Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies

David T. Stanton†

Procter & Gamble Pharmaceuticals, Health Care Research Center, 8700 Mason-Montgomery Road, Mason, Ohio 45040-9462

Received June 18, 1998

The recently developed BCUT metrics of Pearlman were evaluated to determine their utility as measures of molecular structure in quantitative structure−activity relationship (QSAR) and quantitative structure−property relationship (QSPR) studies. These metrics have been found to provide unique information regarding molecular structure and have been found to make significant contributions to resulting equations. The importance of these descriptors is illustrated in a study of inhibitors of dihydrofolate reductase and in a study of the normal boiling points of a diverse set of polar heterocyclic compounds.

INTRODUCTION

The process of developing a meaningful quantitative relationship between the molecular structure of a compound and the properties that it exhibits (both biological and physicochemical) is strongly dependent on how the molecular structure is measured or described. One can employ the most sophisticated statistical techniques and have obtained the most carefully measured property data for a set of compounds but can still fail to find an adequate QSAR/QSPR model because of the lack of necessary structural information. Given the importance of capturing the subtleties of molecular structure to QSAR/QSPR studies, the search for new measures of structural features is a continuing process. There is a wide variety of different measures or *descriptors* of molecular structure. For example, the molecular connectivity parameters of Randić,[1] and later of Kier and Hall,[2,3] provide a great quantity of information concerning the topology of the molecular (size, shape, degree of branching). There are numerous examples of descriptors that capture information concerning the geometry of the molecule.[4,5] The electronic features of the molecule are the focus of a number of other useful descriptors.[6] More recently, sets of whole-molecule descriptors have been reported that combine two or more measures of atom-based properties into a single value. These hybrid molecular descriptors are often found to be valuable in developing good quality QSAR and QSPR models. Examples of the hybrid parameters are the CPSA descriptors,[7] the related hydrogen-bonding CPSA descriptors,[8] and the hydrogen-bonding descriptor of Katritzsky.[9] One of the newest additions to this class of whole-molecule descriptors is the set of BCUT metrics of Pearlman.[10]

The BCUT metrics are extensions of parameters originally developed by Burden.[11] The Burden parameters are based on a combination of the atomic number for each atom and a description of the nominal bond-type for adjacent and nonadjacent atoms. The BCUT metrics expand the number and types of atomic features that can be considered and also provide a greater variety of proximity measures and weighting schemes. The result is a new, whole-molecule descriptor that has shown significant utility in the measurement of molecular diversity and related tasks. Only one report of the use of BCUT metrics in a QSAR or QSPR study has been found.[10] The details for that model were not reported, so it is difficult to assess the true utility of the BCUTs in that analysis. Also, the BCUTs selected for inclusion in the model were selected for the purpose of molecular diversity measurements and not to explain the observed variation in the property in question, so a true evaluation in a QSAR context was not possible. Since the BCUTs appeared to capture sufficient structural detail to yield very useful results in molecular diversity-related tasks, it was of interest to fully evaluate their utility in modeling biological and physicochemical properties.

Two separate studies are reported here. The first is an analysis of a set of substituted benzyl pyrimidine bacterial dihydrofolate reductase (DHFR) inhibitors. This data set was selected because the target involved a specific enzyme−inhibitor interaction. The second study involved the modeling of the normal boiling points (boiling points determined at 760 mmHg) of a large set of diverse and relatively polar organic heterocycles (thiophenes, furans, tetrahydrofurans). This latter data set was selected because it involved a bulk phase property and would require the model to incorporate a greater scope of structural information due to the diversity of the structures involved.

The importance of the BCUT metrics in the QSAR and QSPR models was assessed in two ways. The first was to examine the contribution of individual BCUTs using partial least-squares regression analysis. The loadings of the BCUTs in the latent variables provide a quantitative measure of their contribution. The second assessment involved a determination of the overall predictive strength of the models using external prediction (test) data sets. This provides a way to determine the real-world utility of the models containing BCUT metrics and resolves any questions concerning the possibility of chance correlations occurring. The results of the combined tests should provide a clear demonstration of the utility of the BCUTs in QSAR and QSPR applications.

† E-mail address: stanton@pg.com.

10.1021/ci980102x CCC: $18.00
Published on Web 12/17/1998

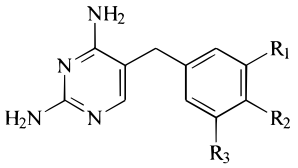

Figure 1. General structural diagram for the substituted-benzyl pyrimidine data set. The substituents at R_1, R_2, and R_3 are described in ref 12.

EXPERIMENTAL SECTION

The structures and biological data used in the study of the DHFR inhibitors were taken from work described by Hirst.[12] The structures of the 74 substituted benzyl pyrimidines involved are available in the original reference. A general structure diagram is provided in Figure 1. The biological data ($\log(1/K_i)$) for the compounds is provided in Table 1. The numbers identifying the structures in this study are consistent with those used by Hirst. One correction was made to the structure data provided in the original work. Compound **38** was corrected to read as *3,4(*$OCH_2CH_2OCH_3)_2$. The structural and boiling point data for the 179-observation boiling point data set used here were received from the Beilstein Institute as part of a previous study.[8] The compounds are identified using Beilstein Registry Numbers (BRN). The normal boiling point data (boiling point at standard pressure, in units of °C) for these compounds are provided in Table 2. All computations were carried out using a Silicon Graphics Indigo-II workstation, with Extreme Graphics, running under the IRIX operating system (version 6.2). The BCUT values were computed using the DiverseSolutions software package (R. S. Pearlman, University of Texas at Austin, Version 3.0.2). All other structural descriptors were calculated using the ADAPT package.[13,14] Model development was also accomplished using ADAPT. Additional model diagnostics were computed using Minitab for Windows (Release 11) and the SCAN (Release 1.1) chemometrics package (both from Minitab, Inc., State College, PA).

Structure Entry. All the structures were sketched into the computer manually and stored in a database using SYBYL Ver. 6.3. Three-dimensional coordinates for each structure were obtained by first applying the CONCORD program,[15] followed by strain-energy minimization using the MAXIMIN program[16] (both part of the SYBYL package). Minimization was conducted with the inclusion of electrostatic terms and using the Gasteiger−Huckel partial atomic charges.[17] The structures and associated partial atomic charge data were then transferred to the ADAPT software package. The structures were also exported from SYBYL in the form of an MACCS SD file[18] for transfer to the DiverseSolutions package.

Calculation of the BCUT Metrics. The set of 85 separate BCUT metrics was computed for each of the structures in both data sets. The metrics calculated were those using nominal bond order values (so-called *Burden numbers*) and were calculated with and without solvent-accessible surface area weighting. These were then exported using the BMF2USER program from the DiverseSolutions package and assembled into a spreadsheet in Excel (Microsoft, Inc., Office-97 version). The BCUTs were then exported as ASCII text tables to be include in the ADAPT data files. For reasons of simplicity, the BCUT metrics are referred to generically in this text using sequential numbers (e.g., BCUT-24), and the details of the metric are provided where needed.

Calculation of the ADAPT Descriptors. Once the structures were imported and stored in the ADAPT data files, a set of 105 whole-molecule descriptors was calculated for each of the structures in both data sets. The calculation of the ADAPT descriptors has been described previously.[19] The set of values calculated included topological, geometric, and electronic descriptor types. The CPSA descriptors, and the related hydrogen-bonding specific descriptors, were also included and were computed using the Gasteiger−Huckel partial atomic charges obtained using SYBYL. The BCUT metrics for each respective data set were then imported yielding a total descriptor set of 190 values for each structure.

Selection of the External Prediction (Test) Sets. To provide an assessment of the true predictive strength of resulting models, a small subset of the available structures were selected randomly and set aside. The DHFR external prediction set comprised 10 structures, and the boiling point prediction set contained 20 structures. These structures were not included in any stage of model development. All remaining structures were placed in the training set.

Objective Feature Selection. Once the descriptors were in place, a process of objective feature selection (OFS) was used to remove from consideration those descriptors that either show little variation over the data set or that are highly correlated with other descriptors. This step of the analysis is considered *objective* because the property data (DHFR inhibition or boiling point) is not considered. Any descriptor that showed 90% or greater of the values to be identical was dropped from further consideration. The correlation of all pairs of remaining descriptors was then examined. When any pair of descriptors exhibited a correlation coefficient of 0.90 or greater, one of the pair was eliminated from consideration. The selection of which of the pair to retain was based on the number of other descriptors to which they were highly correlated. This allowed for the consideration of the unique subset of the descriptors in the model development step. Evaluation of the correlation of all descriptors that remained following OFS was accomplished using the hierarchical cluster analysis, using the method of complete linkage and correlation coefficient distance, as provided in the Minitab package (Release 12).

Model Development. Models were generated from the reduced pool of descriptors obtained from OFS using generalized simulated annealing (GSA).[20] The size of the model (the count of descriptors) was initially determined based on an examination of the rate of decrease of the root-mean-squared (RMS) error obtained from GSA. Models were evaluated for statistical significance using the standard methods and were further evaluated using partial least squares (PLS) regression in the SCAN package.

DISCUSSION

DHFR Inhibition Study. Initial model development was begun using the entire 64-observation training set. Difficulties were noted involving two particular observations, compounds **01** and **10**. Both of these compounds are reported to be anomalies by Hirst. These were set aside, and work was continued using the remaining 62 observations. However, these two compounds will be addressed again later.

Table 1. Observed and Calculated DHFR Inhibition Data ($\log(1/K_i)$) for the 74-Observation Benzyl Pyrimidine Data Set

compound ID	set membership	observed activity ($\log(1/K_i)$)	fitted/predicted activity ($\log(1/K_i)$)	fit/prediction error ($\log(1/K_i)$)	compound ID	set membership	observed activity ($\log(1/K_i)$)	fitted/predicted activity ($\log(1/K_i)$)	fit/prediction error ($\log(1/K_i)$)
01	excluded	3.04	N/A	N/A	38	training	7.22	7.47	−0.25
02	training	5.60	5.99	−0.39	39	training	7.23	7.17	0.06
03	training	6.07	6.20	−0.13	40	training	7.69	8.07	−0.38
04	training	6.18	6.36	−0.18	41	training	7.72	7.52	0.20
05	training	6.20	6.11	0.09	42	ext pred	8.35	8.26	0.09
06	training	6.23	6.02	0.21	43	training	8.38	7.37	1.01
07	training	6.25	6.11	0.14	44	training	8.87	8.58	0.29
08	training	6.28	6.40	−0.12	45	training	7.56	8.21	−0.65
09	training	6.30	6.55	−0.25	46	training	7.74	8.25	−0.51
10	excluded	6.31	N/A	N/A	47	training	7.87	8.06	−0.19
11	ext pred	6.35	5.98	0.37	48	training	7.87	7.69	0.18
12	training	6.39	6.30	0.09	49	training	8.42	8.08	0.34
13	training	6.40	6.60	−0.20	50	training	8.57	8.53	0.04
14	training	6.45	6.77	−0.32	51	training	8.82	8.86	−0.04
15	training	6.46	6.79	−0.33	52	ext pred	8.82	8.77	0.05
16	ext pred	6.47	6.38	0.09	53	training	8.85	8.54	0.31
17	training	6.48	6.68	−0.20	54	training	8.87	8.38	0.49
18	training	6.53	6.79	−0.26	55	training	8.87	8.40	0.47
19	training	6.55	6.61	−0.06	56	training	6.45	6.32	0.13
20	training	6.57	6.44	0.13	57	training	6.60	6.70	−0.10
21	training	6.57	6.42	0.15	58	training	6.84	7.07	−0.23
22	training	6.59	6.94	−0.35	59	training	6.89	6.48	0.41
23	training	6.65	6.86	−0.21	60	training	6.93	7.02	−0.09
24	training	6.70	6.73	−0.03	61	training	7.04	7.17	−0.13
25	training	6.78	6.75	0.03	62	ext pred	7.13	7.11	0.02
26	training	6.82	6.80	0.02	63	training	7.16	7.05	0.11
27	ext pred	6.82	6.68	0.14	64	training	7.20	7.52	−0.32
28	training	6.82	6.67	0.15	65	ext pred	7.41	7.28	0.13
29	training	6.86	6.45	0.41	66	training	7.53	7.95	−0.42
30	training	6.89	6.48	0.41	67	ext pred	7.54	7.36	0.18
31	training	6.89	6.84	0.05	68	training	7.66	7.66	0.00
32	ext pred	6.92	6.80	0.12	69	training	7.71	7.68	0.03
33	training	6.93	6.82	0.11	70	training	7.77	7.45	0.32
34	training	6.96	6.97	−0.01	71	training	7.80	7.89	−0.09
35	training	6.97	6.98	−0.01	72	training	7.82	7.78	0.04
36	training	6.99	6.81	0.18	73	ext pred	7.94	7.76	0.18
37	training	7.02	6.92	0.10	74	training	8.18	8.44	−0.26

A set of 64 descriptors remained following OFS for the 62-observation training set. Included among these were 25 of the initial 85 BCUTs. Cluster analysis of the 64 descriptors provided an interesting comparison of the BCUT descriptors with the more conventional types of descriptors. The dendrogram showing the relationship of the descriptors is shown in Figure 2. A set of 10 clusters was obtained by placing the cut-point as shown in the graph. The cluster membership is described in Table 3. The BCUT descriptors tend to cluster primarily with other BCUTs. Among the conventional QSAR descriptors, the CPSA descriptors cluster with the BCUTs more often than do other descriptor types. When the BCUTs are clustered with descriptors such as topological indices and more general geometric descriptors, the similarity level at which these descriptors are combined is usually small and close to the cut-point (see the enlarged view of cluster-10 in Figure 2). The cluster analysis suggests that the BCUTs carry information that is more like that provided by the CPSA descriptors but that is also unique.

The set of 64 descriptors were used in model development using GSA, with each descriptor initially having equal probability of being included in the model. Based on the initial work, a model containing six descriptors was targeted. An examination of the 10 best models returned by GSA showed that all 10 included three or more BCUTs, and half included four BCUT metrics. The descriptor membership of the top 10 equations is summarized in Figure 3. Three particular BCUT metrics were included in all of the top 10 models. The details for the best model are provided in Table 4. The model yields a very good fit to the observed biological data ($R^2 = 0.871$). It was of interest to determine if the model could be improved by exchanging model descriptors with any of the descriptors previously set aside in the OFS step. An improved model ($R^2 = 0.878$) was obtained by exchanging two of the model descriptors. This new model is summarized in Table 5. The correlation of the estimated and observed DHFR activity values is illustrated graphically in Figure 4. An evaluation of the model shows the overall equation to be significant at the 95% confidence limit with an overall F-value for AOV[23] = 66.09, compared to the critical F-value of 2.27 (6,55, $\alpha = 0.05$). Each descriptor in the model is also significant at the 95% confidence limit with the minimum partial F-value[24] of 12.59 (compared to the same critical value as above). Two of the descriptors in the model show a relatively high pairwise correlation ($r = -0.847$). Since high collinearity among the descriptors of the model can cause it to be unstable, these descriptors warranted further consideration. The variance inflation factors (VIF)[25] are a measure of the degree to which the variance of the coefficients of the descriptors is inflated compared to their values given totally orthogonal descriptors. The VIF values for the model are provided in Table 5. Ideally, VIF values should be below 10 and have an average of about 1. The largest VIF value for this model is 6.5, with a mean

Table 2. Observed and Calculated Normal Boiling Data for the 179-Observation Data Set

BRN number	set membership	observed boiling point (°C)	fitted/predicted boiling point (°C)	fit/prediction error (°C)	BRN number	set membership	observed boiling point (°C)	fitted/predicted boiling point (°C)	fit/prediction error (°C)
1264	training	163.0	155.4	7.6	107055	training	155.1	151.4	3.7
1300	training	115.2	128.2	−13.0	107127	ext pred	160.1	167.4	−7.3
1760	training	172.6	175.5	−2.9	107162	training	165.0	163.4	1.6
2457	training	179.5	202.6	−23.1	107256	training	186.0	189.3	−3.3
2678	training	249.0	243.0	6.0	107512	training	215.8	204.2	11.6
3343	training	229.0	231.6	−2.6	107637	training	184.0	183.7	0.3
3929	training	300.0	295.9	4.1	107647	training	170.1	175.6	−5.5
7564	training	286.0	264.5	21.5	107812	ext pred	148.5	144.2	4.3
15949	training	216.0	221.4	−5.4	107813	training	186.0	187.2	−1.2
79863	training	182.0	195.3	−13.3	107829	ext pred	134.5	135.0	−0.5
80418	training	200.4	197.2	3.2	107910	training	213.5	211.7	1.8
80498	training	253.0	239.0	14.0	107941	training	202.8	209.3	−6.5
81943	training	286.0	303.1	−17.1	108034	training	137.0	134.9	2.1
81973	training	177.0	201.4	−24.4	108140	training	173.2	175.8	−2.6
82647	training	245.5	260.7	−15.2	108441	training	261.0	251.0	10.0
102392	training	120.0	111.0	9.0	108583	training	99.0	110.2	−11.2
102442	training	86.5	93.5	−7.0	108654	training	172.5	169.1	3.4
102449	training	133.1	125.6	7.5	108693	training	179.5	182.3	−2.8
102476	training	91.7	93.1	−1.4	108825	training	196.0	184.0	12.0
102545	training	181.0	164.8	16.2	108880	training	169.0	168.3	0.7
102560	training	108.6	114.4	−5.8	108894	training	226.5	218.8	7.7
102605	training	120.3	114.2	6.1	108898	training	133.5	129.5	4.0
102619	training	115.0	109.8	5.2	108947	training	171.2	174.3	−3.1
102623	training	184.1	170.2	13.9	109091	ext pred	179.5	185.0	−5.5
102690	training	183.0	165.6	17.4	109373	training	246.0	234.8	11.2
102734	training	105.0	126.8	−21.8	109472	training	159.5	161.1	−1.6
102864	ext pred	142.5	143.9	−1.4	109473	ext pred	196.5	206.6	−10.1
102865	training	184.0	176.5	7.5	109654	training	179.5	179.3	0.2
102953	training	114.2	124.0	−9.8	109756	training	232.0	219.6	12.4
103007	training	155.5	147.4	8.1	109899	training	210.6	196.7	13.9
103076	training	120.7	116.0	4.7	109919	training	179.5	219.5	−40.0
103095	training	132.0	129.8	2.2	110036	training	170.0	158.5	11.5
103149	training	136.0	127.9	8.1	110149	training	231.0	216.8	14.2
103259	training	151.5	150.6	0.9	110167	ext pred	240.0	234.0	6.1
103581	training	138.0	127.0	11.0	110744	training	226.5	248.0	−21.5
103587	training	123.5	134.5	−11.0	110936	training	239.0	233.9	5.1
103639	training	202.0	208.6	−6.6	111024	training	138.0	140.4	−2.4
103670	training	170.5	166.6	3.9	111056	training	167.0	192.9	−25.9
103682	training	150.0	145.9	4.1	111060	training	189.0	216.9	−27.9
103694	ext pred	139.0	153.5	−14.5	111162	training	259.0	241.4	17.6
103733	training	63.3	92.5	−29.2	111218	training	224.0	216.9	7.1
103770	training	182.5	171.8	10.7	111416	training	146.5	130.2	16.3
103801	training	186.5	185.9	0.6	111486	training	259.5	242.7	16.8
103989	training	189.5	173.3	16.2	111585	training	232.0	241.3	−9.3
104182	ext pred	100.0	107.9	−7.9	111667	training	239.0	237.3	1.7
104540	training	99.6	112.5	−12.9	112023	training	215.3	210.0	5.3
104652	ext pred	126.8	139.1	−12.3	112135	ext pred	203.0	214.4	−11.4
104751	training	185.0	182.4	2.6	112480	training	258.0	222.4	35.6
104760	training	157.0	151.6	5.4	112487	training	266.5	253.9	12.6
105131	training	180.0	168.6	11.4	112548	ext pred	190.0	190.9	−0.8
105138	ext pred	174.5	161.0	13.5	112550	training	261.5	262.9	−1.4
105219	training	144.5	149.2	−4.7	112640	training	256.0	265.6	−9.6
105248	training	204.3	199.3	5.0	112688	training	162.5	167.5	−5.0
105255	ext pred	136.3	148.2	−11.9	113118	training	207.0	208.8	−1.8
105446	training	137.8	144.4	−6.6	113121	training	258.0	253.5	4.5
105652	training	186.0	207.5	−21.5	113340	training	222.0	214.3	7.7
105681	training	162.2	163.3	−1.1	113994	training	266.5	249.1	17.4
105700	ext pred	159.2	165.8	−6.6	114285	training	228.0	197.5	30.5
105719	training	174.5	179.8	−5.3	114494	training	221.0	236.7	−15.7
105752	training	114.5	128.7	−14.2	114528	training	275.0	272.9	2.1
105794	training	126.0	126.4	−0.4	114964	training	257.5	261.5	−4.0
105796	training	167.1	165.4	1.7	115464	training	238.0	220.7	17.3
105819	training	197.5	201.6	−4.1	116319	training	227.0	221.4	5.6
106217	training	146.5	132.7	13.8	116674	training	192.8	194.9	−2.1
106439	training	107.5	106.5	1.0	116696	training	255.0	240.9	14.1
106440	ext pred	152.9	150.9	2.0	116796	training	165.0	172.6	−7.6
106610	training	139.0	149.6	−10.6	116821	training	170.5	195.3	−24.8
106895	training	187.0	200.9	−13.9	116993	training	219.5	198.0	21.5
106896	training	203.5	213.7	−10.2	117246	ext pred	276.0	284.1	−8.1
106982	training	125.5	123.0	2.5	117547	ext pred	281.0	285.6	−4.6
107045	training	139.0	147.6	−8.6	117598	training	242.0	244.0	−2.0

Table 2 (Continued)

BRN number	set membership	observed boiling point (°C)	fitted/predicted boiling point (°C)	fit/prediction error (°C)	BRN number	set membership	observed boiling point (°C)	fitted/predicted boiling point (°C)	fit/prediction error (°C)
117938	training	233.3	237.0	−3.7	132536	training	210.5	226.3	−15.8
119280	training	304.0	292.8	11.2	138297	training	329.5	329.1	0.4
119466	training	203.0	224.5	−21.5	141070	training	255.0	263.3	−8.3
119629	training	233.0	209.5	23.5	142110	ext pred	201.0	205.4	−4.4
120634	training	283.0	275.1	7.9	142482	training	260.0	273.3	−13.3
121061	training	270.0	269.8	0.2	145486	training	278.5	277.9	0.6
121075	training	230.5	210.6	19.9	148125	training	294.5	272.0	22.5
121603	ext pred	259.0	251.2	7.8	151030	training	282.0	284.1	−2.1
122293	training	207.0	223.4	−16.4	157907	training	292.0	296.0	−4.0
122437	training	260.5	246.6	13.9	164164	training	285.5	267.1	18.4
122594	training	216.0	212.0	4.0	164180	training	282.5	288.6	−6.1
122848	training	236.0	238.5	−2.5	164362	training	293.0	296.9	−3.9
123370	training	170.0	209.5	−39.5	165751	training	321.0	327.0	−6.0
125169	training	306.0	297.8	8.2	174482	training	220.0	239.6	−19.6
125405	training	232.2	246.1	−13.9	211356	training	293.0	284.1	8.9
125786	training	221.5	200.1	21.4	383599	training	153.0	133.2	19.8
130105	training	287.5	317.5	−30.0	383604	training	172.0	205.6	−33.6
130566	training	274.5	262.7	11.8	384639	training	297.0	291.1	5.9
131639	training	263.0	273.8	−10.8					

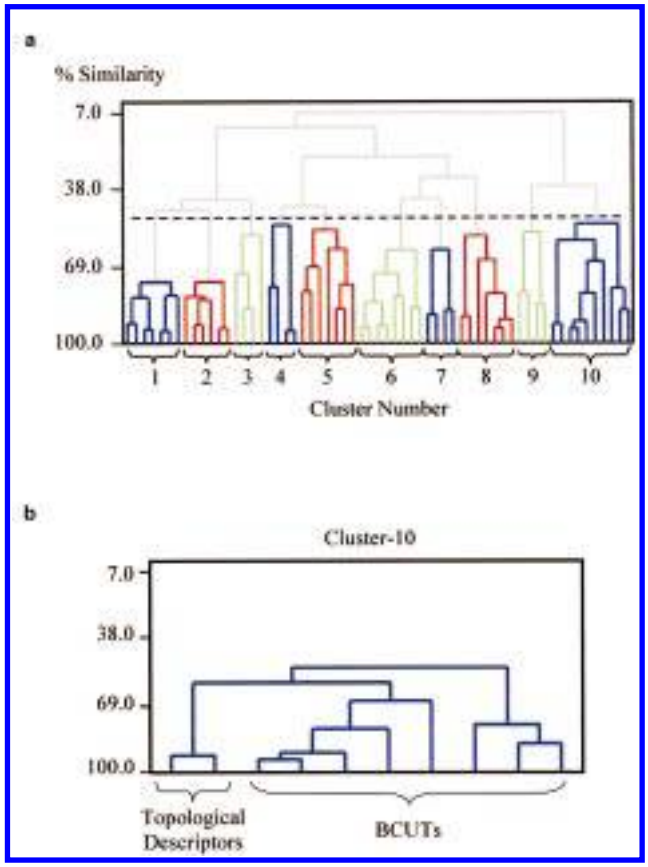

Figure 2. (a, top) A dendrogram showing the correlation of the 64 descriptors remaining after OFS for the DHFR inhibitor data set. (b, bottom) An expanded view of the linkage of the BCUT descriptors and the topological descriptors in cluster 10 of the dendrogram shown in part a. The topological descriptors are not joined to the BCUT descriptors until the similarity level drops below 59%.

Table 3. Cluster Membership for the Ten Clusters Obtained Using Hierarchical Cluster Analysis of the 64-Descriptors Surviving OFS for the DHFR Data Set[a]

cluster number	cluster membership
cluster-1	2-CPSAs
	5-others
cluster-2	3-BCUTs
	1-CPSA
	2-others
cluster-3	1-BCUT
	3-others
cluster-4	2-BCUTs
	1-CPSA
	1-other
cluster-5	2-BCUTs
	4-CPSAs
	1-other
cluster-6	8-others
cluster-7	2-BCUTs
	2-others
cluster-8	4-BCUTs
	2-CPSAs
	1-other
cluster-9	4-CPSAs
cluster-10	8-BCUTs
	2-others

[a] The cluster membership entries show the number and class of descriptors included in a given cluster. The *Others* classifications represents the set of conventional QSAR descriptors that are derived from topological, geometric, and electronic representations of molecular structure that do not include the CPSAs or the BCUTs.

VIF of 3.05. Since these values are within acceptable limits, the collinearity that does not cause great concern. The results of these diagnostics show the model to be statistically significant and robust.

One interesting feature of the model is the importance of the BCUT metrics. The metric BCUT-11 (elecneg_S_burden_1.00_R_H) is the descriptor most highly correlated with the dependent variable (DHFR inhibition) ($r = -0.789$). The correlation is shown graphically in Figure 5. This particular BCUT metric employs the fractional surface-area weighted atomic electronegativity as the atomic property used in the BCUT calculation. Since electronegativity is not a characteristic that is explicitly encoded by any of the ADAPT descriptors used, the inclusion of this BCUT metric is even more interesting. The relationship is obviously nonlinear, and a simple mathematical transform of the BCUT metric that would linearize the relationship could not be found. However, it is clear that finding a way to linearize the relationship between the BCUT metric and the observed

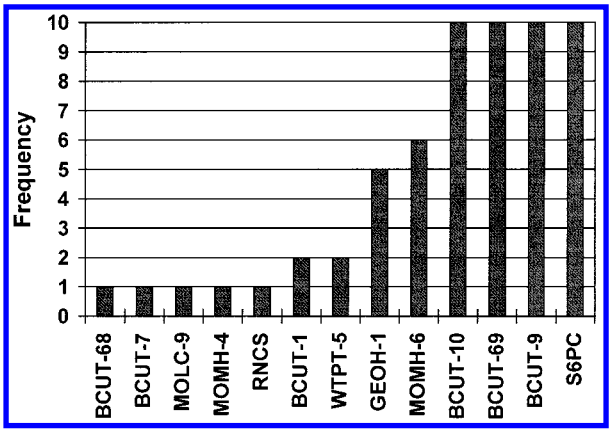

Figure 3. Summary of the frequency of inclusion of descriptors in the 10 best six-variable models of DHFR inhibition activity generated using GSA. Descriptors with a frequency of 1 are included in only one of the 10 models, while descriptors with a frequency of 10 are included in all 10 models.

DHFR activity would yield a very tight correlation. Another way to assess the importance of the BCUTs in this particular model is to examine the contribution of each descriptor using partial least squares (PLS) regression. This analysis was carried out using the PLS implementation in the SCAN software package. It was found that 87.3% of the variation in the observed DHFR activity is accounted for by the first three PLS components (model total = 87.8%). Three different BCUTs were the most highly loaded of all six descriptors in each of the first three components. This suggests that the BCUT metrics play a very significant role in the model, providing the majority of the structural information needed to explain the differences in the observed DHFR activity values.

While these results provide compelling evidence that the BCUT metrics are providing significant structural information to the model, it was interesting to see what could be accomplished using just the conventional ADAPT descriptors. This experiment was conducted by using the same 62-member training set and the 105 available ADAPT descriptors. Model development was carried out in the same fashion noted above. The best six-variable model obtained in this case yielded a poorer fit to the observed data ($R^2 = 0.826$, $s = 0.355$) when compared to the best model developed including the BCUTS ($R^2 = 0.878$). This result further reinforces the notion that the BCUT metrics are contributing unique and important structural information to the model.

The final test of the model was the evaluation of its predictive strength. Activity values for the 10 external prediction set compounds that were selected at random and set aside prior to model development. The model yielded excellent correlation between the predicted and observed DHFR activity values for all 10 compounds ($r = 0.997$). The prediction results are shown in Figure 6. The performance of the model in prediction, coupled with the model statistical diagnostics show this model to be sound.

The last issue to consider in this study was that of the two anomalies noted by Hirst in his work and set aside at the beginning of this study. The first of these is compound **01** (the 3,5-dihydroxy substituted analog), and the second is compound **10** (the 3,5-dihydroxy methyl analog). These two analogues are the only two that place a hydrophilic hydroxyl group in a hydrophobic pocket of the receptor. In other disubstituted analogues, a hydrophilic substituent is paired with a hydrophobic substituent. The phenyl ring can rotate to place the polar group toward solvent and the hydrophobic group toward the enzyme. The unfavorable interactions of the dihydroxy analogues result in much poorer than expected activity. The exclusion of these two compounds from the model development process is justified because they would have created a severe bias in the model.

Boiling Point Study. Model development for the boiling point data set was carried out in the same manner used for the DHFR study. The initial training set comprised 159 structures. The process of OFS for this data set yielded a reduced pool of 78 descriptors. Again, hierarchical cluster analysis was used to examine the relationship between the remaining descriptors. In this case, a set of seven clusters was obtained by selecting the cut-point as shown of the graph in Figure 7. The cluster membership is described in Table 6. As before, the BCUTs tend to be clustered primarily with other BCUTs and then with CPSAs. Also, as before, when topological or geometric descriptors are combined in a cluster with BCUTs, the linkage occurs at a low level of similarity close to the cut-point. The BCUT metrics appear to provide new structural information not provided by the conventional descriptor types. These results show that the relationship between the descriptors and the way they encode information concerning molecular structure is not an artifact of the training set selection but is characteristic of how they capture structural information.

The size of the initial model was selected based on the value of the RMS error of the best models containing 5−15 variables generated using GSA. Based on this criterion, attention was focused on an eight-variable model. The descriptor membership of the top 10 models from GSA is summarized in Figure 8. All the models contain at least two and as many as three BCUT descriptors. The best model of the set was considered for further analysis. The selected model yielded a good fit to the observed boiling point data ($R^2 = 0.945$). The correlation of the observed and estimated normal boiling points for the model is shown in Figure 9. The details of the model are presented in Table 7. An examination of slightly larger models (9 and 10 variables) showed no significant improvement in fit. Evaluation of potentially useful descriptors set aside at the OFS step also failed to yield a significant improvement. Standard regression diagnostics were applied to the model. The equation as a whole was found to be significant at the 95% confidence level (overall F-value = 323.10) compared to a critical F-value of 2.00 (8, 150, $\alpha = 0.05$). Each descriptor was also found to be significant at the 95% confidence level (minimum partial F-value = 18.88, compared to the critical value noted above). The maximum VIF was 3.5, with an average VIF of 2.10 (see Table 7). These values indicate the collinearity among the descriptors is low. Based on these results, the model appears to be quite acceptable for further analysis.

PLS analysis was again used to evaluate the contribution of the BCUT metrics to the model. As in the case of the DHFR study, the BCUT descriptors appear to be providing a large portion of the structural information needed to explain the observed boiling points. The first three PLS components account for 93.5% of the variance in the observed boiling point data, compared to 94.5% for the whole model. It is

Table 4. Details of Best Model for the 62-Observation DHFR Training Set Obtained Using GSA[c]

descriptor	regression coeff	std dev of coeff	partial-*F* value	VIF
sixth-order path-cluster molecular connectivity index (S6PC)[a]	5.150E−01	4.780E−02	116.2	2.8
molecular length (including hydrogens) (GEOH-1)	−4.510E−02	7.650E−03	34.8	1.2
Constant_S_burden_0.25_R_H (BCUT-1)[b]	1.590E+00	4.670E−01	11.6	1.1
Elecneg_S_burden_0.50_R_L (BCUT-9)[b]	−4.530E+00	1.020E+00	19.4	5.1
Elecneg_S_burden_1.00_R_H (BCUT-10)[b]	−7.090E−01	1.440E−01	24.2	1.3
Tabpolar_S_burden_0.50_R_L (BCUT-69)[b]	2.690E+01	3.950E+00	46.3	3.7
intercept	2.830E+01	3.410E+00	68.9	N/A

[a] Reference 2. [b] Reference 10. [c] $R^2 = 0.871$, $R^2_{cv,pls} = 0.840$, $s = 0.306$, overall-F (for AOV) = 61.7, F (6, 55, $\alpha = 0.05$) = 2.27.

Table 5. Details of the Modified Model for the 62-Observation DHFR Training Set Obtained by Examining Descriptors That Were Set Aside as Part of the Objective Feature Selection (OFS) Process[c]

descriptor	regression coeff	std dev of coeff	partial-*F* value	VIF
(S6PC)	4.120E−01	6.350E−02	42.1	5.3
second moment of inertia (MOMH-2)[a]	−1.150E−04	1.840E−05	38.8	1.3
BCUT-1	1.650E+00	4.660E−01	12.6	1.1
BCUT-10	−6.070E−01	1.440E−01	17.5	1.4
Elecneg_S_burden_1.00_R_L (BCUT-11)[b]	−3.620E+00	6.830E−01	28.0	6.5
BCUT-69	2.440E+01	3.280E+00	55.3	2.7
intercept	2.080E+01	2.720E+00	58.8	N/A

[a] Reference 21. [b] Reference 10. [c] $R^2 = 0.878$, $R^2_{cv,pls} = 0.849$, $s = 0.297$, overall-F (for AOV) = 66.1, F (6, 55, $\alpha = 0.05$) = 2.27.

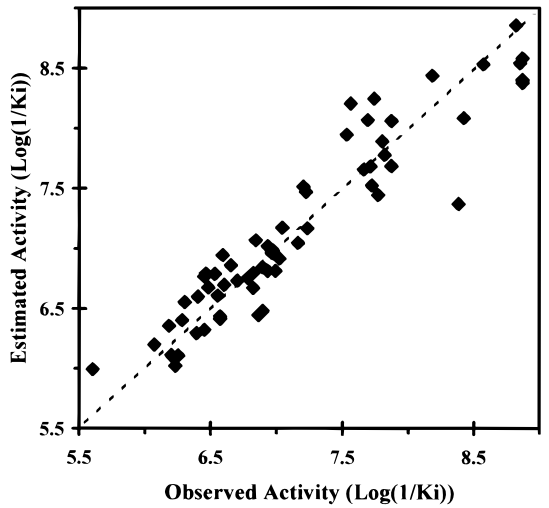

Figure 4. Comparison of the estimated and observed DHFR inhibitor activity values (log(1/K_i)) for the 62-observation data set model (see Table 4).

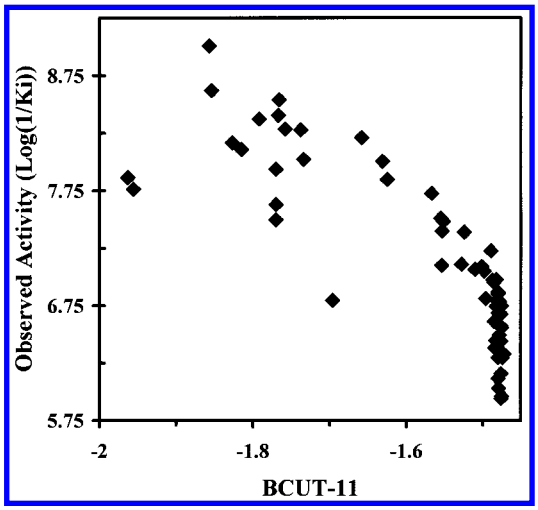

Figure 5. Graph illustrating the relationship observed between the observed inhibitor activity and BCUT-11 (Elecneg_S_burden_1.00_R_H) descriptor values for the 62-observation training set.

interesting to note that a BCUT metric is the highest loaded descriptor in all of the first three components. One BCUT metric in particular (BCUT-27, gastchrg_S_burden_0.25_R_L) yields the largest correlation coefficient with the observed boiling point values ($r = -0.690$). This correlation is illustrated in Figure 10. Gasteiger−Huckel partial atomic charges, weighted by the solvent-accessible surface area of the given atom, are used in the calculation of this particular BCUT metric. The BCUT metric produces two or more separate clusters in the data set, which seem to segregate the data set according to structure class. Furans and thiophenes that contain carbonyl groups appear to be in one cluster, while the saturated ring systems (tetrahydrofurans and tetrahydrothiophenes) are in another class. Such separations are often seen with other descriptors, and this observation is consistent with the notion that the BCUTs are measuring important structural features.

It is interesting to compare the model developed in this study to the one developed previously.[12] While the training sets are not identical, the current data set represents a large subset (67%) of the combined data set used in the past work. The most notable difference between the models is the lack of contribution of the CPSA descriptors in the current equation. The CPSA descriptors were originally conceived to capture structural information important for understanding polar intermolecular interactions. In the past, the CPSA descriptors have played a very significant role in boiling point QSPR models. However, only one CPSA descriptor (FNSA-3) appears in the current model. The BCUT metrics appear to provide the structural information previously provided by the CPSA descriptors. Additionally, it appears that the BCUT metrics do a better job of capturing the needed structural information given that the GSA descriptor selection step had both CPSAs and BCUTs to choose from. To further examine the notion that the BCUTs are providing structural information similar to the CPSA descriptors, an experiment was done to find the best eight-variable model that excluded the BCUT metrics. The best eight-variable developed using only the

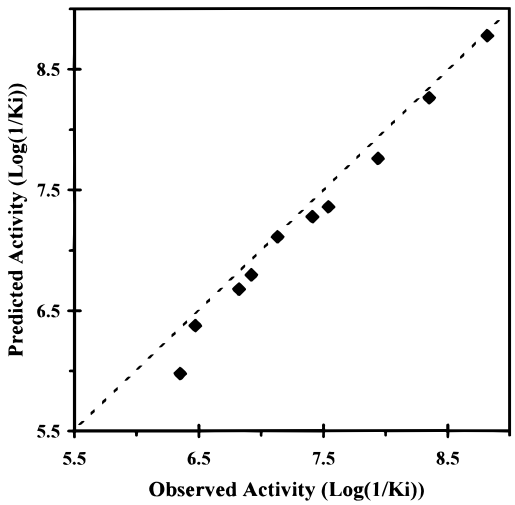

Figure 6. Comparison of the predicted and observed DHFR inhibitor activity values (log(1/K_i)) for the 10-observation external prediction set.

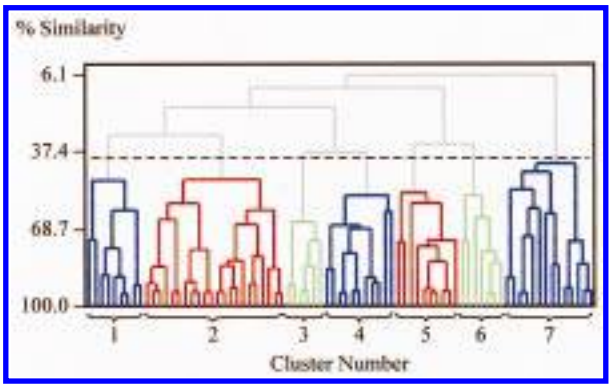

Figure 7. Dendrogram showing the correlation among the 78 descriptors remaining following OFS for the boiling-point data set.

Table 6. Cluster Membership for the Seven Clusters Obtained Using Hierarchical Cluster Analysis of the 78 Descriptors Surviving OFS for the Boiling Point Data Set[a]

cluster number	cluster membership
cluster-1	3-BCUTs
	5-CPSAs
	1-other
cluster-2	22-others
cluster-3	2-BCUTs
	4-CPSAs
cluster-4	2-CPSA
	9-other
cluster-5	8-BCUTs
	1-CPSA
	1-other
cluster-6	3-BCUTs
	2-CPSAs
	2-others
cluster-7	8-BCUTs
	4-CPSAs
	2-others

[a] The cluster membership entries show the number and class of descriptors included in a given cluster.

conventional ADAPT descriptor set and the same 159-member training set yielded a poorer fit to the observed boiling point data ($R^2 = 0.925$), compared to the best model containing BCUT metrics ($R^2 = 0.945$). The non-BCUT model contained three CPSA descriptors, compared to the one CPSA descriptor in the model that included the BCUTs.

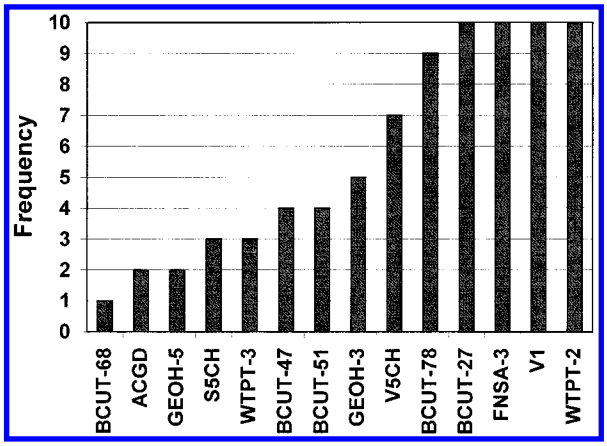

Figure 8. Summary of the frequency of inclusion of descriptors in the 10 best eight-variable models of boiling point generated using GSA.

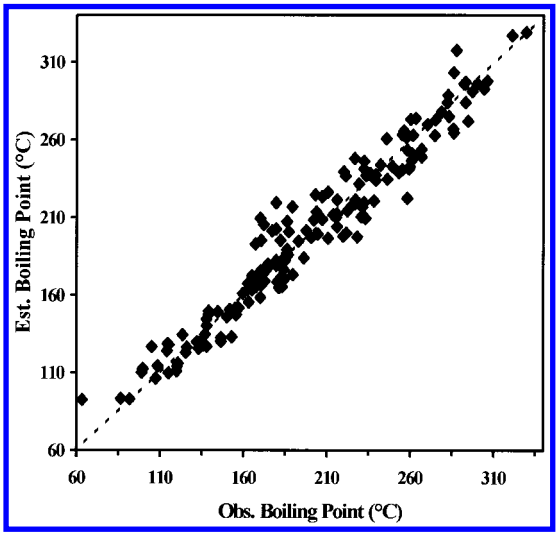

Figure 9. Comparison of the estimated and observed boiling normal boiling point values for the 159-observation training set model (see Table 5).

This is further evidence that the BCUT metrics are providing similar information concerning the structural features responsible for polar intermolecular interactions and providing that information in greater detail than the CPSA descriptors.

Again, the true value and strength of the model can be assessed in external prediction. The model described above was used to predict the normal boiling points for the 20 external prediction set compounds set aside prior to model development. The model performs well, yielding an excellent correlation between the predicted and observed boiling point values ($r = 0.989$). The comparison of the predicted and observed boiling point values is shown in Figure 11. Thus, the boiling point model containing the BCUT descriptors is found to be statistically sound and useful for predicting the normal boiling points for these classes of compounds.

CONCLUSIONS

The QSAR/QSPR studies described above provide strong evidence that the BCUT metrics make very useful and effective descriptors for such studies. It is apparent that they are measuring particular structural features that can be related to the observed properties of a variety of molecules. They also appear to provide unique structural information com-

Table 7. Details of the Eight-Variable Model for the 159-Observation Normal Boiling Point Data Set[e]

descriptor	regression coeff	std dev of coeff	partial-*F* value	VIF
fractional negative surface area (FNSA-3)[a]	−9.345E+02	7.088E+01	173.8	3.5
valence-corrected first-order molecular connectivity index (V1)[b]	4.019E+01	1.597E+00	633.2	2.1
valence-corrected fifth-order chain molecular connectivity index (V5CH)[b]	−1.111E+02	1.748E+01	40.4	1.4
molecular ID/number of atoms (WTPT-2)[c]	3.092E+02	3.597E+01	73.9	1.4
molecular thickness (GEOH-3)	−2.608E+01	6.004E+00	18.9	1.6
Gastchrg_S_burden_0.25_R_L (BCUT-27)[d]	−1.627E+02	1.416E+01	132.1	3.1
Haccept_burden_0.25_R_H (BCUT-51)[d]	2.331E+01	3.684E+00	40.0	1.5
Tabpolar_burden_0.50_R_H (BCUT-78)[d]	2.302E+01	3.090E+00	55.5	2.2
intercept	−7.967E+02	7.011E+01	129.5	N/A

[a] Reference 7. [b] Reference 2. [c] Reference 22. [d] Reference 10. [e] $R^2 = 0.945$, $R^2_{cv,pls} = 0.939$, $s = 13.6$, overall-F (for AOV) = 323.1, F (8, 150, $\alpha = 0.05$) = 2.00.

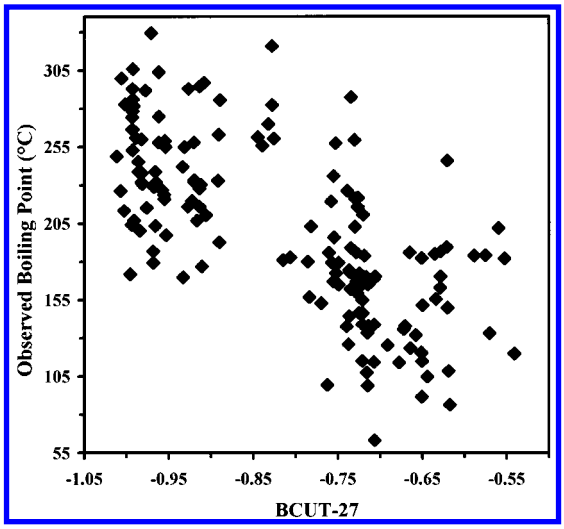

Figure 10. Graph illustrating the relationship between the observed boiling point values and the BCUT-27 descriptor values for the 159-observation training set.

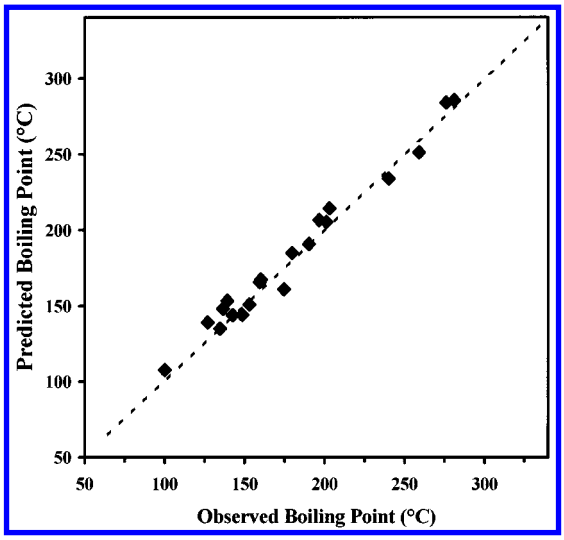

Figure 11. Comparison of the predicted and observed boiling point values for the 20-observation external prediction set.

pared to the more typical descriptors (CPSAs, molecular connectivity indices, etc.) given the results of the cluster analysis of the descriptors, and that they are selected over other descriptors for inclusion in models using the GSA optimization technique. The evidence described above shows that they complement, rather than simply replace, the more typical types of descriptors in that they are usually most highly correlated just with other BCUTs. The BCUT metrics appear to perform better than the CPSA descriptors in capturing structural information important for understanding polar intermolecular interactions, given these two data sets and properties. The metrics that appear to be the most useful are those that employ the solvent accessible surface area as a weighting factor for the atomic property. This is appealing from the point of view of a physical interpretation because the effect of a given atom in such an intermolecular interaction will be proportional to its exposure on the surface of the molecule. When viewed in the light of the past utility of the CPSA descriptors in a variety of studies, this result suggests that the use of BCUT metrics in QSAR and QSPR studies merits further exploration.

ACKNOWLEDGMENT

The author would like to thank Prof. R. S. Pearlman for providing access to the DiverseSolutions software package and for the information provided concerning the calculation of the BCUT metrics. Appreciation is also expressed for the suggestion by the reviewer of the inclusion of the hierarchical cluster analysis of the descriptors.

Supporting Information Available: Structures for the 179-observation boiling point data set. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, 97, 6609−6615.

(2) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, 1976.

(3) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; John Wiley & Sons: New York, 1986.

(4) Rohrbaugh, R. H.; Jurs, P. C. Descriptions of Molecular Shape Applied in Studies of Structure/Activity and Structure/Property Relationships. *Anal. Chim. Acta* **1987**, *199*, 99−109.

(5) Jurs, P. C.; Dixon, S. L.; Egolf, L. M. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; Methods and Principles in Medicinal Chemistry; VCH: New York, 1995; Vol. 2, Chapter 2.

(6) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027−1043.

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

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

(9) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.

(10) Pearlman, R. S.; Smith, K. M. In *3D-QSAR and Drug Design: Recent Advances*; Kubinyi, H., Martin, Y., Folkers, G., Eds.; Kluwer Academic: Dordrecht, Netherlands, 1997; pp 339−353.
(11) Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput, Sci.* **1989**, *29*, 225−227.
(12) Hirst, J. D. Nonlinear Quantitative Structure−Activity Relationship for the Inhibition of Dihydrofolate Reductase by Pyrimidines. *J. Med. Chem.* **1996**, *39*, 3526−3532.
(13) Stuper, A. J.; P. C. Jurs. ADAPT: A Computer System for Automated Data Analysis Using Pattern Recognition Techniques. *J. Chem. Inf. Comput. Sci*. **1975**, *2*, 99−105.
(14) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, R. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, D.C., 1979; pp 103−129.
(15) Pearlman, R. S. In *3D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Amsterdam, 1993; pp 41−79.
(16) Sybyl Version 6.3 Force Field Manual; Tripos: St. Louis, MO, USA, 1996; p 196.
(17) Gasteiger−Huckel partial atomic charges are calculated using the Gasteiger−Marsili method to calculate the σ-electron contributions and the Huckel method for calculating the π-electron contributions. Sybyl Version 6.3 Force Field Manual; Tripos: St. Louis, MO, USA, 1996; p 290.
(18) Dalby, A.; Nourse, J. G.; Hounshell, D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Lauffer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 244−255.
(19) Jurs, P. C.; Hasan, M. N.; Hansen, P. J.; Rohrbaugh, R. H. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: Berlin, 1988; pp 209−233.
(20) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 77−84.
(21) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950; pp 144−156.
(22) Randić, M. On Molecular Identification Numbers. *J. Chem. Comput. Sci.* **1984**, *24*, 164−175.
(23) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*; Irwin: Homewood, IL, 1985; p 240.
(24) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*; Irwin: Homewood, IL, 1985; p 281.
(25) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*; Irwin: Homewood, IL, 1985; p 391.

CI980102X