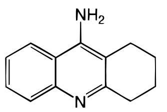
D



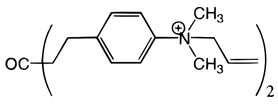
E



F



G



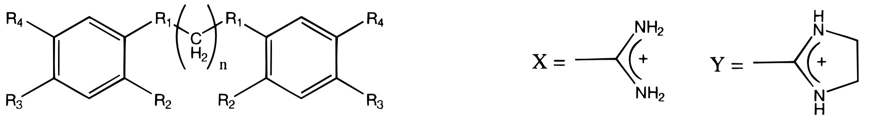
H

no.	structure	R	pIC50	no.	structure	R	pIC50
1	A	$4\text{-}NHCOCH_3\text{-}C_6H_4$	2.684	**31**	C	$2\text{-}CH_3$, 5-OH	5.549
2	A	$4\text{-}NH_2\text{-}C_6H_4$	3.161	**32**	C	3-OH, $4\text{-}CH_3$	5.507
3	A	$4\text{-}Cl\text{-}C_6H_4$	2.090	**33**	C	$3\text{-}OCOCH_3$	5.521

Table 1 (Continued)

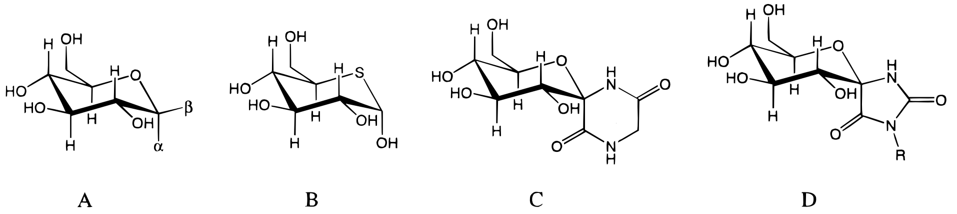
no.	structure	R	pIC$_{50}$	no.	structure	R	pIC$_{50}$
4	A	4-CN-C_6H_4	1.936	**34**	C	2-CH_3, 3-$OCOCH_3$	4.389
5	A	4-NO_2-C_6H_4	2.291	**35**	C	2-CH_3, 5-$OCOCH_3$	5.273
6	A	H	2.754	**36**	C	3-$OCOCH_3$, 4-CH_3	6.181
7	A	CH_3	2.762	**37**	C	3-$OCOCH_2CH_3$	5.224
8	A	$C(CH_3)_3$	2.431	**38**	C	2-CH_3, 3-$OCOCH_2CH_3$	3.123
9	A	CF_3	2.417	**39**	C	2-CH_3, 5-$OCOCH_2CH_3$	4.161
10	B	CH_2OH	2.521	**40**	C	3-$OCOCH_2CH_3$, 4-CH_3	5.424
11	B	CH_2Cl	2.622	**41**	C	3-OCH_3	3.224
12	B	CH_2SH	3.357	**42**	C	2-CH_3, 3-OCH_3	3.622
13	B	$CH_2OCH_2CH_3$	2.936	**43**	C	2-CH_3, 5-OCH_3	3.462
14	B	$CH_2SCH_2CH_3$	2.821	**44**	C	3-OCH_3, 4-CH_3	3.912
15	B	$CH_2OCOCH_2CH_2CH_3$	3.640	**45**	D	$CONHCH_3$	7.244
16	B	$CH_2SCOCH_2CH_2CH_3$	3.900	**46**	D	$CONH(CH_2)_7CH_3$	6.959
17	B	$CH_2CH_2COCH_3$	4.072	**47**	D	$CONH(CH_2)_3CH_3$	6.818
18	B	$CH{=}CH_2$	2.535	**48**	D	$CONHCH_2C_6H_5$	6.337
19	B	$CH_2CH{=}CH_2$	3.072	**49**	D	$CONHC_6H_5$	6.456
20	B	$CH_2CH_2CH{=}CH_2$	2.947	**50**	E	$CONHCH_3$	7.469
21	B	CH_2CH_2Br	4.056	**51**	D	CONH(4-OCH_3-C_6H_4)	5.770
22	B	$CH_2CH_2CH_2Br$	3.224	**52**	D	$CON(CH_3)_2$	6.013
23	B	$CH_2CH_2CH_2I$	3.327	**53**	D	CONH(4-Cl-C_6H_4)	5.745
24	B	$CH_2CH_2CH_2CH_2Br$	3.272	**54**	D	$CON(CH_3)CONHCH_2C_6H_5$	5.201
25	B	$CH_2CH_2CH_2CH_2CH_3$	3.088	**55**	D	$CON(CH_2CH_3)_2$	4.398
26	C	3-NO_2	3.202	**56**	D	$CONHCH(CH_3)_2$	4.456
27	C	3-$NHCOCH_3$	3.000	**57**	F		6.108
28	C	3-NH_3^+	3.717	**58**	C	3-$OCON(CH_3)_2$	7.041
29	C	3-OH	6.012	**59**	G		7.119
30	C	2-CH_3, 3-OH	3.850	**60**	H		8.097

(e) Structure and Affinity Data for the Bisamidine Data Set of Series e



no.	*n*	R$_1$	R$_2$	R$_3$	R$_4$	pIC$_{50}$	no.	*n*	R$_1$	R$_2$	R$_3$	R$_4$	pIC$_{50}$
1	2	O	H	X	H	1.801	**20**	5	O	Cl	X	H	3.153
2	3	O	H	X	H	3.070	**21**	5	O	Br	X	H	3.169
3	4	O	H	X	H	2.799	**22**	3	NH	H	X	H	3.163
4	5	O	H	X	H	3.086	**23**	4	NH	H	X	H	3.173
5	6	O	H	X	H	3.402	**24**	5	NH	H	X	H	3.253
6	3	O	H	H	X	2.215	**25**	6	NH	H	X	H	3.539
7	4	O	H	H	X	2.265	**26**	3	NH	NO_2	X	H	2.316
8	5	O	H	H	X	2.671	**27**	5	NH	NO_2	X	H	2.947
9	6	O	H	H	X	2.985	**28**	2	NH	NH_2	X	H	1.581
10	2	O	NO_2	X	H	1.301	**29**	4	NH	NH_2	X	H	2.104
11	4	O	NO_2	X	H	2.252	**30**	5	NH	NH_2	X	H	2.936
12	5	O	NO_2	X	H	2.700	**31**	6	NH	NH_2	X	H	3.004
13	2	O	NH_2	X	H	1.646	**32**	3	O	H	Y	H	2.751
14	3	O	NH_2	X	H	2.456	**33**	4	O	H	Y	H	2.567
15	4	O	NH_2	X	H	2.121	**34**	5	O	H	Y	H	2.765
16	3	O	OCH_3	X	H	2.748	**35**	3	O	OCH_3	Y	H	2.646
17	4	O	OCH_3	X	H	1.998	**36**	4	O	OCH_3	Y	H	2.193
18	5	O	OCH_3	X	H	2.518	**37**	5	O	OCH_3	Y	H	2.394
19	4	O	Cl	X	H	2.876							

(f) Structure and Affinity Data for the GP Data Set of Series f

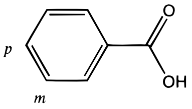


no.	structure	R	p*K*$_i$	no.	structure	R	p*K*$_i$
1	A	α-$CONHCH_3$	1.435	**16**	A	β-CONH-cyclopropyl	2.886
2	A	β-SCH_2CONH_2	1.676	**17**	A	β-$NHCOCH_2NHCOCH_3$	3.004
3	A	β-CH_2CONH-2,4-F_2-C_6H_3	1.724	**18**	A	β-$CONH_2$	3.357
4	A	α-$CONHC_2H_5OH$	1.772	**19**	A	β-$CONHNH_2$	3.398
5	A	β-CH_2NH_3+	1.775	**20**	A	α-$CONH_2$	3.432
6	A	β-*O*-(1-6)-D-glucose	1.788	**21**	A	β-$NHCOCH_2NH_2$	3.432
7	A	β-CH_2N_3	1.818	**22**	A	β-$CONHCH_3$	3.796

Table 1 (Continued)

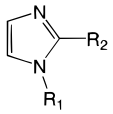
no.	structure	R	pK_i	no.	structure	R	pK_i
8	A	α-CONH-4-OH-C_6H_4	2.252	**23**	D	NH_2	3.836
9	A	β-$CH_2OSO_2CH_3$	2.319	**24**	A	β-$NHCONH_2$	3.854
10	A	β-$SCH_2CONHC_6H_5$	2.444	**25**	A	β-$NHCOC_3H_7$	4.027
11	A	β-$CONHNH_2$	2.523	**26**	C		4.229
12	A	β-CO_2CH_3	2.553	**27**	A	β-$NHCOCH_2Cl$	4.347
13	B		2.699	**28**	A	β-$NHCOCH_3$	4.495
14	A	β-$CONHNHCH_3$	2.745	**29**	A	α-$CONH_2$, *b*-$NHCO_2CH_3$	4.796
15	A	α-OH	2.770	**30**	D	H	5.523

(g) Structure and Hammett Constant Data for the Benzoic Acids of Series g[e]



no.	R	σ	no.	R	σ	no.	R	σ
1	H	0.00	**25**	*m*-SO_2CH_3	0.60	**49**	*p*-SO_2CH_3	0.72
2	*m*-Br	0.39	**26**	*p*-Br	0.23	**50**	*m*-$CH{=}CH_2$	0.05
3	*m*-CF_3	0.43	**27**	*p*-CF_3	0.54	**51**	*m*-CH_2CN	0.16
4	*m*-CH_3	−0.07	**28**	*p*-CH_3	−0.17	**52**	*m*-CHO	0.35
5	*m*-Cl	0.37	**29**	*p*-Cl	0.23	**53**	*m*-CH_2OCH_3	0.02
6	*m*-CN	0.56	**30**	*p*-CN	0.66	**54**	*m*-$COCH_3$	0.38
7	*m*-F	0.34	**31**	*p*-F	0.06	**55**	*m*-$CONH_2$	0.28
8	*m*-I	0.35	**32**	*p*-I	0.18	**56**	*m*-NCS	0.48
9	*m*-NH_2	−0.16	**33**	*p*-NH_2	−0.66	**57**	*m*-$NHCH_3$	−0.30
10	*m*-NO_2	0.71	**34**	*p*-NO_2	0.78	**58**	*m*-$N(CH_3)_2$	−0.15
11	*m*-OCF_3	0.38	**35**	*p*-OCF_3	0.35	**59**	*m*-$OCOCH_3$	0.39
12	*m*-OH	0.12	**36**	*p*-OH	−0.37	**60**	*m*-SCN	0.41
13	*m*-OCH_3	0.12	**37**	*p*-OCH_3	−0.27	**61**	*m*-SO_2NH_2	0.46
14	*m*-SH	0.25	**38**	*p*-SH	0.15	**62**	*p*-$CH{=}CH_2$	−0.02
15	*m*-SCH_3	0.15	**39**	*p*-SCH_3	0.00	**63**	*p*-CH_2CN	0.01
16	*m*-SCF_3	0.40	**40**	*p*-SCF_3	0.50	**64**	*p*-CHO	0.42
17	*m*-$C(CH_3)_3$	−0.10	**41**	*p*-$C(CH_3)_3$	−0.20	**65**	*p*-CH_2OCH_3	0.03
18	*m*-C_2F_5	0.47	**42**	*p*-C_2F_5	0.52	**66**	*p*-$COCH_3$	0.50
19	*m*-CH_2Br	0.12	**43**	*p*-CH_2Br	0.14	**67**	*p*-$CONH_2$	0.36
20	*m*-CH_2Cl	0.11	**44**	*p*-CH_2Cl	0.12	**68**	*p*-NCS	0.38
21	*m*-CH_2I	0.10	**45**	*p*-CH_2I	0.11	**69**	*p*-$NHCH_3$	−0.84
22	*m*-C_2H_5	−0.07	**46**	*p*-C_2H_5	−0.15	**70**	*p*-$N(CH_3)_2$	−0.83
23	*m*-SO_2CF_3	0.79	**47**	*p*-SO_2CF_3	0.93	**71**	*p*-SCN	0.52
24	*m*-SO_2F	0.80	**48**	*p*-SO_2F	0.91	**72**	*p*-SO_2NH_2	0.57

(h) Structure and pK_a Data for the Imidazoles of Series h



no.	R_1	R_2	pK_a	no.	R_1	R_2	pK_a
1	CH_3	Br	3.82	**9**	H	F	2.40
2	CH_3	F	2.30	**10**	H	H	6.99
3	CH_3	H	7.12	**11**	H	CH_3	7.86
4	CH_3	NH_2	8.54	**12**	H	NH_2	8.46
5	CH_3	NO_2	−0.48	**13**	H	NO_2	−0.81
6	H	Br	3.79	**14**	H	C_6H_5	6.48
7	H	Cl	3.55	**15**	H	2-pyridyl	5.36
8	H	C_2H_5	7.73	**16**	H	SCH_3	5.95

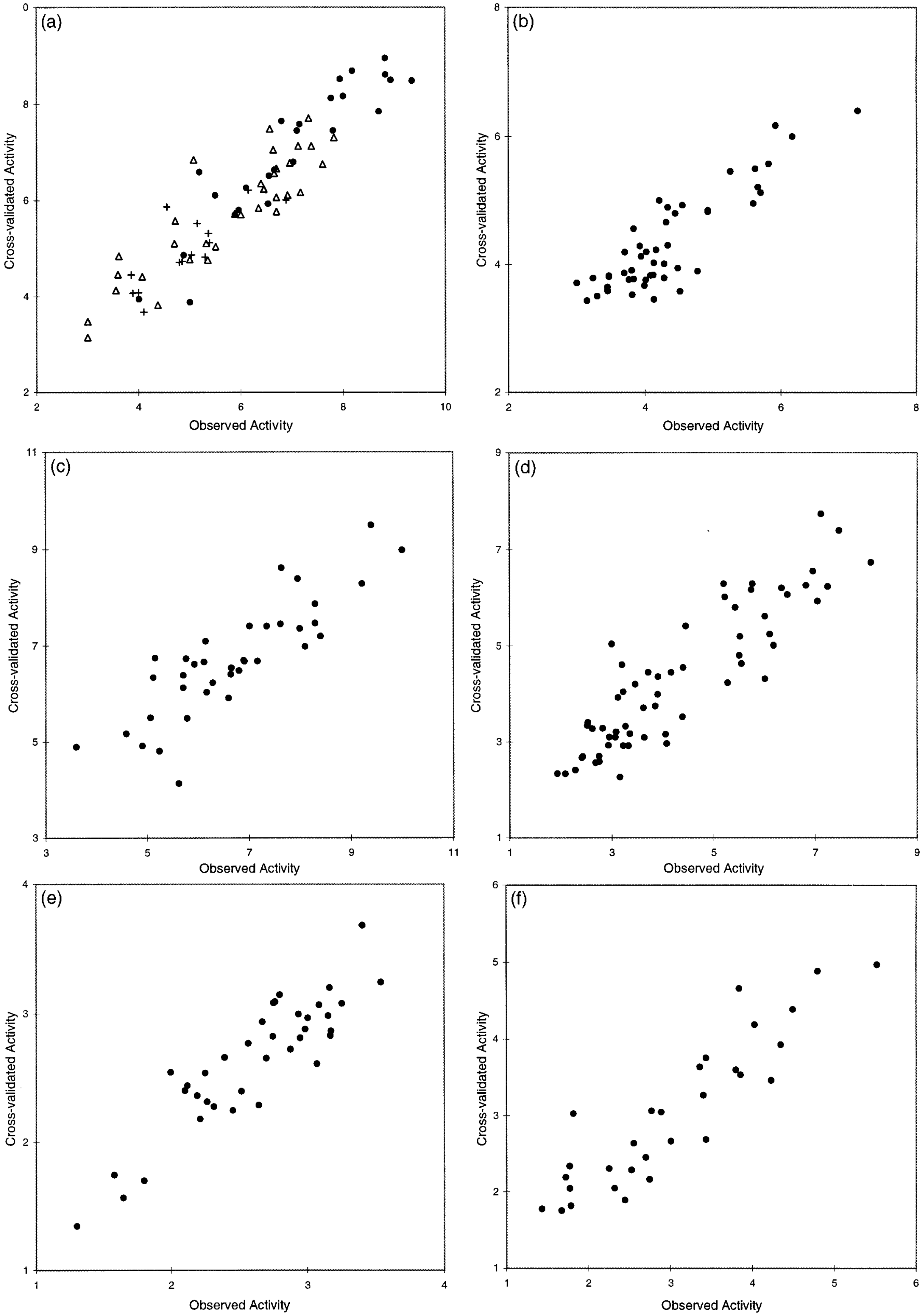
[a] Incorrect structure in Wagener et al.[19] [b] Incorrect activity in Waller and McKinney[18] and Wagener et al.[19] [c] Incorrect structure in the Cerius2 example file. [d] **1**−**37** training set; **38**−**43** test set. [e] **1**−**49** training set; **50**−**72** test set.

For the shape comparison, a grid with an extension of 2 Å beyond the molecular boundary was constructed. A regular grid spacing of 0.5 Å was used. The shape index was based on the Meyer formula[31] (eq 2):

$$S_{AB} = \frac{U_{AB}}{\sqrt{T_A T_B}} \qquad (2)$$

where U is the number of grid points enclosed by the volume that was the union of the two molecules being compared and T_A and T_B are the number of grid points inside their individual volumes.

Genetic Neural Network. A data set containing N training compounds leads to $N \times N$ electrostatic and shape SMs. GNN simulations were performed on each of the SMs for every data set. For the six biological series, an additional GNN run was done on the $N \times 2N$ combined matrix containing the two properties. In the companion study, we have shown that the number of descriptors (n) used in a GNN model is an important parameter.[1] In the current work, we use a procedure similar to that employed with CoMFA to find the model with optimal predictivity. We applied GNN simulations on the same matrix using different values of n, ranging from 1 to a maximum of N/5. The optimal model is the one that gave the smallest standard error (σ) in cross-validation predictions:


(a)
Cross-validated Activity
Observed Activity
0
8
6
4
2
2
4
6
8
10
(b)
Cross-validated Activity
Observed Activity
8
6
4
2
2
4
6
8
(c)
Cross-validated Activity
Observed Activity
11
9
7
5
3
3
5
7
9
11
(d)
Cross-validated Activity
Observed Activity
9
7
5
3
1
1
3
5
7
9
(e)
Cross-validated Activity
Observed Activity
4
3
2
1
1
2
3
4
(f)
Cross-validated Activity
Observed Activity
6
5
4
3
2
1
1
2
3
4
5
6

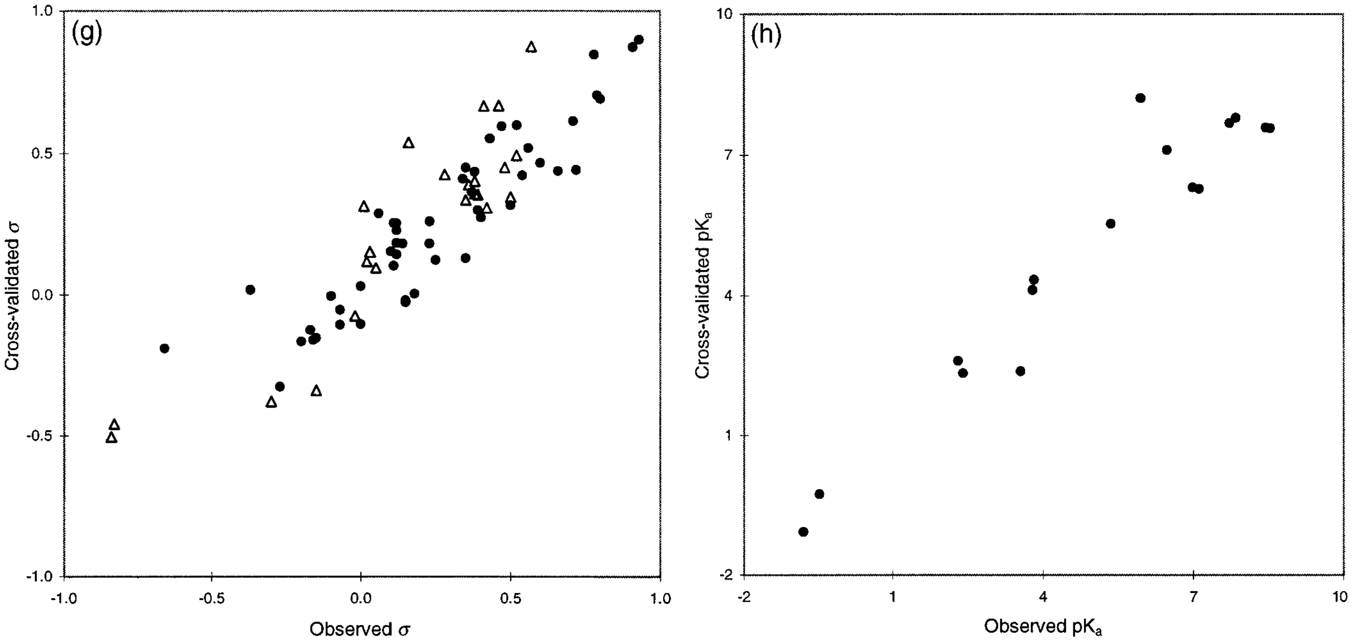


Figure 1. Plots of cross-validated activities against observed values for (a) 73 Ah compounds (series a) (dioxins are shown as solid circles, furans as open triangles, and biphenyls as crosses), (b) 47 DβH inhibitors (series b), (c) 37 inverse agonists of BzR (series c), (d) 60 AChE inhibitors (series d), (e) 37 bisamidines (series e), and (f) 30 GP inhibitors (series f). (g) Plot of cross-validated and predicted Hammett constants against observed values for 72 benzoic acids (series g) (training compounds are shown as solid circles and test compounds as open triangles). (h) Plot of cross-validated pK_a against the observed values for 16 imidazoles (series h).

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(y_{\mathrm{i,observed}} - y_{\mathrm{i,predicted}})^2}{N - n - 1}} \qquad (3)$$

where $y_{\mathrm{i,observed}}$ is the observed property value and $y_{\mathrm{i,predicted}}$ is the predicted value from cross-validated prediction. This formula is related to the standard deviation of the error of predictions in the CoMFA PLS analysis,[32] and it penalizes models with a large number of descriptors. In this way, simpler models that are marginally less predictive are preferred if the increase of predictivity in the more complex model does not justify the higher associated risk of data overfitting.[33-35] All neural networks used in this study had a configuration of n-2-1.

The overall predictive quality of the GNN models was determined by the use of a cross-validated r,[2] which is commonly referred to as q^2 to distinguish from the conventional Pearson correlation coefficient. It is defined as:

$$\text{cross-validated } r^2 \equiv q^2 = 1 - \frac{\sum_{i=1}^{N}(y_{\mathrm{i,observed}} - y_{\mathrm{i,predicted}})^2}{\sum_{i=1}^{N}(y_{\mathrm{i,observed}} - \overline{y_{\mathrm{i,observed}}})^2} \qquad (4)$$

For the maximum possible correlation of the data, q^2 takes a value of 1. A zero value indicates that the predictions are no better than those made randomly; i.e., all compounds were predicted at the average observed output, $\overline{y_{\mathrm{i,observed}}}$. A negative q^2 value, which arises from anticorrelation, is possible.

III. Results

Predictions of Biological Activities. (a) Ah Data Set. Waller and McKinney (WMcK) used the CoMFA method to correlate a set of 78 polyhalogenated aromatic compounds with their Ah receptor binding affinities.[18] Their six-component PLS model yielded a q^2 value of 0.72. In a recent study, Wagener et al. obtained an excellent QSAR for the same data set.[19] They computed a 12-element autocorrelation vector (AV) that encoded molecular surface properties for each of the compounds and used them as input to a neural network (NN). The AV/NN model gave a high q^2 value of 0.83.

In a preliminary analysis of this work, we had identified four duplicate entries in the data set used by the two research groups (PCDF29−32 and PCDF35−38 in Table 2 of WMcK are identical). These redundant entries diminished the significance of cross-validation since the estimated predictivity would almost certainly be too large. For this reason, the four duplicates were removed in this study. In addition, a comparison of the data set with the original literature[36-38] revealed several discrepancies. The entry for PCDF14, whose pEC_{50} value was not reported in the cited sources, appeared to be a compilation error. Also, the activities for PCB8 and PCB10 reported by both groups were incorrect. The recompiled data set is shown in Table 1a, and it consists of 25 dioxin, 34 furan, and 14 biphenyl derivatives. Partly due to the symmetry of the parent compounds, the alignment for these compounds was not well-defined even for compounds within a given congeneric series. In this study, we used the alignment procedure described by WMcK.

The 73 × 73 SMs for electrostatic and shape properties were obtained using the set of aligned structures. Application of the GNN algorithm gave q^2 values of 0.72 and 0.85 for the electrostatic and shape matrices, respectively. This result was in agreement with the CoMFA study, which reported that the steric influence was more important. The GNN model derived from the combined similarity matrix contained two electrostatic and five shape descriptors. This composite model showed no improvement in predictivity (q^2 = 0.85; Figure 1a) relative to the previous GNN model obtained directly from the shape matrix.

The SM/GNN results were significantly better than the CoMFA statistics for this data set. This method also performed marginally superior than the AV/NN method that employed more descriptors for correlation. The principal source of error in this study probably came from the uncertainty of the alignment,[18] which is a major problem in the 3D QSAR methods that required molecular superposition. In the absence of experimental information of the ligand−receptor structures, the current alignment or one that is based on an active analogue approach[39] seemed a reasonable starting point. Nevertheless, other alternative alignments should also be explored because they might lead to more predictive models. In this regard, we are developing a GA-based method that is used in conjunction with a similarity-based approach to obtain molecular alignments.

(b) DβH Data Set. The set of 47 1-(substituted benzyl)imidazole-2(3*H*)-thiones and their inhibitory activity with DβH had been correlated by two different QSAR approaches. Burke and Hopfinger performed molecular shape analysis (MSA) on this set of compounds.[20] They formulated a six-descriptor regression equation involving linear and quadratic variables encoding steric, electrostatic, and hydrophobic characteristics of the ligands (eq 5):

$$\mathrm{pIC}_{50} = -119.6V_0 + 70.6{V_0}^2 + 2.09Q_{3,4,5} - 4.63Q_6 + 0.046{\pi_0}^2 - 0.595\pi_4 + 53.38 \quad (5)$$

where V_0 is a 3D measure denoting the common overlap steric volume against the most active compound, $Q_{3,4,5}$ is the sum of partial atomic charges on atoms 3, 4, and 5, Q_6 is the partial atomic charge on atom 6, π_0 is the molecular lipophilicity, and π_4 is the water/octanol fragment constant of the 4-substituent. Because cross-validated statistics were not reported, we recalculated this regression QSAR based on the published descriptor values and obtained a value of 0.76 for q.[2]

Recently Hahn and Rogers studied the same set of compounds with another 3D QSAR method, the receptor surface model (RSM).[21] They built a regression equation based on two thermodynamic variables (eq 6):

$$\mathrm{pIC}_{50} = 3.762 + 0.296 \times \langle -10.203 - E_{\mathrm{interact}} \rangle + 0.089 \times \langle 26.855 - E_{\mathrm{inside}} \rangle \quad (6)$$

where E_{interact} is the sum of van der Waals and electrostatic interaction energies between the inhibitor and the pseudoreceptor and E_{inside} is the intramolecular energy of the inhibitor within the receptor environment. The broken brackets (⟨ ⟩) denote a spline function which returns the value of the argument if it is positive, and 0 otherwise. The use of splines for the two energy terms is a computationally inexpensive way to introduce nonlinearity in the equation. This QSAR model yielded a q^2 of 0.79.

We utilized the set of DβH inhibitors that was distributed as an example for the RSM module[21,40] in the Cerius2 package. The prealigned coordinates were used without modification to obtain direct comparison with the RSM result. Application of GNN on individual similarity matrices led to a four-descriptor electrostatic model ($q^2 = 0.64$) and a six-descriptor shape model ($q^2 = 0.76$). GNN on the combined matrix led to the selection of one electrostatic and three shape similarity descriptors. This mixed model, despite the use of fewer descriptors, was marginally more predictive ($q^2 = 0.77$; Figure 1b) than the shape model. Overall, the predictive statistics obtained by SM/GNN were comparable to the results of the MSA or RSM methods.

Unlike the Ah data set, the molecular alignment for the DβH inhibitors was well-defined though the individual conformations used for the similarity calculation could be ambiguous. Specifically, an asymmetric substitution pattern on the phenyl ring might result in two rotamers that were similar in conformational energy. In a forthcoming study, we will describe an alternative set of conformations that results in a superior QSAR model.

(c) BzR Data Set. Inverse agonists and antagonists of the benzodiazepine receptor (BzR) have been the subject for a number of 3D QSAR investigations.[15,17,22−24] The current data set was first studied by Allen et al., who reported a CoMFA model involving 37 compounds (**1**−**37**).[22] A few years later, the same research group refined their CoMFA model by incorporating a robust variable selection routine, GOLPE.[23] They tested both the original and the improved QSAR models with six additional analogues (**38**−**43**). In addition, Kroemer et al. used this data set to study the effect of electrostatic parameters on the quality of CoMFA models.[24] Good et al. performed PLS on similarity matrices derived from this data set and obtained some predictive QSARs.[15] Recently, the same set of compounds was examined in the first application of the comparative molecular moments analysis (CoMMA).[17] The key results in the literature are summarized in Table 2c.

The coordinates for these compounds were constructed using the Cerius2 program as described in Method. The compounds were superimposed using the alignment rule specified in the original CoMFA study.[22] GNN applications on individual similarity matrices indicated that, in accordance with the CoMFA results, the complement of shape between the ligand and its receptor was the major determinant of binding (Table 2c). The electrostatic interactions appeared to play a lesser role, though such description was needed to build a QSAR model with optimal predictivity. The top-ranking GNN model that was derived from the combined matrix gave a q^2 of 0.73 (Figure 1c), which was significantly better than those obtained using conventional CoMFA and SM/PLS (i.e., without GOLPE variable selection) or CoMMA methods. Finally, as an external test for the final QSAR, activity predictions were made with six additional compounds reported by Allen et al.[23] The experimental and predicted pIC_{50} values of these compounds are listed in Table 3. The rms prediction error was 0.3. Given that the range of activity for the training set spanned 6.4 log units, the predictions of the test compounds seemed very accurate.

The success of the use of GOLPE in conjunction with the other methods has brought fresh insight on its potential application to the current SM/GNN paradigm. In certain way, GNN and GOLPE have a similar philosophy. Both are feature selection routines that eliminate some of the less relevant input data. This contributes a better signal-to-noise and in general improves the QSAR quality. The utility of GOLPE in the preprocessing of the molecular fields is a promising direction that will be explored in future work.

Table 2. Statistical Data for Series a−h[a]

	combine	electrostatic	shape
(a) Ah			
CoMFA[18]	0.72 (6)	0.66 (8)	0.72 (9)
AV/NN[19]	0.83 (12)		
SM/GNN	0.85 (7)	0.72 (7)	0.85 (6)
(b) DβH Inhibitors			
MSA[20]	0.76 (6)		
RSM[21]	0.79 (2)		
SM/GNN	0.77 (4)	0.65 (4)	0.76 (6)
(c) BzR Inverse Agonists			
CoMFA[22]	0.59 (4)	16%	84%
CoMFA[23]	0.65 (4)	11%	89%
CoMFA/GOLPE[23]	0.82 (5)	48%	52%
CoMFA[24]	0.67 (4)[b]	13%	87%
SM/PLS[15]	0.69 (4)	0.59 (5)	0.60 (3)
SM-GOLPE/PLS[15]	0.72 (3)		
CoMMA[17]	0.39 (2)		
SM/GNN	0.73 (4)	0.61 (5)	0.71 (7)
(d) AChE Inhibitors			
CoMFA[25]	0.62 (3)	23%	77%
CoMFA/q^2-GRS[25]	0.73 (7)	33%	67%
SM/GNN	0.80 (9)	0.57 (3)	0.81 (9)
(e) Bisamidines			
LR I[26]	0.63 (8)		
LR II[26]	0.51 (3)		
SM/GNN	0.80 (6)	0.60 (6)	0.81 (7)
(f) GP Inhibitors			
SM/GNN	0.82 (5)	0.72 (6)	0.81 (6)
(g) Benzoic Acids			
LR[13]		0.91 (1)[c]	
CoMFA[16]		0.89 (6)	
DM/PLS[16]		0.90 (8)	
HSM/PLS[16]		0.75 (2)	
CSM/PLS[16]		0.75 (2)	
CoMMA[17]	0.69 (13)		
SM/GNN		0.83 (4)	0.34 (4)
(h) Imidazoles[d]			
SM/PLS[15]		0.63 (3)	
SM-GOLPE/PLS[15]		0.77 (3)	
CoMMA[17]	0.70 (2)		
SM/GNN		0.92 (3)	0.21 (3)

[a] This shows the q^2 and the associated number of descriptors (i.e., terms or components in linear or PLS regressions, number of descriptors in GNN; shown in parentheses) used in the correlation. For some CoMFA studies, the relative contributions of the electrostatic and steric fields in the PLS are also listed (in percentages). [b] Kroemer et al. have reported many CoMFA models using different charge types and cutoff schemes. The result shown in the table corresponds to the predictivity of a system closest to the one used in the electrostatic similarity calculations (C1bn in Table VIII of Kroemer et al.).

(d) AChE Data Set. Recently Cho et al. have reported a 3D QSAR study for the 60 structurally diverse AChE inhibitors using the CoMFA method.[25] The structures were aligned using the geometry established by a few homogeneous ligands whose X-ray structures with AChE were known. They demonstrated that the quality of the CoMFA model could be substantially improved by making appropriate region selections using a newly developed q^2-GRS routine.[25,41,42] They reported a seven-component PLS model with a cross-validated r^2 of 0.73 and concluded that the steric field was the major contributor of the PLS regression (Table 2d). The aligned coordinates of the AChE inhibitors used in the current work were kindly provided by Cho et al.

The principal role of the steric field was also apparent in the similarity approach. Applying GNN on the shape matrix alone resulted in a very good QSAR that had predictive statistics ($q^2 = 0.81$) already exceeding the CoMFA/q^2-GRS work. The electrostatic counterpart, used by itself, yielded a much worse correlation ($q^2 = 0.57$). Combining the shape and electrostatic information did not improve the model predictivity in this case ($q^2 = 0.80$; Figure 1d); in fact, all nine GNN descriptors chosen from the combined matrix were shape similarity descriptors. This suggested that the electrostatic information only added noise to the input matrix, and subsequently the effectiveness of the genetic search was slightly impaired.

Table 3. Activity Predictions for the Six Additional Inverse Agonists of BzR Using the GNN QSAR Derived from the Combined Similarity Matrix

pIC_{50}	**38**	**39**	**40**	**41**	**42**	**43**
experimental	6.3	7.0	6.3	7.0	6.0	6.5
predicted	6.8	6.8	6.7	6.8	6.3	6.7

(e) Bisamidine Data Set. Montanari et al. reported a number of linear regression (LR) equations which correlated the potency of 37 bisamidine analogues against *Leishmania*.[26] Their initial equation involved six indicator variables and two physicochemical properties and yielded a q^2 value of 0.63 (LR I in Table 2e). Later, they explored the possibility of replacing the indicator variables by topological or similarity descriptors. The three descriptors in their final model were log *P*, an electrotopological state index, and a Carbó similarity index using the least active compound (**4**) as the reference structure. The similarity index had equal contribution from shape and electrostatic attributes, though further studies suggested that shape was the more important component. The final regression model yielded a q^2 value of 0.51 (LR II in Table 2e).

The set of prealigned coordinates of bisamidines, which attained an isohelical conformation that was believed to be the bioactive form,[43,44] was kindly provided by Montanari et al. The dominance of shape factor was confirmed by the GNN analysis, where the predictivity of the shape model exceeds that of the electrostatic by over 0.2 q^2 unit. As with the previous AChE study, GNN on the combined matrix yielded a shape-only model ($q^2 = 0.80$; Figure 1e) that had no improvement in performance over the GNN model derived from the shape matrix.

(f) GP Data Set. We examined a set of 30 α-D-glucose derivatives whose cocrystallized structures with GP had been determined by Johnson et al. This set included some highly potent inhibitors that were discovered recently.[45,46] We took advantage of the X-ray alignment since the bioactive conformations for these ligands were known.

The optimal GNN QSARs from the electrostatic and shape similarity matrices were both six-descriptor models, and they yielded q^2 values of 0.72 and 0.81, respectively. Application of the GNN algorithm on the combined matrix yielded a simpler five-descriptor model that gave a q^2 value of 0.82 (Figure 1f). The three new potent compounds (**23**, **26**, and **30**) fit well in the current model. Their removals from the GNN analysis in fact made the resulting model marginally less predictive ($q^2 = 0.79$).

The predictivity of the SM/GNN QSAR was comparable to the models obtained in previous 3D QSAR studies[4,12] on different sets of glucose analogues. Using the GRID[47] force field and the GOLPE[48] selection on the field values, the researchers had reported PLS models with values of q^2 ranging from 0.76 to 0.81.

Prediction of Physicochemical Parameters. (g) Hammett Constants of Substituted Benzoic Acids. The Hammett constant (σ) is one of the most common electrostatic parameters in traditional 2D QSAR. It is derived from the ionization constants of substituted benzoic acids and reflects the inherent polar effect of a given substituent relative to hydrogen.[49] Numerous compilations of the Hammett constants are available for a large variety of substituents.[50]

In an earlier analysis of a series of 27 benzoic acids, Sotomatsu et al. observed that σ was linearly correlated with the sum of partial atomic charges of the two oxygen atoms and the hydrogen atoms of the carboxylic group.[51] Kim and Martin extended this work with 49 meta- or para-substituted benzoic acids.[13] Because they did not report the q^2 value of the model, we reanalyzed the equation based on the published partial atomic charges. The one-descriptor LR was highly predictive and yielded a q^2 value of 0.91. This work was followed by a number of structure−properties studies, all of which involved PLS analyses on matrices that were derived from electrostatic field[13,15,16] or molecular moments (CoMMA; $q^2 = 0.69$).[17] Here, we cite the set of results reported by Martin et al., who have performed PLS on the field values (CoMFA) and two types of similarity matrices (SM/PLS) as well as a distance matrix (DM/PLS).[16] Their six-component CoMFA model yielded a q^2 value of 0.89. Performing PLS on either the Hodgkin (HSM) or the Carbó (CSM) similarity matrices both led to a two-component model with a q^2 value 0.75, though with a distance matrix a very predictive eight-component model was obtained ($q^2 = 0.90$).

The GNN method was applied to the electrostatic and shape similarity matrices for the 49 benzoic acids (**1**−**49**). The fact that σ predominantly conveys electrostatic information was clearly reflected by the much better correlation obtained from the electrostatic matrix ($q^2 = 0.83$) relative to the shape matrix ($q^2 = 0.34$). A further test for the predictive ability of the electrostatic QSAR was to predict σ of some analogues that were not used in model construction (**50**−**72**). Most of the predicted σ values for 23 additional substituted benzoic acids were very accurate; two of the largest prediction errors came from the test compounds that had substantially lower σ values than any example in the training set (Figure 1g). The rms error and the correlation coefficient (r^2) for the test set predictions were 0.19 and 0.81, respectively.

On an absolute scale, the cross-validated statistics and the test set predictions provided by the electrostatic SM/GNN model were very good. On relative terms, the present result was superior to the CoMMA and SM/PLS benchmarks but was inferior to the LR, CoMFA, and DM/PLS models. The significant improvement achieved by the use of a distance-based matrix along with the possibility of a better DM/GNN QSAR was intriguing, though more studies are needed to determine whether such improvement is general or specific to this data set.

(h) pK_a of Imidazoles. The GNN method was employed for the prediction of the pK_a values of 16 imidazoles. The QSAR based on the shape similarity matrix gave a very weak correlation ($q^2 = 0.21$), whereas the electrostatic matrix yielded an excellent result ($q^2 = 0.92$; Figure 1h). The electrostatic model was substantially superior to the corresponding SM/PLS models (with or without GOLPE field selection) and the CoMMA results.

Table 4. Significant Results for the Eight Data Sets[a]

data set	N	n	r^2	q^2	randomization test q^2
(a) Ah	73	7	0.89	0.85	0.14 ± 0.07
(b) DβH	47	4	0.80	0.77	0.14 ± 0.12
(c) BzR	37	4	0.83	0.73	0.17 ± 0.10
(d) AChE	60	9	0.86	0.80	0.16 ± 0.10
(e) bisamidines	37	6	0.86	0.80	0.33 ± 0.11
(f) GP	30	5	0.90	0.82	0.31 ± 0.20
(g) benzoic acids	49	4	0.88	0.83	0.04 ± 0.09
(h) imidazoles	16	3	0.96	0.92	−0.26 ± 0.33

[a] N is the number of training compounds, n is the number of GNN descriptors in the model, r^2 is the Pearson correlation coefficient for the training set, and q^2 is the correlation coefficient for the cross-validated predictions. The q^2 values for the randomization test are derived from 20 multiple runs using data sets with differently scrambled output.

IV. Conclusion

The biological activities of six molecular series and the physicochemical properties of two molecular series have been correlated using the new SM/GNN approach. The predictive statistics of GNN models from individual shape or electrostatic SMs provide a qualitative measure of the relative importance of the two effects. A composite GNN model that incorporates both properties, which is sometimes necessary for optimal predictivity, can be obtained from the combined similarity matrix. The major results of this study are summarized in Table 4. The cross-validated statistics and test set predictions of the QSAR models are generally very good. The q^2 values range from 0.73 to 0.85 for the seven larger data sets and exceed 0.90 for the imidazoles. Their significance is validated by a standard randomization test,[52] which suggests that the results are unlikely to be a chance correlation.[53] Moreover, most of the SM/GNN results compare favorably with the benchmarks obtained by the state-of-the-art QSAR methods. The consistency in performance is very encouraging given the great structural variety in the different data sets. This demonstrates the usefulness of similarity as a descriptor and the robust nature of GNN as a correlation tool.

The gain in predictivity obtained by replacing PLS with GNN for the analysis of SM comes at a higher computational cost, though the SM/GNN calculations are not expensive by current standards. The generation of the field and similarity matrices for a data set containing, say, 50 drug-size (i.e., 20−40 atoms) molecules typically takes a few minutes on a modest workstation (e.g., 175-MHz R4400 Silicon Graphics Indigo2). A GNN simulation on the resulting matrix requires approximately 1−2 CPU hours. Furthermore, once the model has been constructed, activity predictions of new analogues can be made very rapidly with a trained neural network. This fast processing time makes it particularly suitable for screening a large number of potential drug candidates that may be obtained from structure-based design methods, database searches, or combinatorial libraries. This aspect of application is an impetus for the development of this method.

A very useful feature of CoMFA is the depiction of the important regions where molecular interactions may take place. This is possible because CoMFA used linear combinations of field values as descriptors and their spatial reference to the molecule remains intact. This is not the case for similarity-based descriptors because the spatial reference of the field is destroyed when the similarity index is calculated. In this regard, a region selection regime that is similar to the CoMFA/q^2-GRS routine may be useful. In the procedure, the full grid is initially divided into a number of subgrids that may contain parts of the molecules. The GNN method is applied to each of the SMs derived from individual subgrids. The important molecular regions are identified by the locations of the subgrids that can generate good SM/GNN models. By providing a focus for possible structural modifications, the drug optimization process may be expedited.

Finally, the current method is no exception to the fact that molecular alignment is an inherent, and often critical, element of many 3D QSAR methods. Two methodological extensions have been considered to improve this aspect of the problem. The first involves a GA-based alignment technique that is used in conjunction with molecular similarity. Preliminary work suggests that significant improvements in predictivity can be obtained with alternative alignments.[54] The second concerns some novel similarity indices that are derived from spatially invariant properties rather than standard molecular fields. For example, one can obtain a similarity measure based on the elements of the autocorrelation vectors proposed by Wagener et al. or the molecular moments descriptors used in CoMMA. This new variant of SM/GNN does not require molecular superposition; the approach will be addressed more thoroughly in future investigations.

The results of this study suggest that the SM/GNN method is a general, efficient, and robust approach to obtain 3D QSAR with good correlative and predictive statistics for a variety of chemical classes and properties.

Acknowledgment. We are grateful to several research groups for providing the coordinates of some of the data sets used in this study. They include Dr. K. Watson and Prof. L. Johnson for the glycogen phosphorylase inhibitors, Dr. S.-J. Cho and Prof. A. Tropsha for the acetylcholinesterase inhibitors, and Dr. C. Montanari and Dr. M. Tute for the bisamidines. We thank Molecular Simulations Inc. for the software support. This work is supported in part by a GAOLI grant supplement from the National Science Foundation.

References

(1) So, S.-S.; Karplus, M. Three-dimensional quantitative structure–activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. *J. Med. Chem.* **1997**, *40*, 4347–4359.
(2) So, S.-S.; Karplus, M. Evolutionary optimization in quantitative structure–activity relationship: an application of genetic neural network. *J. Med. Chem.* **1996**, *39*, 1521–1530.
(3) So, S.-S.; Karplus, M. Genetic neural networks for quantitative structure–activity relationships: improvements and application of benzodiazepine affinity for benzodiazepine/GABA$_A$ receptors. *J. Med. Chem.* **1996**, *39*, 5246–5256.
(4) Watson, K. A.; Mitchell, E. P.; Johnson, L. N.; Cruciani, G.; Son, J. C.; Bichard, C. J. F.; Fleet, G. W. J.; Oikonomakos, N. G.; Kontou, M.; Zographos, S. E. Glucose analogue inhibitors of glycogen phosphorylase: from crystallographic analsysis to drug prediction using GRID force-field and GOLPE variable selection. *Acta Crystallogr.* **1995**, *D51*, 458–472.
(5) Martin, J. L.; Johnson, L. N.; Withers, S. G. Comparison of the binding of glucose and glucose 1-phosphate derivatives to T-state glycogen phosphorylase b. *Biochemistry* **1990**, *29*, 10745–10757.
(6) Martin, J. L.; Veluraja, K.; Ross, K.; Johnson, L. N.; Fleet, G. W.; Ramsden, N. G.; Bruce, I.; Orchard, M. G.; Oikonomakos, N. G.; Papageorgiou, A. C. Glucose analogue inhibitors of glycogen phosphorylase: the design of potential drugs for diabetes. *Biochemistry* **1991**, *30*, 10101–10116.
(7) Johnson, L. N.; Snape, P.; Martin, J. L.; Acharya, K. R.; Barford, D.; Oikonomakos, N. G. Crystallographic binding studies on the allosteric inhibitor glucose-6-phosphate to T state glycogen phosphorylase b. *J. Mol. Biol.* **1993**, *232*, 253–267.
(8) Oikonomakos, N. G.; Kontou, M.; Zographos, S. E.; Tsitoura, H. S.; Johnson, L. N.; Watson, K. A.; Mitchell, E. P.; Fleet, G. W.; Son, J. C.; Bichard, C. J.; et al. The design of potential antidiabetic drugs: experimental investigation of a number of beta-D-glucose analogue inhibitors of glycogen phosphorylase. *Eur. J. Drug Metab. Pharmacokinet.* **1994**, *19*, 185–192.
(9) Oikonomakos, N. G.; Kontou, M.; Zographos, S. E.; Watson, K. A.; Johnson, L. N.; Bichard, C. J.; Fleet, G. W.; Acharya, K. R. *N*-Acetyl-beta-D-glucopyranosylamine: a potent T-state inhibitor of glycogen phosphorylase. A comparison with alpha-D-glucose. *Protein Sci.* **1995**, *4*, 2469–2477.
(10) Watson, K. A.; Mitchell, E. P.; Johnson, L. N.; Son, J. C.; Bichard, C. J.; Orchard, M. G.; Fleet, G. W.; Oikonomakos, N. G.; Leonidas, D. D.; Kontou, M. Design of inhibitors of glycogen phosphorylase: a study of alpha- and beta-C-glucosides and 1-thio-beta-D-glucose compounds. *Biochemistry* **1994**, *33*, 5745–5758.
(11) Oikonomakos, N. G.; Zographos, S. E.; Johnson, L. N.; Papageorgiou, A. C.; Acharya, K. R. The binding of 2-deoxy-D-glucose 6-phosphate to glycogen phosphorylase b: kinetic and crystallographic studies. *J. Mol. Biol.* **1995**, *254*, 900–917.
(12) Cruciani, G.; Watson, K. A. Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. *J. Med. Chem.* **1994**, *37*, 2589–2601.
(13) Kim, K. H.; Martin, Y. C. Direct prediction of dissociation constants (pK_a's) of clonidine-like imidazolines, 2-substituted imidazoles, and 1-methyl-2-substituted-imidazoles from 3D structures using a comparative molecular field analysis (CoMFA) approach. *J. Med. Chem.* **1991**, *34*, 2056–2060.
(14) Kim, K. H.; Martin, Y. C. Direct prediction of linear free energy substituent effects from 3D structures using comparative molecular field analysis. 1. Electronic effects of substituted benzoic acids. *J. Org. Chem.* **1991**, *56*, 2723–2729.
(15) Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. *J. Med. Chem.* **1993**, *36*, 2929–2937.
(16) Martin, Y. C.; Lin, C. T.; Hetti, C.; DeLazzer, J. PLS analysis of distance matrices to detect nonlinear relationships between biological potency and molecular properties. *J. Med. Chem.* **1995**, *38*, 3009–3015.
(17) Silverman, B. D.; Platt, D. E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* **1996**, *39*, 2129–2140.
(18) Waller, C. L.; McKinney, J. D. Comparative molecular field analysis of polyhalogenated dibenzo-*p*-dioxins, dibenzofurans, and biphenyls. *J. Med. Chem.* **1992**, *35*, 3660–3666.
(19) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of molecular surface properties for modeling *corticosteroid* binding globulin and cytosolic *Ah* receptor activity by neural networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.
(20) Burke, B. J.; Hopfinger, A. J. 1-(Substituted-benyzl)imidazole-2(3*H*)-thione inhibitors of dopamine *β*-hydroxylase. *J. Med. Chem.* **1990**, *33*, 274–281.
(21) Hahn, M.; Rogers, D. Receptor surface models. 2. Application to quantitative structure–activity relationships studies. *J. Med. Chem.* **1995**, *38*, 2091–2102.
(22) Allen, M. S.; Tan, Y.-C.; Trudell, M. L.; Narayanan, K.; Schindler, L. R.; Martin, M. J.; Schultz, C.; Hagen, T. J.; Koehler, K. F.; Codding, P. W.; Skolnick, P.; Cook, J. M. Synthetic and computer-assisted analyses of the pharmacophore for the benzodiazepine receptor inverse agonist site. *J. Med. Chem.* **1990**, *33*, 2343–2357.
(23) Allen, M. S.; LaLoggia, A. J.; Dorn, L. J.; Martin, M. J.; Costantino, G.; Hagen, T. J.; Koehler, K. F.; Skolnick, P.; Cook, J. M. Predictive binding of beta-carboline inverse agonists and antagonists via the CoMFA/GOLPE approach. *J. Med. Chem.* **1992**, *35*, 4001–4010.
(24) Kroemer, R. T.; Hecht, P.; Liedl, K. R. Different electrostatic descriptors in comparative molecular field analysis: a comparison of molecular electrostatic and Coulomb potentials. *J. Comput. Chem.* **1996**, *17*, 1296–1308.
(25) Cho, S. J.; Garsia, M. L. S.; Bier, J.; Tropsha, A. Structure-based alignment and comparative molecular field analysis of acetylcholinesterase inhibitors. *J. Med. Chem.* **1996**, *39*, 5064–5071.

(26) Montanari, C. A.; Tute, M. S.; Beezer, A. E.; Mitchell, J. C. Determination of receptor-bound drug conformations by QSAR using flexible fitting to derive a molecular similarity index. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 67−73.

(27) Cerius2, Version 2.0; Molecular Simulations Inc., San Diego, CA.

(28) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 5832−5842.

(29) Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.

(30) Hodgkin, E. E.; Richards, W. G. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1987**, *14*, 105−110.

(31) Meyer, A. M.; Richards, W. G. Similarity of molecular shape. *J. Comput.-Aided Mol. Des.* **1991**.

(32) Kubinyi, H.; Abraham, U. Practical problems in PLS analyses. *In* 3D *QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers B.V.: Leiden, The Netherlands, 1993; pp 717−728.

(33) Andrea, T. A.; Kalayeh, H. Applications of neural networks in quantitative structure−activity relationships of dihydrofolate reductase inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.

(34) So, S.-S.; Richards, W. G. Application of neural networks: quantitative structure−activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl)pyrimidines as DHFR inhibitors. *J. Med. Chem.* **1992**, *35*, 3201−3207.

(35) Manallack, D. T.; Ellis, D. D.; Livingstone, D. J. Analysis of linear and nonlinear QSAR data using neural networks. *J. Med. Chem.* **1994**, *37*, 3758−3767.

(36) Bandiera, S.; Safe, S.; Okey, A. B. Binding of polychlorinated biphenyls classified as either phenobarbitone-, 3-methylcholanthrene- or mixed-type inducers to cytosolic *Ah* receptor. *Chem.-Biol. Interact.* **1982**, *39*, 259−277.

(37) Safe, S. H. Comparative toxicology and mechanism of action of polychlorinated dibenzo-p-dioxins and dibenzofurans. *Annu. Rev. Pharmacol. Toxicol.* **1986**, *26*, 371−399.

(38) Safe, S. H. Polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs), dibenzofurans (PCDFs), and related compounds: environmental and mechanistic considerations which support the development of toxic equivalency factors (TEFs). *Crit. Rev. Toxicol.* **1990**, *21*, 51−88.

(39) Dammkoeler, R. A.; Karasek, S. F.; Shands, E. F. B.; Marshall, G. R., Constrained search of conformational hyperspace. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 3−21.

(40) Hahn, M. Receptor surface models. 1. Definition and construction. *J. Med. Chem.* **1995**, *38*, 2080−2090.

(41) Cho, S. J.; Tropsha, A., Cross-validated R^2-guided region selection for comparative molecular field analysis: a simple method to achieve consistent results. *J. Med. Chem.* **1995**, *38*, 1060−1066.

(42) Cho, S. J.; Tropsha, A.; Suffness, M.; Cheng, Y.-C.; Lee, K.-H. Antitumor agents. 163. Three-dimensional quantitative structure−activity relationship study of 4′-*O*-demethylepipodophyllotoxin analogues using the modified CoMFA/q^2-GRS approach. *J. Med. Chem.* **1996**, *39*, 1383−1395.

(43) Lowe, P. R.; Sansom, C. E.; Schwalbe, C. H.; Stevens, M. F. G. Crystal structure and molecular modelling of the antimicrobial drug pentamidine. *J. Chem. Soc., Chem. Commun.* **1989**, 1164−1165.

(44) Donkor, I. O.; Tidwell, R. R.; Jones, S. K. Pentamidine congeners. 2. 2-Butene-bridged aromatic diamidines and diimidazolines as potential anti-*Pneumocystis carinii* pneumonia agents. *J. Med. Chem.* **1994**, *37*, 4554−4557.

(45) Bichard, C. J. F.; Mitchell, E. P.; Wormald, M. R.; Watson, K. A.; Johnson, L. N.; Zographos, S. E.; Koutra, D. D.; Oikonomakos, N. G.; Fleet, G. W. J. Potent inhibition of glycogen phosphorylase by a spirohydantoin of glucopyranose: first pyranose analogues of hydantocidin. *Tetrahedron Lett.* **1995**, *36*, 2145−2148.

(46) Krülle, T. M.; Watson, K. A.; Gregoriou, M.; Johnson, L. N.; Crook, S.; Watkin, D. J.; Griffiths, R. C.; Nash, R. J.; Tsitsanou, K. E.; Zographos, S. E.; Oikonomakos, N. G.; Fleet, G. W. J. Specific inhibition of glycogen phosphorylase by a spirodiketopiperazine at the anomeric position of glucopyranose. *Tetrahedron Lett.* **1995**, *36*, 8281−8294.

(47) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important molecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(48) Cruciani, G.; Clementi, S. GOLPE: philosophy and applications in 3D QSAR. In *Advanced Computer-Assisted Techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH Publishers, Inc.: New York, 1994; Vol. 3, pp 61−88.

(49) Jurs, P. C.; Dixon, S. L.; Egolf, L. M. Molecular concepts: representations of molecules. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH Publishers, Inc.: New York, 1995; Vol. 2, pp 15−38.

(50) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; John Wiley & Sons, Inc.: New York, 1979.

(51) Sotomatsu, T.; Murata, Y.; Fujita, T. Correlation analysis of substituent effects on the acidity of benzoic acids by the AM1 method. *J. Comput. Chem.* **1989**, *10*, 94−98.

(52) Wold, S.; Eriksson, L. Statistical validation of QSAR results. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH Publishers, Inc.: New York, 1995; Vol. 2, pp 309−318.

(53) Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure−activity relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(54) So, S.-S.; Karplus, M. Manuscript in preparation.

JM970488N