Development of a Quantitative Structure−Property Relationship Model for Estimating Normal Boiling Points of Small Multifunctional Organic Molecules

David T. Stanton[†]

Health Care Research Center, Procter & Gamble Pharmaceuticals, 8700 Mason-Montgomery Road,
Mason, Ohio 45040-8006

Received May 22, 1999

Computer-assisted quantitative structure−property relationship techniques are applied in the development of a robust and accurate model of normal boiling points (boiling at 760 mmHg) for a very diverse set of 268 small organic molecules. Most of the molecules included in this study contain two or more functional groups. The final model yields a tight fit to the training set data ($R^2 = 0.963$), with a fit error of 6.5%. More importantly, the model is also shown to perform well in external prediction. The mean prediction error for boiling points for a 78-member external test set was 12.3 °C, or 8.3%. A detailed analysis of the small number of compounds that were either outliers or not well predicted illustrates areas for potential improvement of the methodology used.

INTRODUCTION

Knowledge of the physical properties of organic compounds is necessary for the design, development, and manufacture of products in which they are used. The suitability of a particular compound for a given purpose will depend on its physicochemical properties. In the process of designing new products, a variety of alternatives for various product components is considered. The scope of the selection will be limited to those compounds for which physical property data are available. While several compilations of experimentally derived physical property values of organic compounds exist, data may not be available for all compounds being considered. In many cases it is impractical or impossible to measure the property of interest because insufficient material exists to test, there are toxicity issues, or the number of compounds for which data are needed is very large. Thus, a reliable means of estimating the needed property value would be an advantage. The physical property that is considered in this work is the normal boiling point (the boiling point at 760 mmHg).

A variety of methods for estimating normal boiling points have been reported in the literature. One subset of these requires a measure of some other physical property, such as boiling points at reduced pressure[1] or chromatographic retention indices.[2,3] While such methods are effective, they are also limited because these other experimentally determined property values may not be available. This becomes particularly troublesome in cases where one needs to estimate the boiling points for a large number of compounds. The most useful methods of estimating boiling point will be those that are based directly on descriptions of molecular structure.

Group contribution (or group additivity) methods represent one class of property estimation approaches. Some additivity methods are quite simple and can be implemented using only paper and pencil for small numbers of structures.[4] Other group contribution methods are more complicated and are best implemented using computers.[5−8] While such methods have the advantage of requiring only knowledge of the structure of the compounds of interest, they suffer from two important drawbacks. First, by definition, group contribution methods work only when the values for all significant fragments of the test molecule have been tabulated. If the test molecule contains an important structural fragment that does not have a corresponding contribution value, then an estimate of the boiling point cannot be made. The second problem with such methods is related to the additivity principle itself. The group contribution property estimates are computed by summing the product of the counts of fragments found in the test structures and their tabulated contribution values. However, it is very difficult to quantify the interactions of all possible combinations of fragments in a molecule, and such interactions often have a significant influence on the physical property of interest.

Other methods have been developed that employ a different approach to estimating normal boiling points based strictly on molecular structure. These are termed quantitative structure−property relationships (QSPR). The basis of such relationships is the assumption that compounds of similar structure will exhibit similar properties. The key to the success of QSPR is accurate measurement of the structural features that modulate the observed property. The structural characteristics of a molecule can be measured on either a substructure or a whole-molecule basis. These measures, or *descriptors*, encode structural characteristics that fall broadly into three classes: (1) topological, (2) geometric, and (3) electronic. Several examples of these methods have been published. Methods that employ multiple linear regression for selection of descriptors and generation of predictive equations for normal boiling points of large sets of hetrocyclic organic compounds (furans, tetrahydrofurans, thiophenes, pyrans, and pyrroles) have been described.[9,10] Later, it was shown that the accuracy of similar models could be improved by employing artificial neural network tech-

[†] E-mail: stanton@pg.com.

10.1021/ci990311x CCC: $19.00 © 2000 American Chemical Society
Published on Web 11/23/1999

niques.[11] These QSPR models were also shown to provide greater accuracy than group contribution techniques applied to the same set of compounds.[12−14] Similar approaches have been described for estimating the normal boiling points of fluorocarbons,[15] nonassociating compounds (compounds with no hydrogen-bonding potential) from the Toxic Substance Control Act (TSCA) inventory,[16] and other diverse sets of organic compounds.[17−19]

On the basis of past successes described for QSPR techniques applied for generating predictive equations for normal boiling points, this approach was chosen for the present work. The compounds of interest in this study can be generally categorized as being very diverse, containing a variety of functional groups (alcohols, ethers, aldehydes, esters, etc.). Most of the compounds contain two or more functional groups. While it would be desirable to employ one of the existing equations, experience indicates that the scope of such models is not yet broad enough to be generally applicable.[10] Given the scope of the diversity and the size of the potential target set of compounds being considered in this work, it seemed appropriate to develop a model based on known data for compounds similar to those for which predicted values will be sought.

METHODOLOGY

The general procedure used to analyze the data set and to develop and evaluate the predictive models has been described previously.[9] Deviations from the general process are detailed below. The structures for all compounds were first assembled as a SYBYL database (Tripos, Inc., version 6.2). All subsequent calculations were performed using the ADAPT software package[20,21] running on a Silicon Graphics Inc. Indigo-2 (R4400) workstation running under the IRIX (version 5.3) operating system. A short FORTRAN program was written to convert and store the structures from a Tripos multi-mol (Tripos MOL-format[22]) in ADAPT data file format.

The Data Set. The set of 372 structures involved in this study comprised a large and diverse set of functional-group types (alcohols, ethers, aldehydes, esters, etc.). Examples of some of the compounds involved are given in Figure 1. All structures are identified by the Chemical Abstracts Service (CAS) registry number. [All registry numbers have been provided by the author.] The structures were obtained in the form of SMILES[23] strings, and were stored as a single SYBYL database. Reasonable low-energy conformations were obtained by first converting all the 2D structures to 3D conformations using CONCORD,[24] followed by strain-energy minimization using the Tripos force field,[25] including electrostatic terms. The atomic partial charges used in the energy-minimization step were calculated in SYBYL using the Gasteiger−Huckel method.[26]

Experimental Boiling Point Data. The experimentally determined normal boiling point data used in this study were taken from five separate sources.[27−31] When two or more values were obtained for a given structure, the average of the values was taken as the “true” normal boiling point. The boiling points for all the compounds considered in this study spanned a range of 410 °C (21.0−431.0 °C), with a mean of 192.3 °C (see Table 1).

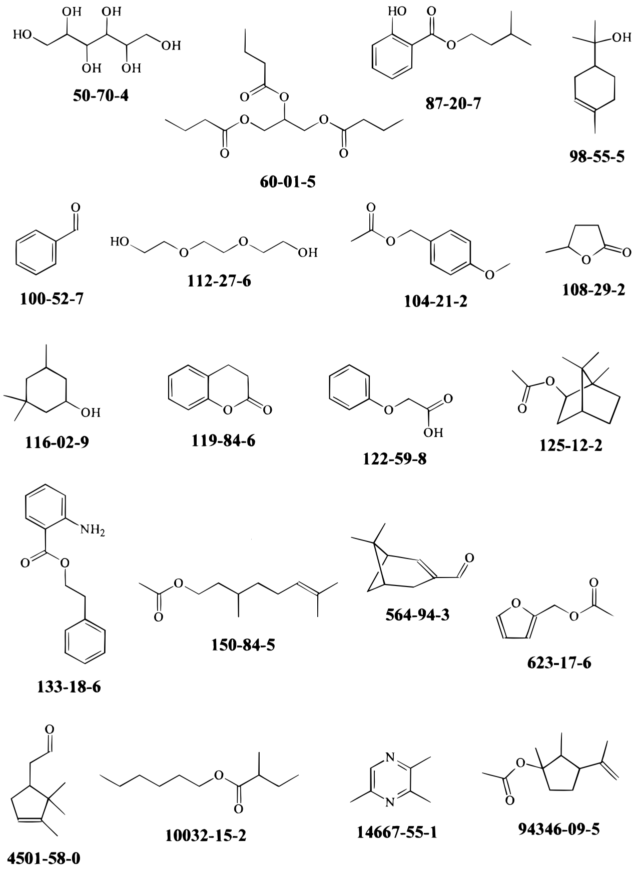

Figure 1. Example structures of compounds used to develop the normal boiling point model.

Descriptor Generation and Analysis. Once the energy-minimized structures had been stored in the ADAPT data files, a set of 129 individual molecular structure descriptors was calculated for each of the 372 structures in the data set. The descriptor set was selected to capture important topological, geometric, and electronic structural features. The topological descriptors are derived using graph-theoretical approaches to defining chemical structures (chemical graph theory).[32,33] Additional 2-dimensional information is captured as counts of specific structural fragments (i.e., counts of carbon and heteroatoms, counts of single, double, triple, and aromatic bonds, etc.). Geometric descriptors capture the 3-dimensional aspects of structure, such as surface area and volume[34] and width, length, and thickness.[35] Electronic descriptors provide information concerning the distribution of charge in the molecule.[36] Additionally, some descriptors employ structural representations that capture two or more of these structural feature types (e.g., surface area and partial atomic charge). Of particular interest were the charged partial surface area (CPSA) descriptors[37] and the related hydrogen-bonding-specific descriptors,[10] which have been shown to be useful in capturing information concerning structural features that modulate polar intermolecular interactions. The partial charges used in the calculation of the CPSA and related descriptors were obtained using the program CHARGE.[36]

Before any of the descriptors were used in the development of regression equations, they were first subjected to a process of objective feature selection (OFS). In this process, the descriptor values for a given set of structures are examined without considering the response variable (boiling point). Those descriptors that show little or no variation over the

Table 1. Observed and Calculated Normal Boiling Point Data for Both the 268-Member Training Set and the 78-Member External Prediction Set

CAS reg no.	set membership	obs bp (°C)	estim/ pred bp (°C)	residual (°C)	CAS reg no.	set membership	obs bp (°C)	estim/ pred bp (°C)	residual (°C)	CAS reg no.	set membership	obs bp (°C)	estim/ pred bp (°C)	residual (°C)
100-06-1	training	258.0	239.5	18.5	108-95-2	training	182.0	167.7	14.3	124-07-2	training	239.0	239.0	0.0
10031-82-0	prediction	252.0	242.7	9.3	109-06-8	training	129.0	133.2	−4.2	124-10-7	prediction	99.0	96.1	2.9
10031-87-5	training	163.0	149.5	13.6	109-15-9	training	245.0	227.2	17.8	124-19-6	prediction	199.0	197.9	1.1
10032-13-0	training	213.0	207.7	5.3	109-21-7	training	164.0	163.8	0.2	125-12-2	prediction	121.0	120.2	0.8
10032-15-2	training	212.0	199.9	12.1	109-43-3	prediction	158.0	154.1	3.9	128-37-0	training	265.0	277.0	−12.0
100-42-5	training	145.0	147.2	−2.2	109-52-4	training	185.0	181.0	4.0	133-18-6	training	226.0	268.6	−42.6
100-47-0	training	191.0	204.2	−13.2	109-60-4	training	102.0	113.0	−11.0	135-02-4	training	238.0	225.3	12.7
100-51-6	training	205.0	199.4	5.6	109-94-4	prediction	229.0	229.5	−0.5	136-60-7	prediction	103.0	112.6	−9.6
100-52-7	training	178.0	174.6	3.4	110-19-0	training	117.0	127.3	−10.3	138-86-3	training	176.0	171.1	4.9
100-66-3	training	154.0	149.3	4.7	110-38-3	training	245.0	242.1	2.9	140-11-4	training	210.0	216.4	−6.4
100-86-7	training	215.0	217.6	−2.6	110-39-4	training	224.0	241.0	−17.0	140-26-1	prediction	143.0	139.7	3.3
101-41-7	prediction	163.0	149.5	13.5	110-40-7	training	312.0	298.8	13.2	140-29-4	training	234.0	216.2	17.8
101-48-4	training	194.0	227.3	−33.3	110-43-0	training	151.0	142.3	8.7	140-39-6	prediction	158.0	154.1	3.9
101-81-5	prediction	212.0	199.9	12.1	110-45-2	training	123.0	151.2	−28.2	140-67-0	training	216.0	228.6	−12.6
101-84-8	prediction	217.0	201.7	15.3	110-62-3	training	103.0	107.6	−4.6	14073-97-3	training	208.0	196.8	11.2
101-97-3	training	228.0	221.1	6.9	110-74-7	prediction	269.0	276.4	−7.4	140-88-5	training	99.0	108.4	−9.3
102-04-5	prediction	265.0	276.9	−11.9	110-86-1	training	115.0	116.6	−1.6	141-10-6	prediction	235.0	237.7	−2.7
102-13-6	prediction	259.0	282.5	−23.5	110-89-4	training	106.0	117.8	−11.8	141-12-8	training	236.0	231.1	4.9
102-76-1	training	259.0	262.4	−3.4	110-93-0	training	173.0	169.0	4.0	141-78-6	training	77.0	88.7	−11.7
103-25-3	training	239.0	225.5	13.5	110-98-5	prediction	145.0	153.4	−8.4	141-79-7	training	130.0	119.4	10.6
103-26-4	training	262.0	250.1	11.9	111-11-5	prediction	132.0	127.7	4.3	141-97-9	prediction	233.0	231.1	1.9
103-28-6	prediction	331.0	324.0	7.0	111-12-6	prediction	344.0	319.8	24.2	142-62-1	prediction	205.0	224.2	−19.2
103-36-6	training	271.0	267.7	3.3	111-14-8	training	223.0	220.6	2.4	142-92-7	training	170.0	178.7	−8.7
103-37-7	prediction	247.0	250.4	−3.4	111-27-3	training	157.0	154.2	2.8	143-07-7	training	299.0	300.6	−1.6
103-38-8	prediction	234.0	232.2	1.8	111-70-6	training	176.0	177.5	−1.5	143-08-8	prediction	223.0	232.0	−9.0
103-45-7	training	235.0	238.6	−3.6	111-71-7	training	153.0	153.9	−0.9	14667-55-1	training	171.0	171.6	−0.6
103-48-0	training	250.0	250.8	−0.8	111-81-9	prediction	54.0	89.5	−35.5	148-24-3	training	267.0	279.1	−12.1
103-50-4	prediction	240.0	236.2	3.8	111-87-5	prediction	312.0	298.8	13.2	150-78-7	prediction	205.0	205.0	0.0
103-52-6	training	260.0	256.3	3.8	112-05-0	prediction	80.0	113.0	−33.0	150-84-5	training	234.0	233.4	0.7
103-58-2	training	282.0	261.7	20.3	112-06-1	prediction	233.0	256.1	−23.1	15356-70-4	training	216.0	215.0	1.0
103-82-2	training	265.0	252.2	12.8	112-12-9	prediction	194.0	183.1	10.9	1632-73-1	training	199.0	204.7	−5.7
103-93-5	training	237.0	224.0	13.0	112-14-1	training	211.0	216.9	−5.9	16409-45-3	training	228.0	257.9	−29.9
104-09-6	training	222.0	226.6	−4.6	112-23-2	prediction	218.0	192.9	25.1	1797-74-6	training	239.0	245.7	−6.7
104-50-7	training	234.0	198.3	35.7	112-27-6	prediction	248.0	251.7	−3.7	18479-51-1	prediction	229.0	229.5	−0.5
104-53-0	training	222.0	226.2	−4.2	112-30-1	prediction	195.0	195.1	−0.1	18479-57-7	training	195.0	184.4	10.7
104-61-0	prediction	245.0	253.0	−8.0	112-31-2	training	208.0	215.3	−7.3	18479-58-8	training	192.0	194.1	−2.1
104-65-4	training	252.0	266.3	−14.3	112-38-9	training	275.0	288.7	−13.7	2021-28-5	prediction	225.0	229.5	−4.5
104-67-6	prediction	295.0	319.1	−24.1	1124-11-4	training	190.0	187.8	2.2	20777-49-5	training	233.0	244.9	−11.9
10482-56-1	training	217.0	208.9	8.1	112-42-5	training	243.0	258.7	−15.7	21722-83-8	training	222.0	219.1	2.9
104-87-0	training	204.0	197.5	6.5	112-43-6	training	250.0	269.8	−19.8	2198-61-0	prediction	184.0	173.9	10.1
104-93-8	training	175.0	172.5	2.5	112-53-8	prediction	259.0	256.1	2.9	2216-51-5	training	216.0	215.0	1.0
105-13-5	training	259.0	239.6	19.4	115-18-4	training	98.0	87.5	10.5	2315-68-6	training	231.0	230.7	0.3
105-37-3	training	99.0	96.1	2.9	115-95-7	prediction	192.0	201.2	−9.2	2345-24-6	training	229.0	250.8	−21.8
105-53-3	training	199.0	197.9	1.1	116-02-9	training	198.0	211.0	−13.0	2445-77-4	training	187.0	180.5	6.5
105-54-4	training	121.0	120.2	0.8	116-53-0	prediction	230.0	227.4	2.6	24817-51-4	training	256.0	256.3	−0.3
105-57-7	training	103.0	112.6	−9.6	118-61-6	prediction	178.0	196.2	−18.2	2623-23-6	training	227.0	257.9	−30.9
105-66-8	training	143.0	139.7	3.3	119-36-8	training	223.0	247.8	−24.8	2639-63-6	training	205.0	204.1	0.9
105-68-0	prediction	260.0	256.3	3.7	119-61-9	prediction	278.0	292.5	−14.5	27215-95-8	prediction	189.0	188.2	0.8
105-85-1	training	235.0	237.7	−2.7	119-84-6	prediction	231.0	234.6	−3.6	3188-00-9	prediction	207.0	204.7	2.3
105-87-3	training	233.0	231.1	1.9	120-14-9	training	283.0	261.2	21.8	3208-16-0	training	92.0	96.7	−4.7
106-21-8	training	205.0	224.2	−19.2	120-57-0	prediction	259.0	272.3	−13.3	334-48-5	prediction	269.0	276.4	−7.4
106-22-9	training	223.0	232.1	−9.1	121-98-2	training	250.0	228.8	21.2	3404-61-3	prediction	147.0	124.7	22.3
106-23-0	training	205.0	205.0	0.0	122-00-9	prediction	196.0	206.3	−10.3	3452-97-9	training	193.0	183.2	9.8
106-24-1	prediction	237.0	224.0	13.0	122-03-2	training	235.0	233.1	1.9	35154-45-1	prediction	123.0	117.8	5.2
106-25-2	training	225.0	229.5	−4.5	122-57-6	training	261.0	260.7	0.3	35158-25-9	prediction	202.0	189.0	13.0
106-27-4	training	184.0	173.9	10.1	122-59-8	training	285.0	272.2	12.8	36653-82-4	training	334.0	341.0	−7.0
106-30-9	training	189.0	188.2	0.8	122-67-8	prediction	171.0	168.0	3.0	3848-24-6	training	128.0	152.1	−24.1
106-32-1	training	207.0	204.7	2.3	122-78-1	prediction	191.0	232.4	−41.4	409-02-9	prediction	174.0	180.3	−6.3
106-33-2	prediction	243.0	217.5	25.5	122-97-4	training	235.0	245.3	−10.3	41519-18-0	training	204.0	196.9	7.1
106-35-4	training	147.0	124.7	22.3	122-99-6	training	245.0	248.9	−3.9	4180-23-8	prediction	200.0	158.1	41.9
106-36-5	training	123.0	117.8	5.3	123-11-5	prediction	204.0	197.5	6.5	431-03-8	training	88.0	114.5	−26.5
106-44-5	training	202.0	189.0	13.0	123-19-3	training	145.0	127.5	17.5	4351-54-6	training	163.0	181.4	−18.4
106-65-0	training	198.0	187.3	10.7	123-25-1	training	218.0	222.0	−4.0	4437-51-8	training	124.0	130.6	−6.6
106-68-3	training	167.0	150.2	16.8	123-32-0	prediction	175.0	172.5	2.5	4501-58-0	training	202.0	203.2	−1.2
106-73-0	training	172.0	165.3	6.7	123-35-3	prediction	259.0	239.6	19.4	4602-84-0	prediction	198.0	187.3	10.7
107-75-5	training	241.0	233.1	7.9	123-38-6	training	47.0	58.6	−11.6	470-82-6	training	176.0	173.9	2.1
107-87-9	training	101.0	94.6	6.4	123-51-3	training	131.0	131.7	−0.7	473-54-1	training	203.0	202.9	0.1
107-92-6	training	163.0	159.6	3.4	123-66-0	training	168.0	164.1	3.9	4747-07-3	training	123.0	131.7	−8.7
108-10-1	training	117.0	105.2	11.8	123-72-8	training	75.0	82.2	−7.2	4864-61-3	prediction	167.0	150.2	16.8
108-21-4	training	89.0	96.7	−7.7	123-76-2	training	245.0	232.6	12.4	498-81-7	training	205.0	203.5	1.5
108-48-5	prediction	286.0	248.9	37.1	123-86-4	training	125.0	136.3	−11.3	499-75-2	training	238.0	238.8	−0.8
108-64-5	training	132.0	127.7	4.3	123-92-2	training	142.0	150.6	−8.6	501-52-0	training	280.0	274.7	5.3
108-88-3	training	111.0	107.5	3.6	124-04-9	training	338.0	321.4	16.6	503-74-2	training	176.0	166.7	9.3
108-89-4	training	144.0	137.8	6.2	124-06-1	training	295.0	307.4	−12.4	50-70-4	training	431.0	444.9	−13.9

Table 1 (Continued)

CAS reg no.	set membership	obs bp (°C)	estim/ pred bp (°C)	residual (°C)	CAS reg no.	set membership	obs bp (°C)	estim/ pred bp (°C)	residual (°C)	CAS reg no.	set membership	obs bp (°C)	estim/ pred bp (°C)	residual (°C)
513-86-0	prediction	172.0	165.3	6.7	611-13-2	training	181.0	175.6	5.4	78-69-3	training	155.0	183.0	−28.0
5146-66-7	training	228.0	230.0	−1.9	61692-84-0	training	183.0	180.4	2.6	78-70-6	training	193.0	191.5	1.5
515-00-4	training	221.0	229.5	−8.5	617-35-6	training	144.0	137.4	6.6	78-83-1	training	108.0	101.7	6.3
5292-21-7	training	243.0	236.1	6.9	619-01-2	prediction	163.0	159.6	3.4	78-84-2	training	63.0	73.8	−10.8
53398-85-9	training	203.0	202.3	0.8	620-79-1	training	276.0	263.4	12.7	79-09-4	training	141.0	134.8	6.2
534-22-5	training	64.0	79.3	−15.3	622-45-7	training	173.0	173.8	−0.8	79-20-9	training	57.0	69.8	−12.8
5349-62-2	training	250.0	249.0	1.0	623-17-6	training	176.0	196.8	−20.8	79-31-2	training	153.0	145.0	8.0
536-78-7	training	166.0	157.0	9.0	623-42-7	prediction	117.0	105.2	11.8	79-92-5	training	160.0	153.6	6.4
5392-40-5	prediction	114.0	133.4	−19.4	624-24-8	training	128.0	121.1	6.9	80-26-2	training	220.0	219.5	0.5
539-30-0	training	185.0	199.2	−14.2	628-63-7	training	146.0	159.2	−13.2	868-57-5	training	105.0	98.7	6.3
539-82-2	training	144.0	144.7	−0.7	629-33-4	training	155.0	175.4	−20.4	87-20-7	training	277.0	302.1	−25.1
539-90-2	prediction	51.0	79.5	−28.5	634-36-6	training	241.0	221.3	19.7	87-91-2	training	280.0	271.0	9.0
5405-41-4	training	170.0	195.2	−25.2	637-65-0	prediction	207.0	140.7	66.3	89-83-8	training	232.0	241.6	−9.6
542-55-2	training	98.0	128.0	−30.0	6378-65-0	prediction	89.0	96.7	−7.7	90-05-1	training	205.0	211.7	−6.7
544-63-8	training	326.0	325.1	0.9	638-11-9	prediction	145.0	153.4	−8.4	91-62-3	training	259.0	248.5	10.5
547-63-7	training	91.0	85.1	5.9	638-49-3	prediction	132.0	127.7	4.3	925-78-0	training	187.0	170.1	16.9
554-12-1	training	79.0	71.7	7.3	64-17-5	training	78.0	79.1	−1.1	928-95-0	training	157.0	160.9	−3.8
556-24-1	training	114.0	106.4	7.6	644-49-5	training	136.0	128.6	7.4	928-96-1	training	156.0	160.9	−4.9
556-82-1	training	140.0	140.8	−0.8	6485-40-1	training	191.0	232.4	−41.4	93-04-9	training	274.0	268.3	5.7
562-74-3	prediction	241.0	233.1	7.9	65-85-0	training	249.0	238.0	11.0	93-15-2	training	254.0	254.8	−0.8
564-94-3	training	220.0	218.1	1.9	659-70-1	training	192.0	192.4	−0.4	93-16-3	training	263.0	259.9	3.1
56-81-5	training	289.0	274.1	15.0	66-25-1	training	130.0	130.5	−0.5	93-51-6	training	221.0	227.0	−6.0
57-10-3	training	351.0	345.3	5.7	6728-26-3	training	145.0	145.7	−0.7	93-58-3	training	198.0	190.1	7.9
57-11-4	training	375.0	361.4	13.6	67-56-1	training	65.0	46.7	18.3	93-89-0	training	212.0	214.2	−2.2
5837-78-5	training	155.0	140.4	14.7	67-63-0	training	82.0	94.2	−12.2	94-30-4	training	263.0	246.5	16.5
5870-93-9	training	224.0	220.6	3.4	6789-88-4	training	272.0	278.5	−6.5	94346-09-5	training	181.0	220.1	−39.1
589-35-5	training	151.0	144.4	6.7	692-86-4	training	258.0	266.6	−8.6	95-48-7	training	191.0	182.5	8.6
589-66-2	prediction	101.0	94.6	6.4	71-23-8	training	97.0	92.1	4.9	96-17-3	training	91.0	91.9	−0.9
589-98-0	training	177.0	168.5	8.5	71-36-3	training	118.0	112.7	5.3	97-53-0	training	254.0	262.6	−8.6
590-01-2	training	147.0	142.0	5.0	71-41-0	training	137.0	135.0	2.0	97-61-0	training	196.0	181.0	15.0
590-86-3	training	91.0	96.4	−5.4	71-43-2	training	80.0	78.2	1.8	97-62-1	training	112.0	108.1	3.9
5910-89-4	training	156.0	152.2	3.8	7452-79-1	training	131.0	120.7	10.3	97-85-8	training	147.0	149.4	−2.4
591-93-5	training	26.0	57.6	−31.6	75-07-0	training	21.0	37.2	−16.2	97-96-1	training	117.0	104.1	13.0
592-41-6	training	63.0	70.8	−7.8	7549-33-9	training	277.0	265.2	11.8	97-99-4	training	178.0	173.8	4.2
592-84-7	training	106.0	136.0	−30.0	76-22-2	training	204.0	186.9	17.1	98-01-1	training	162.0	163.1	−1.1
5989-27-5	training	176.0	171.1	4.9	7774-44-9	training	223.0	209.9	13.1	98-55-5	training	219.0	208.9	10.1
5989-54-8	training	178.0	171.1	6.9	7778-87-2	training	208.0	204.4	3.6	98-85-1	training	204.0	208.3	−4.3
600-07-7	training	176.0	158.9	17.1	7779-70-6	training	260.0	263.2	−3.2	98-86-2	training	202.0	196.4	5.6
600-14-6	training	111.0	122.9	−11.9	7779-81-9	training	176.0	180.4	−4.4	99-86-5	training	175.0	159.6	15.4
60-01-5	training	297.0	266.2	30.9	7785-26-4	training	156.0	150.7	5.3	99-87-6	training	171.0	173.2	−2.2
60-12-8	training	219.0	224.5	−5.5	7785-70-8	training	157.0	150.1	6.9					
606-45-1	training	248.0	218.5	29.6	7786-58-5	training	250.0	243.2	6.8					

structures involved in a given training set are eliminated.The minimum limit of variation at this step was 10%. All remaining descriptors were examined for high pairwise correlation. If a given pair of descriptors yielded a correlation coefficient of 0.90 or greater, one of the pair was eliminated from further consideration. Selection of the descriptor to be retained was based on past experience and utility in model development. Descriptors remaining after OFS was performed were taken into the model development step.

Initial Training Set. A subset of 78 of the original 372 structures (21%) was selected at random and set aside to act as an external test (prediction) set (mean boiling point 199.9 °C, range 51.0−344.0 °C). The remaining 294 structures formed the initial model training set (mean boiling point 190.3 °C, range 21.0−431.0 °C). The OFS process was applied to the 129 available structural descriptors. A final subset of 45 descriptors remained following OFS.

The method of generalized simulated annealing (GSA)[38] was used to select the descriptors for the initial equations. Equations containing 9−19 descriptors were evaluated according to RMS error for the training set. The goal of this evaluation was to select the model that provided the best fit to the training set while using the smallest number of descriptors. This was done by plotting the RMS error as a function of the number of descriptors in the best GSA model of a given size. A 12-descriptor model was selected on this basis for further evaluation.

The results of the initial boiling point model are shown graphically in Figure 2. The model itself is summarized in Table 2. This initial 12-variable model provided a strong fit to the training set data, yielding an R^2 value of 0.921 and a standard deviation of regression (s) of 18.7 °C. The plot of the fitted and observed boiling points for the initial model given in Figure 2 shows the correlation to be reasonably linear over a range of 410 °C. Several of the observations deviate from the correlation. One particularly visible outlier is **104-21-2** (Figure 1) that has a reported boiling point of 114.0 °C. The estimated value produced by the initial model for this compound is 253.3 °C, yielding an error of estimate of −139.3 °C. To diagnose the error, the normal boiling point for **104-21-2** was estimated using the ACD/ILab on-line property estimation program available via the Internet.[40] The boiling point obtained using this method was estimated to be 251.4 ± 15.0 °C. This result is in good agreement with the value produced by the model. A second source reports the boiling point for this compound to be 270 °C,[41] which is also in good agreement with the model estimate. This particular type of error was commonly observed in several

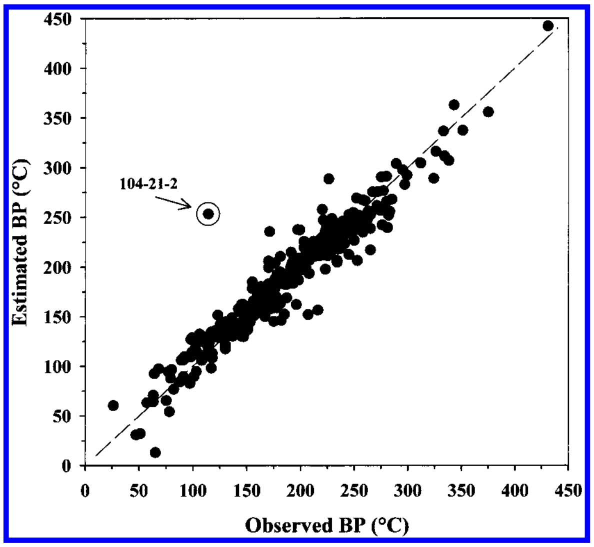

Figure 2. Comparison of the estimated and observed boiling points for the initial 294-observation model. One obvious outlier, 104-21-2, is identified.

Table 2. Details of the Initial Boiling Point Model Developed Using the 294-Observation Training Set Prior to Outlier Detection and Removal ($R^2 = 0.921$, $s = 18.74$, $N = 294$, overall F (for AOV) = 273.1, $F(12, 281, \alpha = 0.05) = 1.787$)

mol descriptor ID	regressn coeff	std dev of regressn coeff	partial F value
WNSA-2[a]	1.2901	0.1450	79.12
S6P[b]	24.12	5.691	17.97
RSHM[c]	293.4	50.39	33.90
WNSA-3[a]	−5.488	0.7566	52.62
ALLP-4[d]	90.67	7.568	143.6
DPSA-1[a]	0.4513	0.03445	171.6
PNSA-1[a]	2.369	0.1675	200.1
GEOH-3[e]	−49.62	7.119	48.59
FPSA-3[a]	1885	232.1	66.03
FPSA-1[a]	303.8	47.35	41.15
RPCS[a]	−3.704	0.5510	45.20
RNCG[a]	50.65	15.00	11.40
INTERCEPT	−579.0	48.33	145.0

[a] From the collection of CPSA descriptors.[37] [b] Sixth-order path molecular connectivity descriptor.[33] [c] From a collection of hydrogen-bond-specific CPSA descriptors.[10] [d] The total weighted number of paths in the structure within the specified lower and upper length limits (1 and 46, respectively).[39] [e] Molecular thickness, calculated using the ADAPT program DGEO.[20]

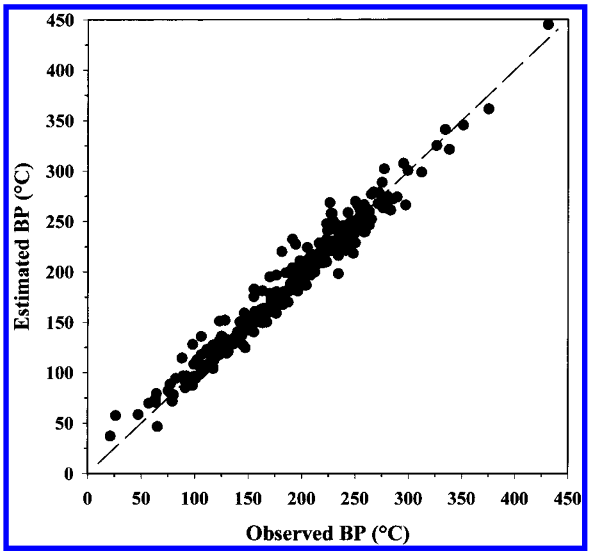

Figure 3. Comparison of the estimated and observed boiling points for the final 268-observation model.

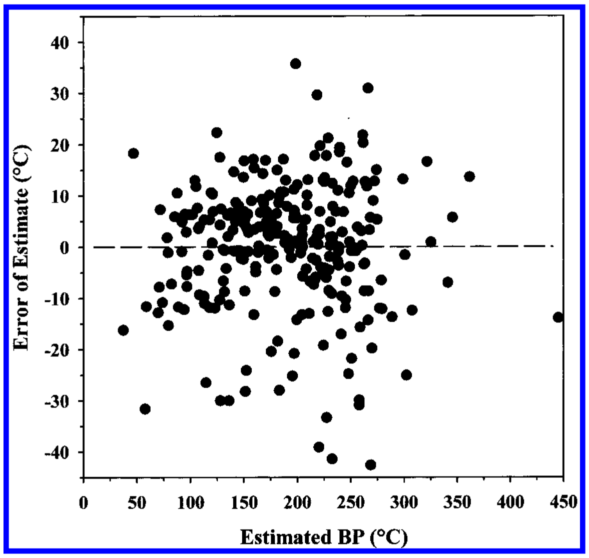

Figure 4. Residual plot for the final normal boiling point model.

other studies of this type.[9−11] The observation of large negative residuals in boiling point modeling studies is suggestive of either experimental or typographic errors involving the original boiling point data. It has been observed that it is not uncommon to find boiling points measured at reduced pressure mistakenly reported as normal boiling points (at 760 mmHg). It is very difficult to detect this type of error at the time the data set is assembled, especially if only one source of data is available. Such errors can produce a large bias in model development during the descriptor selection step. In the past, this problem was addressed by employing robust regression techniques[42,43] that are especially useful for identifying statistical outliers.

The initial model was analyzed using the robust regression techniques. This analysis identified a set of 26 compounds as outliers. These were removed from the training set and set aside for later analysis. The modeling process was repeated from the OFS step using the remaining 268-structure training set. The new 12-variable model yielded an improved fit to the data ($R^2 = 0.963$), and a significantly reduced standard deviation of regression ($s = 12.7$ °C). The correlation of the estimated and observed boiling points for the new model is illustrated graphically in Figure 3, and the scatter of the error is shown in the residual plot shown in Figure 4. Observed and fitted (estimated) boiling point values for the training set are shown in Table 1. The details of the model are provided in Table 3.

A review of the descriptors involved in the model show that much of the structural information that is important to explaining the observed boiling points is derived from the CPSA type of descriptors [PNSA-1 (PNSA = partial negative surface area), FPSA-3 (FPSA = fractional positive surface area), FNSA-1 (FNSA = fractional negative surface area), FNSA-3, WNSA-2 (WNSA = weighted negative surface area)]. One of the descriptors is from the set of CPSA-related hydrogen-bonding-specific descriptors [CNTH (count of hydrogen bond donor groups), and two atomic partial charge-related descriptors [QPOS (magnitude of charge on most positive atom) and RNCG (relative negative charge)]. The inclusion of these descriptors and the related CPSA descrip-

Table 3. Details of the Final Training Set Model Developed Using the 268-Observation Training Set ($R^2 = 0.963$, $s = 12.66$, $N = 268$, overall F (for AOV) = 552.6, $F(12, 255, \alpha = 0.05) = 1.790$)

mol descriptor id	regressn coeff	std dev of regressn coeff	partial F value	variance inflatn factor
S2[a]	54.63	2.287	570.4	16.7
FPSA-3[b]	899.0	205.0	19.24	9.3
S3C[c]	−50.64	4.259	141.4	7.3
FNSA-3[b]	−1024	72.61	199.1	4.5
WNSA-2[b]	1.761	0.09292	359.3	10.0
CNTH[d]	34.58	3.608	91.86	8.2
PNSA-1[b]	2.227	0.1556	204.9	41.2
FNSA-1[b]	−567.1	48.66	135.8	34.3
QPOS[e]	−121.3	12.55	93.45	1.9
L/B[f]	30.47	3.070	98.53	6.1
GEOH-2[g]	17.78	2.478	51.44	3.2
RNCG[b]	42.95	9.892	18.85	2.5
INTERCEPT	−150.4	12.80	138.1	N/A

[a] Second-order molecular connectivity.[33] [b] From the collection of CPSA descriptors.[37] [c] Third-order cluster molecular connectivity.[33] [d] From a collection of hydrogen-bond-specific CPSA descriptors.[37] [e] Sum of the positive partial atomic charges in the molecule. [f] Length-to-breadth ratio.[44] [g] Molecular width, calculated using the ADAPT program DGEO.[20]

tors in the model suggests the strong influence of polar intermolecular interactions on the observed boiling points. Molecular size and shape are also important as suggested by the presence of the two molecular connectivity indices S2 (X^2) and S3C ($X^3{}_c$),[33] and two geometry-based descriptors (GEOH-2 and L/B (length-to-breadth ratio)). These results are similar to what has been observed in the past for a collection of structures of this type, and are consistent in a physical sense with what one would intuitively expect.

Internal validation has shown the model both statistically significant and robust. The overall F value for analysis of variance[45] was 552.6 with a critical F value of 1.79 ($F(12, 256, \alpha = 0.05)$), and all the partial F values[46] are greater than the critical value of 1.79. An examination of both the fit plot (Figure 3) and the residual plot (Figure 4) shows a tight fit to the training set data with no apparent pattern or bias indicated in the errors of the estimates. Another internal test of the model is the calculation of the jack-knifed estimates for the training set members (cross-validation).[47] In this test, each observation is held out, one at a time, and the coefficients of the model are recomputed. The boiling point for the structure held out is then estimated using the model developed on the $n - 1$ observation training set. This test examines the homogeneity of the training set and determines if any single structure has undue influence in determining the model coefficients. A plot showing the correlation of the jack-knifed-estimated and observed boiling points is shown in Figure 5. The jack-knifed estimates yield a cross-validated R^2 of 0.955. These results reinforce the idea that the model is robust, and further shows that no single observation has undue influence on the model. Next, attention was turned to the detection of any problems related to collinearity among the descriptors in the model. The variance inflation factor (VIF) for a given descriptor is a measure of the effect of collinearity among the descriptors on the model.[48] It is suggested that the VIF for a given descriptor should be below 10, and that the mean VIF value for the model should be around 1. The VIF values for each descriptor are given in Table 3. The greatest single VIF value was 41.8 (for the descriptor PNSA-1), with a mean VIF value of 12.5. In working with other large and diverse sets of compounds, it has been observed that VIF values such as these are common, and the resulting models still perform well in prediction. However, as a further check of the potential for problems regarding collinearity, the descriptors were subjected to a further analysis using partial least-squares (PLS) regression.[49,50] The goal of such an analysis is 2-fold. First, the PLS analysis results will indicate problems due to collinearity among the descriptors by finding that the model can be validated using fewer latent variables in the PLS model than were used in the ordinary least-squares (OLS) model. Second, the weighting of the individual original descriptors within the latent variables acts as an indication of the relative importance of these descriptors in the structure−property relationship implied in the model.

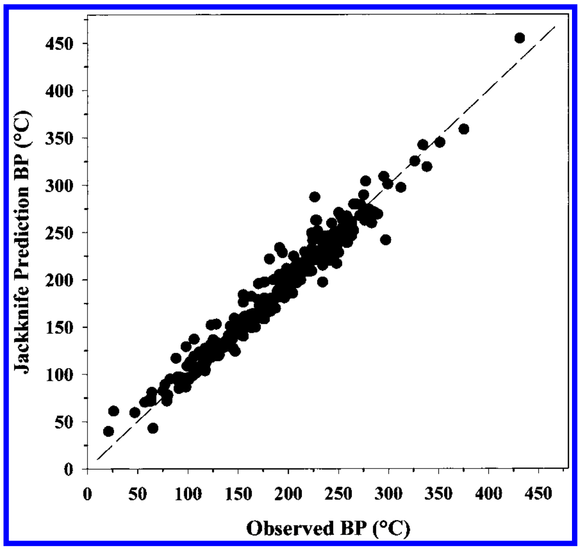

Figure 5. Comparison of the observed and jack-knife (cross-validation) predicted boiling point values for the final model.

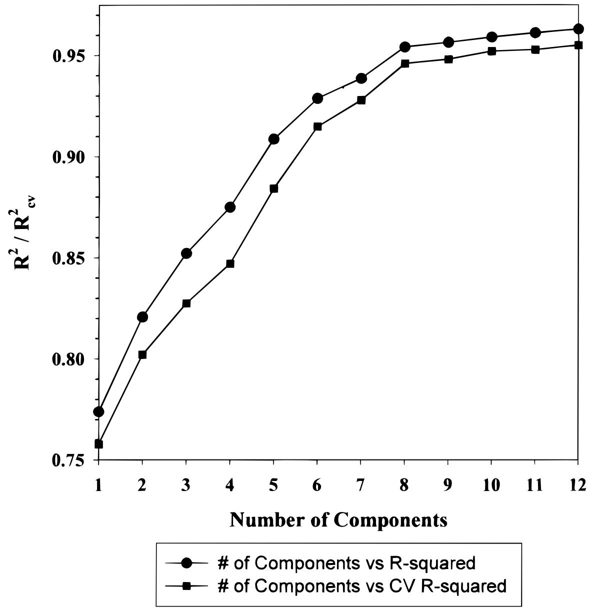

Figure 6. Comparison of the increase of R^2 from OLS and $R_{cv}{}^2$ from PLS regression as the number of PLS components used in the model increases.

Results of the PLS analysis are shown graphically in Figure 6. It is clear that the fitted and the cross-validated (leave-1-out method) R^2 values are correlated and that they

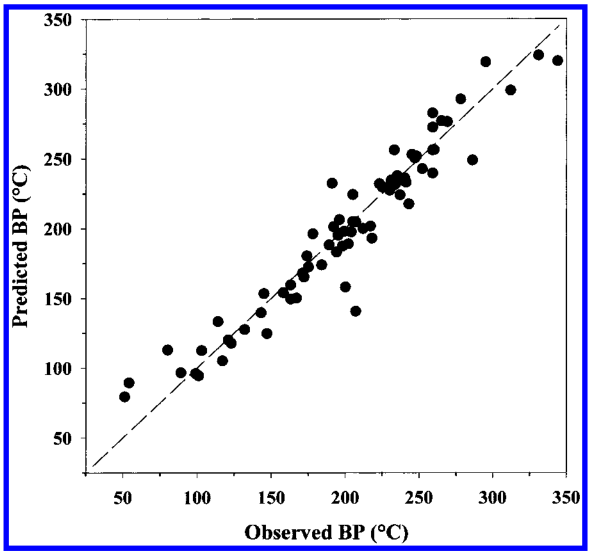

Figure 7. Comparison of the predicted and observed normal boiling points for the 78-observation external prediction set.

do not plateau before reaching the 12th component. In situations where collinearity was truly a problem in the ordinary least-squares (OLS) model, the PLS cross-validated R^2 value would be observed to either plateau or decrease. It is also gratifying to note that the final cross-validated R^2 values from both the OLS and PLS analyses are in such good agreement. When taken together, these results suggest that while some collinearity exists it does not degrade the validity of the model.

The final test of the model is accomplished by evaluating its strength as a predictive tool. Therefore, the model was applied to predict the normal boiling points of the 78-observation external validation (test) set. The model performed well in external prediction, yielding a correlation coefficient for the predicted and observed boiling points of 0.962. The results for the external prediction set are shown graphically in Figure 7, and a comparison of the observed and predicted boiling point values is given in Table 1.

As a final step, attention was directed to the 26 compounds indicated as outliers in the initial boiling point model training set. Care must be exercised when dealing with outliers. One can always obtain an improved statistical fit to data by removing outlying observations. The existence of the outliers could just be the result of an inadequate model. It is therefore important to determine why a given observation is an outlier, and to possibly learn something about the limitations of the final equation. The boiling points for the 26 outlying observations were predicted using the final 12-variable model. A comparison of the predicted and observed boiling points yielded a correlation coefficient of 0.731. The comparison is shown graphically in Figure 8. While the model actually seems to have performed well for some of these compounds, there are others where the performance is quite poor. It was important to determine why these compounds were considered outliers in the first place.

Table 4 lists the observed and predicted boiling points and the prediction errors for the 26 outliers. The structures for these compounds are given in Figure 9. The last column in the table lists the *leverage* values for each of the observations in the initial model. Leverage values are defined as the values that comprise the diagonal component of the Hat matrix derived during the OLS regression analysis process.[51] It can be thought of as a measure of the influence of a given observation on its own prediction. Since the leverage values are derived only from the descriptor data (the independent variables), this diagnostic tool focuses only on the structure of the compounds in the training set and not on the property being modeled. The critical value for identifying an observation as having high leverage is typically taken as $2p/n$, where p is the number of parameters in the model (the number of coefficients, including the y intercept) and n is the number of observations (structures) in the training set. In the case

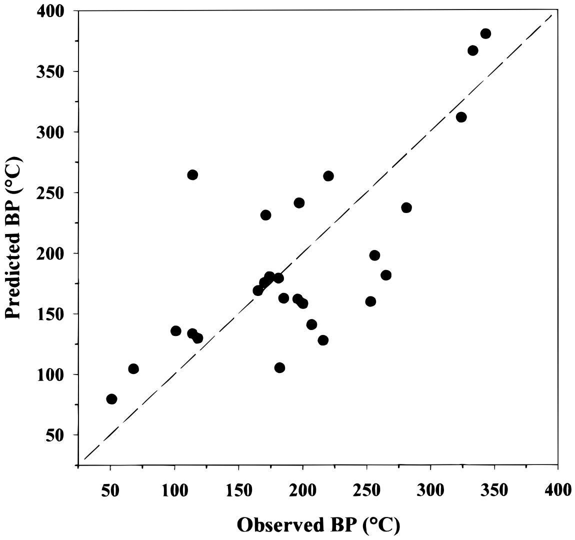

Figure 8. Comparison of the predicted and observed normal boiling points for the set of 26 outliers that were set aside after examination of the initial boiling point model.

Table 4. Observed and Predicted Values for the 26 Outliers Detected during the Development of the Initial Boiling Point Model[a]

CAS reg no.	obsd bp (°C)	pred bp (°C)	predictn error	leverage
15707-24-1	181	179.1	1.9	0.015
110-41-8	171	231.1	−60.1	0.023
625-55-8	68	104.5	−36.5	0.025
104-21-2	114	264.3	−150.3	0.027
706-14-9	281	236.9	44.2	0.029
90-02-8	197	241.0	−44.0	0.034
98-00-0	170	175.4	−5.4	0.047
122-91-8	220	263.1	−43.1	0.048
99-85-4	182	105.3	76.7	0.049
586-62-9	185	162.5	22.5	0.054
108-29-2	207	140.7	66.3	0.058
83-34-1	265	181.3	83.7	0.068
127-17-3	165	168.8	−3.8	0.082
134-20-3	256	197.6	58.4	0.082
120-72-9	253	159.6	93.4	0.085
64-19-7	118	129.7	−11.7	0.091
120-51-4	324	311.4	12.6	0.094
493-01-6	196	161.8	34.2	0.096
143-28-2	333	366.2	−33.2	0.101
106-49-0	200	158.1	41.9	0.128
541-35-5	216	127.8	88.2	0.174
107-22-2	51	79.5	−28.5	0.195
106-46-7	174	180.3	−6.3	0.237
107-19-7	114	133.4	−19.4	0.239
123-95-5	343	380.1	−37.1	0.240
64-18-6	101	135.7	−34.7	0.269

[a] The predicted values were calculated using the final (268-observation) boiling point model.

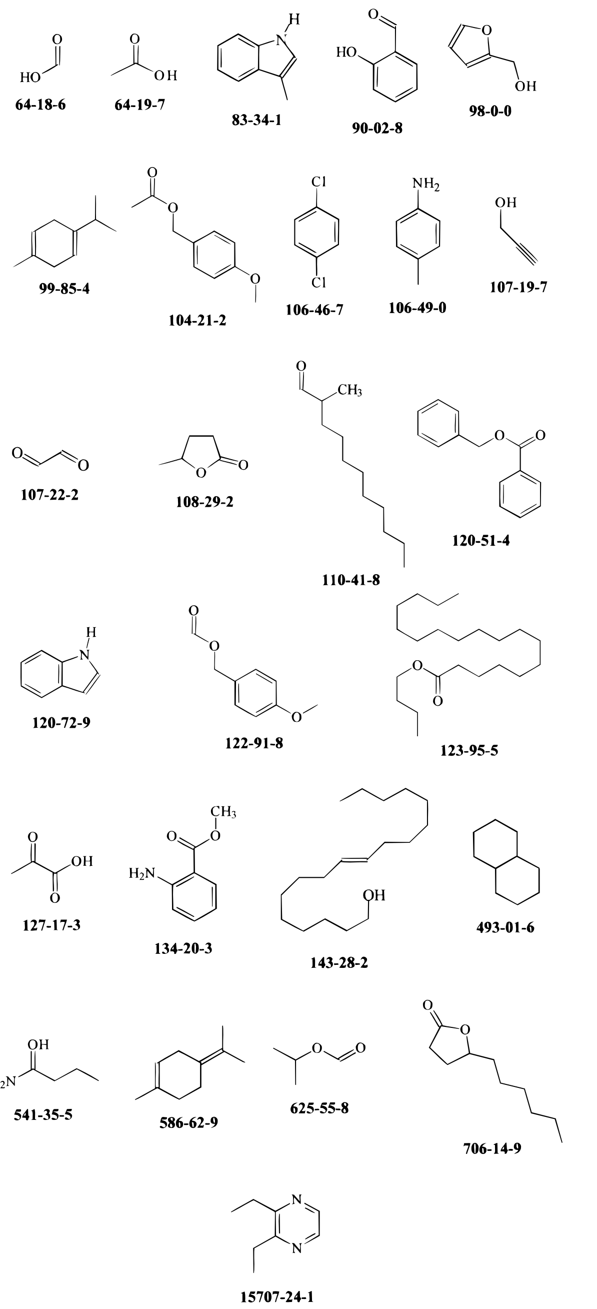

Figure 9. Structures of the 26 outliers detected during the model development process.

of the initial model, $p = 13$ and $n = 294$, yielding a critical leverage value of 0.088. Only 11 of the 26 outliers exceed this value. The structures for these 11 compounds are observed to be atypical compared to the majority of the training set. Several are either very small or are very large, placing them on the fringes of the training set, which increases their influence on each step in the modeling process. One of the high-leverage compounds (106-46-7) contained two chlorine atoms, and was the only structure in the training set that contained any halogen atoms. Another outlier (493-01-6) contains no heteroatoms, and is quite unlike the rest of the training set in that respect. Finally, two other high-leverage outliers (106-49-6, 541-35-5) contain amine groups, another functional group that was not well represented in the original data set. Therefore, it is not surprising that these structures were detected as structural outliers and were justifiably removed.

The most interesting and informative results come from the structures yielding low leverage values. Since these structures have low leverage values, they are not considered structurally dissimilar to the training set with respect to the 12 descriptors used in the model. The identification of these compounds as outliers by robust regression is based on the dependent variable, the boiling point. Most of the prediction errors for these outliers are large (RMS error for this subset 65 °C). The source of error for this type of outlier can be attributed either to the observed boiling point data or to structural features of the molecule that are not properly encoded in the model but that have a large influence on the observed boiling point. The sign of the error is sometimes instructive. As was mentioned above, errors in the observed boiling points are often introduced when the pressure data are not carefully considered. Such errors manifest themselves as large negative residuals, as that observed for 104-21-2. There are five outliers listed in Table 4 that have both low-leverage values and large negative prediction errors. Efforts to obtain additional data for these compounds show that the reported boiling points used for three of these (104-21-2, 110-41-8, and 122-91-8) are apparently incorrect. The error for 104-21-2 has already been discussed. The special case of 122-91-8 is discussed below. The discrepancy with 110-41-8 was harder to detect. The ACD/ILAB approach produced a boiling point of 171 °C, listing it as an experimental value and not an estimate. However, another source reported the boiling point for the compound as 114 °C at 10 mmHg.[40] This value was corrected to standard pressure using a published method[52] that yielded an estimate of 253 °C. A second estimate was calculated using the method of Pearson.[4] Starting with the reported boiling point for 2-methylpentanal of 117 °C,[40] an estimated boiling point of 254 °C was obtained. Given that the two estimation methods, using different techniques, arrive at nearly the same value, and that these values are in agreement with the current model, the evidence suggests that the reported normal boiling point of 171 °C is incorrect.

Boiling points of the remaining two compounds with large negative residuals (90-02-8 and 625-55-8) were confirmed, and the model overestimates the boiling points for these compounds. In the case of 625-55-8, the boiling point could not be explained outside the observed variation in the model. The range of the residuals for the training set compounds is $-42.6 + 35.7$ °C, and the residual for 625-55-8 falls in that range. Additionally, with the other outliers removed and 625-55-8 recombined with the training set, the compound is no longer detected as an outlier using robust regression analysis. One compound, 90-02-8 (2-hydroxybenzaldehyde), clearly illustrates the impact that intramolecular hydrogen bonding has on observed properties. The carbonyl oxygen of the aldehyde group and the hydrogen atom from the adjacent hydroxyl group can form a hydrogen bond that produces a stable six-membered ring. The result is that the intermolecular interactions are weaker in the bulk phase. This produces a reduction of the observed boiling point compared to 3-hydroxybenzaldehyde (normal bp 247 °C[40]), which is structurally very similar, but which cannot form the intramolecular hydrogen bond. So the source of the error for 90-02-8 is identified as the lack of information in the model concerning the identification and geometric arrangement of structural features responsible for intramolecular hydrogen bonding. Another compound that potentially exhibits a decreased boiling point due to intramolecular interactions is 122-91-8. The terminal hydrogen of the aldehyde side chain is reported to interact with the π-electrons of the phenyl ring, forming a hydrogen bond.[53] It was initially thought that the observed boiling point for this compound was incorrect, on the basis of the estimated boiling point produced using ACD/ILAB (calculated bp 254.8 ± 15 °C). While the ACD estimate of the boiling point agrees with the new model's estimate, both estimation methods may ignore the potential for formation of this type of intramolecular hydrogen bond. In addition, it is reported that both the folded and unfolded conformers of 122-91-8 coexist, so that the effect of the interaction is diminished somewhat compared to the strong intramolecular interaction observed for 90-02-8. However, whether the error lies with the experimental data, or with a limitation of the model, the outlier detection process worked well in identifying 122-91-8 as an unusual case.

Outliers exhibiting both low leverage and large positive prediction errors usually point to deficiencies in the model. Since the leverage is low, the compounds do not look like structural outliers with respect to the descriptors in the model. The fact that the boiling point is much higher than expected suggests that there are some intermolecular interactions not properly captured by the descriptors used in the model. An increase in the number or the strength of such interactions would be manifested by higher than expected boiling points. However, if the structural features responsible for these interactions are not well represented in the training set, and are not accounted for by the descriptors used, then the compounds will not appear as structural outliers in robust regression based on leverage values. An examination of the compounds yielding low-leverage values and large positive prediction errors shows that many have relatively unusual functional groups with respect to the rest of the training set. For example, there are several amine-containing compounds included in this subset. Amines have two hydrogens that can be involved in hydrogen-bonding interactions, whereas hydroxyl groups, which are much more common in the training set, provide only one donatable hydrogen. The extra interactions on the part of the amines can explain the unexpected increase in the boiling points. Also, several of the compounds with large positive estimation errors have relatively high leverage values. The critical value for leverage is considered a rule of thumb only, but the magnitude of the leverage value does suggest that such compounds are relatively unusual compared to the training set.

Outliers detected during the initial phase of model development can be used to diagnose problems in both the data set and the model itself. However, even given the few deficiencies observed, the new model still performs quite well for a wide variety of structure types, and yields accurate predictions of normal boiling points for these types of compounds.

CONCLUSIONS

The overall goal of this work was to develop a mathematical model based solely on measures of molecular structure that would provide accurate predictions of normal boiling points of small multifunctional organic compounds. The model described above was found to be internally valid and to yield accurate predictions for a large external set of 78 structures. The mean error for the training set boiling point estimates was 12.3 °C, or 6.5% of the mean boiling point for the data set. The mean prediction error for the external prediction set was found to be 16.7 °C, or 8.3% of the mean boiling point for those compounds. These results are quite good given the large diversity of the structures involved and the degree of polar intermolecular interactions that are expected in the liquid phase. Not only does the model provide accurate boiling point values for new structures, but the descriptors involved capture information about the structural features that is consistent with what one would expect for data sets of this type. Both aspects are required to develop confidence in the results obtained using the model. Additionally, an analysis of the compounds found to be outliers during the modeling process provides information concerning possible deficiencies in both the model and the original data set. Such information allows attention to be focused on areas for future improvement of the model.

ACKNOWLEDGMENT

I express my appreciation to Dennis R. Bacon, Cheryl Glendening, and Raymond A. Crawford of Procter & Gamble (P&G) for assembling the data used in this study. Appreciation is also expressed to Joel D. Oliver and Raymond A. Crawford of P&G for reviewing this manuscript and providing constructive feedback.

Supporting Information Available: Structures for the 372 compounds used in this study (training and prediction set structures and outliers). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Pailhes, F. Estimation of the Boiling Temperature at Normal Pressure for Organic Compounds from their Chemical Formula and a Known Boiling Temperature at Low Pressure. *Fluid Phase Equilib.* **1988**, *41*, 97−107.

(2) Bermejo, J.; Blanko, C. G.; Guillen, M. D. Capillary Gas Chromatography of Chloro Derivatives of 1,4-Dimethylbenzene. *J. Chromatogr.* **1985**, *331*, 237−243.

(3) White, C. M. Prediction of the Boiling Point, Heat of Vaporization, and Vapor Pressure at Various Temperatures for Polycyclic Aromatic Hydrocarbons. *J. Chem. Eng. Data* **1986**, *31*, 198−203.

(4) Pearson, D. E. A Method of Estimating the Boiling Points of Organic Liquids. *J. Chem. Educ.* **1951**, *28*, 60−62.

(5) Joback, K. G.; Reid, R. C. Estimation of Pure-Component Properties from Group Contributions. *Chem. Eng. Commun*. **1987**, *57*, 233−243.

(6) Lai, W. Y.; Chen, D. H.; Maddox, R. N. Application of a Nonlinear Group Contribution Model to the Prediction of Physical Constants. 1. Predicting Normal Boiling Points with Molecular Structure. *Ind. Eng. Chem. Res.* **1987**, *26*, 1072−1079.

(7) Copeman, T. W.; Mathias, P. M.; Klotz, H. C. Industrial Use of Group Contribution Methods for Estimation of Physical Properties. Presented at the Beilstein Workshop on the Estimation of Physical Data for Organic Compounds, Schloss Korb, Italy, May 1988.

(8) Le, T. D.; Weers, J. G. Group Contribution-Additivity and Quantum Mechanical Models for the Predicting the Molar Refractions, Indices of Refraction, and Boiling Points of Fluorochemicals. *J. Phys. Chem.* **1995**, *99*, 13909−13916.

(9) Stanton, D. T.; Jurs, P. C., Hicks, M. G. Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans, and Thiophenes. *J. Chem. Inf. Comput. Sci.* **1991**, *3l*, 301−310.

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

(11) Egolf, L. M.; Jurs, P. C. Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 616−625.

(12) Egolf, L. M.; Wessel, M. D.; Jurs, P. C. Prediction of Boiling Points and Critical Temperatures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 947−956.

(13) Wessel, M. D.; Jurs, P. C. Prediction of Normal Boiling Points of Hydrocarbons from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 68−76.
(14) Wessel, M. D.; Jurs, P. C. Prediction of Normal Boiling Points for a Diverse Set of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 841−850.
(15) Le, T. D.; Weers, J. G. QSPR and GCA Models for Predicting the Normal Boiling Points of Fluorocarbons. *J. Phys. Chem*. **1995**, *99*, 6739−6747.
(16) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Comparative Study of Topological and Geometric Parameters in Estimating Normal Boiling Point and Octanol/Water Partitions Coefficient. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 1054−1060.
(17) 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.
(18) Ivanciuc, O,; Ivanciuc, T.; Balaban, A. T. Quantitative Structure−Property Relationship Study of Normal Boiling Points for Halogen-/ Oxygen-/ Sulfur-Containing Organic Compounds Using the CODES-SA Program. *Tetrahedron* **1998**, *54*, 9129−9142.
(19) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure−Property Relationship. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 28−41.
(20) Stuper, A. J.; P. C. Jurs. ADAPT: A Computer System for Automating Data Analysis using Pattern-Recognition Techniques. *J. Chem. Inf. Comput. Sci*. **1976**, *2*, 99−105.
(21) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, R. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, pp 103−129.
(22) *Sybyl Version 5.5 Theory Manual*; Tripos: St. Louis, MO, 1992; p 3027.
(23) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(24) Pearlman, R. S. In *3D QSAR in Drug Design*; Kubini, H., Ed.; ESCOM: Amsterdam, 1993; pp 41−79.
(25) *Sybyl Version 6.3 Force Field Manual*; Tripos: St. Louis, MO, 1996; p 196.
(26) 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, 1996; p 290.
(27) Properties of Organic Compounds Database, CD-ROM Version 5.0, CRC, Boca Raton, FL.
(28) Flavors and Fragrance, Aldrich Chemical Co., Milwaukee, WI, 1995.
(29) Design Institute for Physical Property Data, American Institute of Chemical Engineers, STN database online search.
(30) Beilstein Handbook of Organic Chemistry, Beilstein Information Systems, STN database online search.
(31) *Perfume and Flavor Chemicals*; Steffen Arctander, Las Vegas, NV, 1969; Vols. I and II.
(32) Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc*. **1975**, *97*, 6609−6615
(33) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, 1976.
(34) Pearlman, R. S. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980.
(35) Brugger, W. E.; Stuper, A. J.; Jurs, P. C. Generation of Descriptors from Molecular Structures. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 105−110.
(36) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure−Property Relationships. *J. Comput. Chem.* **1992**, *13*, 492−504.
(37) 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* (2), 2323−2329.
(38) Sutter, J. M.; Jurs, P. C. Selection of Molecular Descriptors for Quantitative Structure−Activity Relationships. *Data Handl. Sci. Technol.* **1995**, *15*, 111−132.
(39) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for all Self-Avoiding Paths for Molecular Graphs. *Comput. Chem.* **1979**, *3*, 5−13.
(40) Calculated using ACD/I-Lab (July 1997 version, www.acdlabs.com/activelab/). Advanced Chemistry Development, Inc., 133 Richmond St. W., Suite 605, Toronto, Ontario M5H 3L2, Canada.
(41) *CRC Handbook of Chemistry and Physics*, 78th ed.; Lide, D. R., Frederikse, H. P. R., Eds.; CRC: Boca Raton, FL, 1997; Section 3-1.
(42) Massart, D. L.; Kaufman, L.; Rousseeuw, P. J.; Leroy, A. Least-median of squares: A robust method for outlier and model error detection in regression and calibration. *Anal. Chim. Acta* **1986**, *187*, 171−179.
(43) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*; John Wiley and Sons: New York, 1987.
(44) Kaliszan, R. *Quantitative Structurechromatographic Retention Relationships*; John Wiley & Sons: New York, 1987; p 128.
(45) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*, 2nd ed.; Irwin: Homewood, IL, 1985; p 240.
(46) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*, 2nd ed.; Irwin: Homewood, IL, 1985; p 281.
(47) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*; John Wiley and Sons: New York, 1987; p 226.
(48) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*, 2nd ed.; Irwin: Homewood, IL, 1985; p 391.
(49) Stahle, L.; Wold, S. In *Progress in Medicinal Chemistry*; Ellis, G. P., West, G. B., Eds.; Elsevier: Amsterdam, 1988; Vol. 25, pp 291−338.
(50) Glen, W. G.; Dunn, W. J., III; Scott, D. R. Principal Components Analysis and Partial Least Squares Regression. *Tetrahedron Comput. Methods* **1989**, *2*, 349−376.
(51) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*, 2nd ed.; Irwin: Homewood, IL, 1985; p 402.
(52) *CRC Handbook of Chemistry and Physics*, 53rd ed.; Weast, R. C., Ed.; CRC: Boca Raton, FL, 1972; p D-144.
(53) Suezawa, H.; Mori, A.; Sato, M.; Ehama, R.; Akai, I.; Sakakibara, K.; Hirota, M.; Nishio, M.; Kodama, Y. Evidence for the Presence of CH-π-Interacted ap-Conformers of Benzyl Formates. *J. Phys. Org. Chem.* **1993**, *6*, 399−406.

CI990311X