Prediction of Critical Temperatures and Pressures of Industrially Important Organic Compounds from Molecular Structure

Brian E. Turner, Chandra L. Costello, and Peter C. Jurs*

Department of Chemistry, Penn State University, 152 Davey Laboratory, University Park, Pennsylvania 16802

Received January 23, 1998

Quantitative−structure property relationships methods are used to develop mathematical models to predict critical temperatures and pressures of a diverse set of organic compounds taken from the Design Institute for Physical Property Data (DIPPR) database. Each compound is represented with calculated molecular structure descriptors that encode its topological, electronic, geometrical, and other features. Subsets of descriptors are selected with simulated annealing and genetic algorithms. Models to predict the critical properties are constructed using multiple linear regression analysis and computational neural networks with errors comparable to the experimental errors of the critical property data.

INTRODUCTION

Quantitative structure−property relationship (QSPR) methods have been used to develop models for the prediction of two fundamental physical properties of organic compounds: critical temperature and pressure. The critical temperature is the temperature above which a gas cannot be liquefied, and the minimum pressure required for liquefaction at this temperature is the critical pressure. Knowledge of these critical properties for organic compounds is very important to industrial chemical engineers. The values of critical temperatures and pressures are needed in equation of state calculations of thermodynamic and transport properties. In addition, these physical properties are important in high-pressure phase equilibrium processes. Such data are needed, for example, for studies of enhanced oil recovery and supercritical fluid extraction.[1]

As a result of the importance of these properties, a number of different approaches have been taken in order to calculate the values of critical temperatures and pressures. One common approach is to use knowledge of other physical properties. An example of this method is the estimation of the critical point using the Peng−Robinson equation of state.[1] This method uses the van der Waals volume, the normal boiling point, and the acentric factor of a compound as variables. A drawback of such a method is the reliance on experimental data which can be expensive or time-consuming to gather.

An alternative approach is to build models to predict critical temperatures of organic compounds using structural information alone.[2] This method employs quantitative structure−property relationships (QSPRs) to develop its predictions. Different methodologies have been reported to develop QSPRs. One such approach has been through the use of the Automated Data Analysis and Pattern recognition Toolkit (ADAPT).[3] This interactive software system utilizes multiple linear regression analysis (MLR) and computational neural networks (CNN) as model-building routines to produce QSPRs. A goal of the present study is to use these methods to improve on the prediction capabilities using QSPR for critical temperatures.

In general, development of a QSPR involves three steps: structural encoding, feature selection, and model building. Structural encoding involves the use of numerical descriptors to encode the structural features of a compound. Feature selection is then employed to determine which subset of the descriptors best relates to the property of interest. Models built from the best subset of descriptors form a direct link between descriptors and the property of interest. Finally, validation determines the level of the model's predictive capabilities for unknown compounds.

Recently, several changes in the ADAPT software have lead to the belief that improvements have been made in the QSPR methodology.[4−6] Several new approaches for numerically encoding a compound's structure have been implemented. For example, development of carbon type descriptors allows for structural descriptors such as the number of secondary sp^3 carbons to be determined, and hydrogen-bonding descriptors encode the ability of compounds to engage in hydrogen bonding interactions. The development of better descriptors is a continuing goal of QSPR studies.

In addition to new descriptors, the ADAPT software system has had updates in its approach to subjective feature selection.[7,8] Previously, the method of leaps-and-bounds regression was applied in order to build regression models. However, new developments have allowed a genetic algorithm and simulated annealing to replace the old feature selection methods. These new means of feature selection improve results in QSPR models because of their capability to systematically search through a larger number of descriptors and discard from consideration any variables that do not contribute to the model being generated.

The development of new models for predicting critical temperature is the first step here. Once assessments of the methodology are made, the second goal of the study is to develop a predictive equation for critical pressures using the same set of compounds. Together, the two equations can support the prediction of the critical point for various types of organic compounds.

S0095-2338(98)00005-5 CCC: $15.00
Published on Web 05/20/1998

METHODOLOGY

The Data Set. Critical temperatures and pressures were obtained for a set of 165 compounds from the DIPPR database.[9] The functional groups found among the data set compounds include alcohols, ketones, esters, carboxylic acids, aldehydes, phenols, ethers, nitriles, and amines. The hydrocarbons include alkanes, alkenes, benzenes, cycloalkanes, and cycloalkenes. A complete list of the compounds in the data set and their corresponding experimental critical temperatures and pressures is shown in Table 1. The data set contained 76 hydrocarbons, 61 oxygen-containing compounds, 23 nitrogen-containing compounds, and 5 halogen-containing compounds. The compounds range in size from propylene (molecular weight of 42) to quinoline (molecular weight of 129). For the linear regression studies, the data set was divided into a training set (tset) of 147 compounds for model building and an external prediction set (pset) of 18 compounds for model validation. Both subsets were chosen to ensure that a diverse set of compounds was present. For the computational neural network (CNN) studies, a cross-validation set (cvset) of 15 compounds was chosen, leaving 132 compounds in the tset, and the pset remained the same. The cvset is a subset of compounds used to help find an optimal set of weights and biases during CNN training, and it is also used to avoid overtraining of the CNN.

The range of values for both critical temperatures and pressures was large, with the critical temperatures ranging from 365 to 782 K and the critical pressures ranging from 11.94 to 55.47 atm. The experimental uncertainties associated with the values used in this data set are determined by the members of the DIPPR staff. Calculations show that the experimental critical temperatures and pressures in this study have a 1.3% and 1.5% error, respectively. These uncertainties indicate that the ideal errors associated with our models should be approximately 8 K and 0.5 atm.

Molecular Modeling. The 165 compounds were sketched into HyperChem[10] and transferred to the workstation in the form of connection tables. A more extensive energy minimization was done using MOPAC[11] with the PM3 Hamiltonian.[12] The resulting three-dimensional conformations were necessary so that they could be used later to calculate geometrical descriptors.

Descriptor Generation. Once the structures were entered, molecular structure descriptors were generated. Three categories of descriptors were calculated: topological, geometrical, and electronic. Topological descriptors include values such as the total number of atoms, the number of double bonds, molecular connectivity indices, and path counts. Geometrical descriptors encode structural features such as the surface area of a compound, its moments of inertia, and its molecular volume. Electronic descriptors provide information about such features as the charge on various atoms within the molecule or dipole moment. In addition to these descriptors, hybrids known as charged partial surface area (CPSA) descriptors use both geometrical and electronic information to encode the compounds' ability to engage in polar interactions.[13] Finally, other hybrid descriptors account for the effects of hydrogen bonding. A total of 192 descriptors was calculated for each compound in the data set.

Descriptor Screening. The process of descriptor generation produced a large number of descriptors for each compound. However, to reduce the likelihood of chance correlations being found during model development, the number of descriptors had to be reduced to be less than 60% of the number of observations. The reduction of the number of descriptors involved a series of steps. First, all descriptors with 90% identical values were removed because they did not offer valuable information that allowed for discrimination between compounds. Next, when pairs of descriptors were found to have pairwise correlations greater than 0.95, one of them was removed, while the other remained in order to maintain a small pool of information rich descriptors. Finally, the descriptor pool was reduced using vector space descriptor analysis (VSDA)[14] by ranking the descriptors based on an orthogonalization procedure. The descriptors with the lowest ranking were removed until the pool was small enough.

Regression Analysis. After a reduced pool of descriptors was developed, multiple linear regression analysis was used to develop a model that predicted the property of interest:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n$$

In this equation, Y is the value of the physical property of a compound, b_0 is the y-intercept of the regression model, b_1 through b_n are the various coefficients for the descriptors determined by the regression model, and X_1 through X_n are the specific descriptor values for the compound.

To formulate regression models, two methods of feature selection were utilized. A genetic algorithm and simulated annealing allowed for regression models of various sizes to be built from the larger pool of descriptors. Models of different sizes were investigated until the inclusion of an additional descriptor did not significantly improve the model statistics. Once developed, an attempt to improve the model was done using an interactive regression analysis (IRA) routine. This program allowed individual, specific descriptors to be added or removed in order to seek the optimum model statistics. When adding or changing a descriptor no longer improved the rms error of the model, it was determined that the best subset had been found.

Once a model was developed, regression diagnostics were performed in order to determine if any compounds were statistical outliers. Examples of these tests include studentized residual, leverage value, and Cook's distance.[15] Four out of six of these statistical tests must fail in order for a compound to be classified as an outlier. If an outlier is found, the compound is dropped from the tset, and the model coefficients are recalculated.

Model Validation. Once a regression model was selected, several investigations were conducted to determine if the model was statistically valid. The first thing examined was basic model statistics such as F, T, and p-values. Next, the model was tested for multicollinearities using the variance inflation factor (VIF).[16] A VIF greater than 10 for a descriptor generally indicates the presence of an unacceptably large multicollinearity with other descriptors. Graphical approaches were then used to search for bias. To do this, experimental versus predicted plots and residual plots were constructed and examined for trends. Finally, an external pset, which was not used during feature selection or

Table 1. Experimental and Predicted Values of Each Compound for All Four Models

	compound name	exptl pressure	linear predicted	CNN model predicted	exptl temp	linear predicted	CNN model predicted
1	3-chloropropene	46.48	46.66	46.72	514.5	505.8	508.8
2	propylene	45.52	44.40	45.68	364.8	381.5	389.5
3	allyl alcohol[b]	55.47	50.55	52.38	545.0	532.4	538.5
4	ethyl formate	46.80	43.94	43.66	508.4	514.5	516.1
5	methyl acetate	46.88	44.75	44.27	506.8	492.7	503.6
6	propionic acid[a]	45.57	48.28	47.56	604.0	600.1	592.0
7	*n*-propyl chloride	45.20	43.66	44.33	503.1	495.7	501.5
8	propane[b]	41.92	40.26	40.96	369.8	366.5	384.6
9	methyl ethyl ether	43.42	42.51	42.15	437.8	432.4	434.2
10	propanol	51.02	47.53	48.95	536.7	521.6	531.5
11	methylal	39.00	41.34	41.12	480.6	487.3	491.2
12	isopropylamine	44.80	47.24	45.32	471.8	465.1	467.1
13	*n*-propylamine	46.78	44.31	45.93	496.9	494.3	501.7
14	trimethylamine	40.20	45.43	39.39	433.3	443.3	438.1
15	1,3-butadiene	42.73	44.43	44.95	425.4	449.1	436.2
16	*n*-butyronitrile	37.40	43.32	41.85	582.3	595.4	595.2
17	1-butene	39.67	39.96	40.77	419.6	433.4	427.7
18	*cis*-2-butene	41.51	39.67	39.93	435.6	430.4	424.3
19	*trans*-2-butene	40.49	38.89	39.85	428.6	430.2	424.7
20	isobutene	39.47	43.71	41.41	417.9	417.0	410.4
21	methyl ethyl ketone	41.00	41.58	40.90	535.5	522.1	523.2
22	tetrahydrofuran	51.20	47.91	50.15	540.1	520.9	525.0
23	*n*-butyric acid	40.11	42.24	41.73	628.0	626.6	623.3
24	ethyl acetate	38.29	38.05	37.49	523.3	514.2	527.8
25	isobutyric acid	40.00	43.25	41.29	609.1	621.7	612.8
26	methyl propionate	39.52	39.40	39.88	530.6	531.3	532.8
27	*n*-propyl formate[b]	40.10	40.24	39.67	538.0	543.5	538.5
28	*n*-butane	37.47	36.20	36.96	425.2	423.9	421.7
29	isobutane	36.00	37.18	35.12	408.1	390.3	398.0
30	*n*-butanol	43.65	43.75	44.58	562.9	557.6	554.5
31	*sec*-butyl alcohol	41.39	43.82	41.85	536.0	546.9	550.4
32	*tert*-butyl alcohol	39.20	39.07	39.41	506.2	509.9	506.8
33	diethyl ether	35.92	37.27	36.45	466.7	460.4	460.9
34	isobutyl alcohol	42.39	42.76	41.43	547.7	534.3	541.1
35	*n*-butylamine	41.45	40.32	41.84	531.9	531.8	529.6
36	diethylamine	36.60	37.53	38.88	496.6	490.4	494.1
37	pyridine	55.60	51.18	52.08	620.0	614.6	607.9
38	cyclopentane	44.43	41.99	45.14	511.8	519.4	515.7
39	1-pentene	34.83	36.00	35.89	464.8	471.8	467.5
40	diethyl ketone	36.90	37.03	37.03	561.0	562.2	556.2
41	2-pentanone	36.46	36.27	35.52	561.1	549.0	548.7
42	ethyl propionate	33.18	34.14	34.02	546.0	549.1	551.0
43	isobutyl formate	38.30	34.93	34.26	551.3	561.3	552.3
44	*n*-propyl acetate	33.16	34.77	33.99	549.4	561.4	552.3
45	valeric acid	37.60	36.91	36.48	651.0	648.8	649.7
46	isopentane	33.37	33.27	32.91	460.4	459.6	457.5
47	neopentane	31.57	29.24	30.60	433.8	432.4	428.4
48	*n*-pentane	33.26	32.71	33.48	469.6	464.7	462.3
49	2-methyl-2-butanol	38.29	37.74	37.49	545.1	548.3	548.8
50	3-methyl-1-butanol	38.29	39.55	37.83	579.5	579.5	573.0
51	1-pentanol	38.29	40.07	39.88	586.1	591.5	581.6
52	bromobenzene	44.60	45.54	46.39	670.1	679.8	672.8
53	chlorobenzene	44.60	45.20	45.46	632.3	649.7	643.6
54	benzene	48.34	48.27	49.44	562.2	560.0	558.1
55	phenol[a,c]	60.50	NA	NA	694.3	681.2	689.4
56	aniline	52.40	45.54	48.33	699.0	679.5	688.2
57	2-methylpyridine	43.23	43.57	43.50	621.0	635.8	635.1
58	cyclohexane[b]	42.93	39.94	41.51	560.4	554.3	554.5
59	cyclohexanone	38.00	40.26	39.38	629.1	635.8	632.5
60	cyclohexane	40.22	37.41	38.60	553.5	551.6	553.3
61	1-hexene	30.99	32.37	31.36	504.0	505.7	507.4
62	methylcyclopentane	37.35	35.89	36.35	532.8	547.8	547.1
63	cyclohexanol	37.00	41.55	37.12	625.1	635.7	632.0
64	2-hexanone	32.80	31.98	31.27	587.0	580.4	576.8
65	methyl isobutyl ketone	32.30	32.35	32.11	571.4	567.0	564.2
66	ethyl *n*-butyrate	29.11	28.04	29.57	571.0	568.6	572.1
67	isobutyl acetate[b]	29.71	29.06	29.67	561.0	548.8	552.6
68	2,2-dimethylbutane	30.40	29.34	31.02	488.8	494.1	493.6
69	2,3-dimethylbutane	30.86	31.71	31.00	500.0	497.0	499.0
70	*n*-hexane[b]	29.85	29.48	30.29	507.4	500.3	503.9
71	2-methylpentane	29.71	28.98	29.81	497.5	490.1	491.1
72	3-methylpentane	30.83	30.47	31.17	504.4	504.9	506.0
73	diisopropyl ether	28.42	30.57	30.85	500.0	507.4	508.0

Table 1 (Continued)

	compound name	exptl pressure	linear predicted	CNN model predicted	exptl temp	linear predicted	CNN model predicted
74	di-*n*-propyl ether	29.88	30.45	28.50	530.6	522.5	522.5
75	diisopropylamine	31.58	30.44	32.16	523.1	532.9	532.9
76	di-*n*-propylamine	35.83	30.88	31.98	555.8	553.2	549.2
77	benzaldehyde	45.89	44.58	46.17	695.0	668.8	687.2
78	toluene	40.55	40.20	39.95	591.8	589.6	587.2
79	*m*-cresol	45.00	43.35	47.39	705.8	701.8	704.9
80	*o*-cresol	49.40	44.37	46.08	697.5	690.7	700.6
81	*m*-toluidine	41.00	39.68	39.70	709.1	699.0	706.5
82	*o*-toluidine	37.00	41.81	42.11	694.1	700.1	707.7
83	ethylcyclopentane	33.53	32.91	32.63	569.5	582.5	580.7
84	2,3-dimethylpenane	28.70	28.14	28.25	537.3	532.4	534.0
85	*n*-heptane	27.04	26.47	27.34	540.3	532.4	540.5
86	2-methylhexane	26.98	25.80	27.56	530.4	522.5	527.6
87	3-methylhexane	27.77	26.96	28.55	535.3	532.2	533.1
88	ethylbenzene	35.62	35.94	34.44	617.2	620.1	614.4
89	*m*-xylene	34.95	34.85	33.69	617.0	613.3	608.4
90	*p*-xylene	34.65	35.98	34.74	616.3	616.0	610.6
91	2,6-xylenol	42.44	40.93	41.55	701.0	715.8	709.6
92	*N*,*N*-dimethylaniline	35.80	34.46	32.13	687.1	673.9	683.0
93	*cis*-1,2-dimethylcyclohexane	29.00	32.15	30.36	606.1	609.7	604.1
94	*trans*-1,2-dimethylcyclohexane	29.00	30.74	29.73	596.1	605.3	600.0
95	*cis*-1,3-dimethylcyclohexane	29.00	28.89	28.35	591.1	595.8	592.0
96	*trans*-1,3-dimethylcyclohexane	29.00	31.82	30.11	598.0	615.7	612.6
97	*trans*-1,4-dimethylcyclohexane	29.00	30.01	29.46	590.1	598.0	593.7
98	ethylcyclohexane	30.00	29.93	29.22	609.1	610.0	606.5
99	isobutyl isobutyrate[b]	25.76	23.91	26.24	602.0	609.0	599.5
100	2,3-dimethylhexane	25.96	23.98	26.30	563.4	557.1	555.1
101	2-methyl-3-ethylpentane	26.65	24.28	26.22	567.0	561.2	558.8
102	*n*-octane[b]	24.57	23.68	24.50	568.8	560.7	571.1
103	2,2,3-trimethylpentane	26.94	27.01	27.39	563.5	559.8	558.5
104	2,2,4-trimethylpentane[b]	25.34	24.69	26.49	544.0	529.3	536.5
105	2,3,3-trimethylpentane	27.83	29.52	28.62	573.5	586.8	584.4
106	2-ethyl-1-hexanol	26.94	27.11	28.16	640.3	629.7	632.1
107	quinoline[d]	45.99	45.11	47.15	782.2	NA	NA
108	cumene	31.67	32.61	31.14	631.1	637.3	633.1
109	*o*-ethyltoluene	30.00	32.54	31.50	651.1	649.4	647.2
110	29.00	29.00	32.54	30.97	640.1	644.7	640.8
111	mesitylene	30.86	29.58	29.72	637.4	635.1	629.9
112	n-propylbenzene	31.58	31.49	30.05	638.4	640.6	635.4
113	1,2,3-trimethylbenzene[b]	34.09	34.85	34.93	664.5	654.8	661.8
114	1,2,4-trimethylbenzene	31.90	32.98	31.92	649.1	647.1	646.8
115	*n*-propylcyclohexane	27.70	27.30	27.36	639.1	631.9	628.9
116	3,3-diethylpentane	26.40	24.96	26.17	610.0	613.9	613.2
117	*n*-nonane	22.60	21.13	21.89	595.6	587.1	596.6
118	2,2,3,3-tetramethlpentane[b]	27.00	28.25	27.65	610.8	598.9	602.0
119	1,2,3,4-tetrahydronaphthalene	35.73	36.28	35.57	720.1	710.7	709.0
120	*n*-butylbenzene	28.49	27.24	27.36	660.5	663.5	660.8
121	*p*-cymene	28.00	30.22	29.20	653.1	661.0	666.5
122	isobutylbenzene[b]	30.00	29.33	28.18	650.1	651.7	648.2
123	*n*-decane	20.82	18.83	19.43	618.5	610.7	618.8
124	*n*-nonadecane	11.94	15.97	15.46	658.2	656.2	660.7
125	*n*-tetradecane	15.49	12.76	13.33	692.4	689.8	694.3
126	*n*-octadecane	12.53	13.05	12.32	745.3	748.8	742.4
127	*sec*-butyl chloride	38.49	38.46	37.26	520.6	529.9	533.8
128	methyl isopropyl ether	38.29	39.38	38.38	464.5	467.5	470.8
129	*sec*-butylamine	39.48	40.51	40.87	514.3	519.7	521.9
130	*tert*-butylamine	37.90	36.40	38.45	483.9	494.4	492.8
131	2-methyl-1-butene	33.56	35.62	34.50	465.0	472.6	467.6
132	2-methyl-2-butene	33.56	34.93	33.99	471.0	469.0	463.7
133	3-methyl-1-butene	34.70	35.46	34.28	450.4	466.8	463.7
134	methyl isopropyl ketone	38.00	38.71	38.98	553.0	554.4	550.4
135	methyl *n*-butyrate	34.28	33.90	34.04	554.5	552.8	556.7
136	piperidine[b]	45.90	43.48	44.28	594.0	586.3	579.9
137	ethyl propyl ether	33.26	33.79	32.53	500.2	493.3	494.4
138	methyl *tert*-butyl ether	33.85	34.61	35.02	497.1	496.7	500.5
139	3-methylpyridine	43.23	43.55	43.39	645.0	636.5	636.9
140	1,5-hexadiene	33.06	34.35	32.47	507.0	510.9	511.1
141	hexanenitrile[d]	28.82	34.61	32.45	622.1	NA	NA
142	3-hexanone	32.77	33.05	32.74	582.8	587.1	579.4
143	ethyl isobutyrate	30.00	30.36	30.71	553.1	572.3	563.6
144	*n*-propyl propionate	30.20	28.39	29.72	578.0	576.9	573.4
145	1-hexanol	34.64	36.40	35.25	611.3	620.6	611.4
146	2-hexanol	33.56	35.12	33.49	586.2	605.2	593.9

Table 1 (Continued)

	compound name	exptl pressure	linear predicted	CNN model predicted	exptl temp	linear predicted	CNN model predicted
147	4-methyl-2-pentanol	34.25	35.28	32.78	574.4	594.8	584.6
148	acetone[a]	46.40	50.56	45.65	508.2	471.2	484.5
149	2-propanol[a]	53.06	51.03	48.64	508.3	493.7	498.2
150	ethyl vinyl ether[a]	40.17	38.73	38.71	475.1	477.2	476.9
151	isovaleric acid[a]	38.39	37.81	37.95	634.0	635.3	635.6
152	*n*-butyl acetate[a]	30.69	29.55	29.01	579.1	565.8	576.1
153	*p*-cresol[a]	50.83	44.65	50.45	704.6	707.0	707.0
154	*n*-toluidine[a]	39.48	41.03	43.43	693.1	704.3	709.3
155	1-heptene[a]	27.93	28.73	27.38	537.3	537.4	543.9
156	methylcyclohexane[a]	34.26	32.76	31.58	572.2	578.5	578.2
157	2,2,3-trimethylbutane[a]	30.99	29.13	28.69	531.2	527.2	529.8
158	*o*-xylene[a]	36.85	37.41	37.53	630.4	624.0	618.5
159	1-octene[a]	25.17	25.19	24.15	566.6	565.4	573.6
160	*m*-ethyltoluene[a]	28.00	30.27	29.22	637.1	642.9	638.1
161	*p*-diethylbenzene[a]	27.66	30.01	28.93	658.0	671.6	675.9
162	1-decene[a]	21.40	18.91	19.80	617.0	614.6	620.5
163	propionitrile[a]	41.30	47.88	46.39	564.4	560.3	559.6
164	4-methylpyridine[a]	46.00	43.11	42.82	646.1	634.0	634.8
165	*n*-pentyl formate[a]	30.84	31.19	28.36	576.0	596.1	582.6

[a] Member of the prediction set. [b] Member of the cross-validation set. [c] Outlier of the pressure model. [d] Outlier of the temperature model.

generation of the model, was used for validation by determining its rms error and comparing it to that of the tset. If the rms error values varied greatly, the model was determined to be invalid. If any of the methods of validation failed, a different model was generated and tested.

Neural Networks. Once valid linear models were found using MLR, steps were taken to see if prediction results could be improved by the use of CNN. Typically, superior models can be found using CNNs because they implement nonlinear relationships and because they have more adjustable parameters than the linear models. Previous papers have presented the use of CNNs for QSPR studies in some detail.[17] The present work used fully connected, three-layer, feed-forward CNNs which were trained using a quasi-Newton algorithm. The data set was divided into a tset, a cvset, and a pset of 132, 15, and 18 compounds, respectively. These data were then submitted to analysis by networks of various architectures, using the same set of descriptors, to produce the best results with the fewest adjustable parameters. Since the number of input neurons is fixed at the number of descriptors being used and the single output neuron is fixed, investigations of neural network architecture were confined to seeking the best number of hidden layer neurons to use. The cvset was used to locate the optimal set of weights and biases for a given network system and also to avoid overtraining.

All computations were performed on a DEC 3000 AXP Model 500 workstation utilizing ADAPT software.[3]

RESULTS AND DISCUSSION

Critical Temperature Investigations. The first modeling step of this study was the development of a linear model to predict critical temperatures from structural information alone. Two nitrogen-containing compounds (quinoline and hexanenitrile) were determined to be statistical outliers in preliminary modeling, and they were removed from the training set. The eight descriptors comprising the best linear model are shown in Table 2. This model produced an rms error of 9.16 K and a correlation coefficient (R) of 0.993 for the training set compounds. The external prediction set had a rms error of 12.44 K. However, a large percentage of the pset error was associated with one compound, acetone, and when it was removed from the pset the rms error dropped to 9.13 K. Thus, this linear model validates quite well.

Table 2. Critical Temperature Model Developed from ADAPT Descriptors Using Multiple Linear Regression

explanation	descriptor	coefficients
the dipole moment	DPOL	13.24 ± 1.39
total charge weighted partial positive surface area	PPSA 2	-0.1025 ± 0.0139
simple path 3 connectivity index	S3P	27.69 ± 2.66
number of oxygens	NO	-40.68 ± 1.96
secondary sp^3 carbons	2SP3	4.272 ± 0.527
cube root gravitation index	GRAV 3	59.37 ± 2.51
$\sum$ ((S.A. * charge of acceptor atoms)/ number of acceptor atoms)	SCAA 2	-7.427 ± 0.192
average charge on positively charged carbons	ACPC	167.2 ± 15.9
y intercept	CONS	73.22

The linear model described above showed improved results compared to that of a previous study which used the same data set of 165 compounds with an 18 compound pset and a 147 compound tset.[2] The previous study of this data set incorporated the experimental boiling point of each compound as a descriptor. That model had an rms error of 8.36 K, R of 0.994, and was validated with an external pset rms error of 8.75 K. In addition to the model using boiling points, another model was constructed using structural information alone. That model had an rms error of 11.51 K and R of 0.989. However, the overall rms error for the pset was not reported. Neither of the original models contained any statistical outliers.

The eight descriptors in the present linear model span the range of descriptor types. Three are topological descriptors (S3P, NO, 2SP3), two are electronic descriptors (DPOL, ACPC), one is a geometric descriptor (GRAV 3), one is a charged partial surface area descriptor (PPSA 2), and one is a hydrogen-bonding descriptor (SCAA 2). The three most important descriptors of the set, by their F-values, are S3P, PPSA 2, and SCAA 2. These eight descriptors, taken as a group, successfully encode the compounds' structures.

Comparison of the two structural models proves to be valuable. The models contain two descriptors in common:

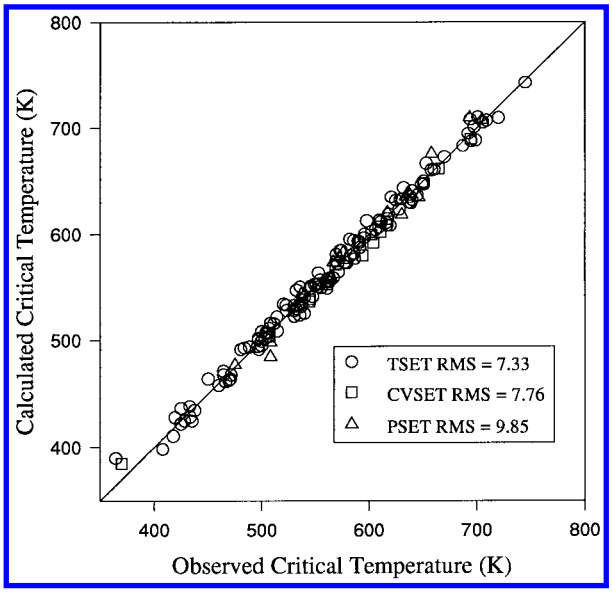

Figure 1. Critical temperature calculated vs observed plot for the training, cross-validation, and prediction sets using an eight-descriptor computational neural network model.

the number of oxygens and the average charge on positively charged carbons. This indicates the possible importance of the relationship between these descriptors and critical temperature. In addition, both models contained descriptors of some type of path connectivity and charge normalized by surface area. Regardless of the similarities, the results of the two models greatly differ with a 36% decrease in training set rms error between the new and the old model. It is believed that the best explanation for the decreased error comes from the presence of new descriptors. Three descriptors not previously available are used here: 2SP3, GRAV 3, and SCAA 2. These new descriptors evidently capture structural information in a new way that was not possible with the descriptors available to the previous study.

Having shown that a linear model could be constructed from structural descriptors alone, the next step involved using a CNN to develop a nonlinear model based on the same set of descriptors. The eight descriptors were tested with several CNN architectures, and it was determined that a 8-4-1 architecture (8 input neurons for the eight descriptors, 4 hidden neurons, and 1 output neuron for a total of 41 adjustable parameters) yielded the best results, considering the accuracy of prediction and the number of adjustable parameters. The CNN calculations were successful in improving the results. The observed rms errors were 7.33 K for the tset, 7.76 K for the cvset, and 9.85 K for the pset. However, once again, acetone contributed a large percentage of the error in the prediction set. Once acetone was removed the rms error of the pset was reduced to 8.34 K. The 7.33 K rms error for the training set corresponds to a 1.3% relative error at the mean of the critical temperature values (573.5 K). The calculated versus experimental critical temperatures of the CNN model are shown in Figure 1. The plot shows that the model is of high quality throughout the range of the critical temperature values.

Examination of the rms errors for both the linear and CNN models illustrates the improved results from the previous study. The present study, using structural information alone, was able to achieve results better than those determined previously using a physical property as a descriptor. In addition, another indication of excellent results is that the errors are very near the predicted error of 8 K determined by the reported experimental values.

Table 3. Critical Pressure Model Developed from ADAPT Descriptors Using Multiple Linear Regression

explanation	descriptor	coefficients
the charge on the most negative atom	QNEG 1	-14.57 ± 1.83
partial positive surface area	PPSA 1	-0.1244 ± 0.0059
weighted partial negative surface area	WNSA 2	0.4344 ± 0.052
the Wiener number	ALLP 5	0.03641 ± 0.0047
number of paths length 3	N3P 5	0.7220 ± 0.087
second major moment (width)	GEOM 2	-8.536 ± 1.03
first major moment of inertia/ second major moment of inertia	MOMI 4	10.76 ± 1.26
sum of surface area of acceptor atoms/number of acceptor atoms	SAAA 2	0.1201 ± 0.021
y intercept	CONS	53.46

Critical Pressure Investigations. The next step was to build a model to predict the critical pressure. This together with the critical temperature would allow the critical point to be determined. The same data set of compounds was employed with the same training, cross-validation, and prediction sets.

As before, an eight-descriptor model proved to be the most effective for prediction. However, phenol proved to be an outlier in this phase of the study. The eight descriptors comprising the final model are shown in Table 3. This linear model produced a rms error of 2.07 atm and R of 0.964. The pset rms error was 2.80 atm. The majority of the model's error was found in the prediction of compounds with high critical pressure values.

The eight descriptors chosen for modeling of critical pressure span the range of descriptor types available for use. Two of the descriptors are topological (ALLP 5, N3P 5), 1 is electronic (QNEG 1), 2 are geometric (GEOM 2, MOMI 4), 2 are charged partial surface area descriptors (PPSA 1, WNSA 2), and 1 is a hydrogen-bonding descriptor (SAAA 2). Only one descriptor is found in both the critical temperature and critical pressure models, the hydrogen-bonding descriptor SAAA 2. The three most important descriptors of the set, by their F-values, are PPSA 1, QNEG 1, and ALLP 5. This set of eight descriptors, taken as a group, successfully encodes the structures of the compounds.

Once the linear model was developed, a CNN was used to improve the overall results. The best CNN architecture was determined to be 8-5-1 (51 adjustable parameters). The best CNN model found gave a tset rms error of 1.51 atm, a cvset rms error of 1.35, and an external pset error of 2.39. The 1.51 atm rms error for the training set corresponds to a 4.4% relative error at the mean of the critical pressure values (33.7 atm). Although the pset error still remains well above that of the tset, it can be seen in Figure 2 that the general predictive capability of the CNN model is quite good. The predictions in the higher pressure ranges remain good. This plot shows more scatter than that for the critical temperatures; for this set of compounds, the critical pressures are more difficult to model.

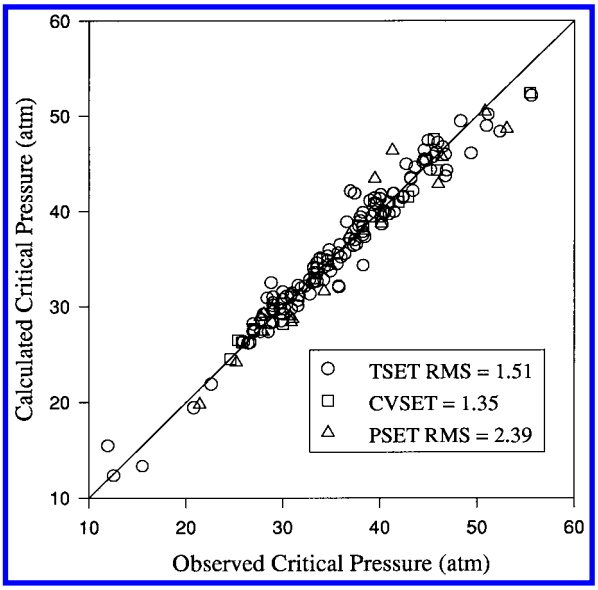

Figure 2. Critical pressure calculated vs observed plot for the training, cross-validation, and prediction sets using an eight-descriptor computational neural network model.

CONCLUSIONS

This study shows that it is possible to develop good quantitative structure−property relationships for the prediction of critical temperature and pressure of structurally diverse organic compounds using structural information alone. The models provide some insight into what structural features are related to these specific physical properties. Linear models chosen using the genetic algorithm and simulated annealing feature selection methods were able to predict these two critical properties. Additionally, using CNNs to build nonlinear models based on these same sets of descriptors produced even better models with good predictive ability.

ACKNOWLEDGMENT

We thank Matthew Wessel, Jon Sutter, Brooke Mitchell, Stephen Johnson, and Heidi Engelhardt for continued assistance throughout this project. This work was supported in part by the Design Institute for Physical Property Data (DIPPR) Project 931: Data Prediction Methods. Support was also received from the National Science Foundation REU Program.

REFERENCES AND NOTES

(1) Bea, H.-K.; Lee, S.-Y. A Method for the Prediction of Critical Temperature and Pressure of Pure Fluids. *Fluid Phase Equilibria* **1991**, *66*, 225−232.

(2) 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.

(3) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979; p 83.

(4) Johnson, S. R.; Jurs, P. C. Acute Mammalian Structure for a Diverse Set of Substituted Anilines Using Regression Analysis and Computational Neural Networks. In *Computer-Assisted Lead Finding and Optimization*; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; Verlag Helvetica Chimica Acta: Basel, 1997.

(5) Mitchell, B. E.; Jurs, P. C. Prediction of Autoignition Temperatures of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 538−547.

(6) Engelhardt, H. L.; Jurs, P. C. Prediction of Supercritical Carbon Dioxide Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 478−484.

(7) Wessel, M. D. *Computer-Assisted Development of Quantitative Structure−Property Relationships and Design of Feature Selection Routines.* Ph.D. Dissertation, Pennsylvania State University, University Park, PA, 1996.

(8) Sutter, J. M.; Jurs, P. C. Selection of Molecular Structure Descriptors for Quantitative Structure−Activity Relationships. In *Adaption of Simulated Annealing to Chemical Problems*; Kalivas, J. H., Ed.; Elsevier Science Publishers B. V.: Amsterdam, 1995.

(9) Design Institute for Physical Property Data (DIPPR). *Physical and Thermodynamic Properties of Pure Chemicals: Data Compilation*; Daubert, T. E., Danner R. P., Eds.; Hemisphere Publishing: New York, 1989: Vol. 1−4.

(10) Hypercube, Inc., Waterloo, ON.

(11) Stewart, J. P. P. MOPAC 6.0, *Quantum Chemistry Program Exchange*; Indiana University, Bloomington, IN, Program 455.

(12) Stewart, J. P. P. MOPAC, A Seimemperical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.

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

(14) Russell, C. J.; Dixon, S. L.; Jurs, P. C. Computer Assisted Study of the Relationship Between Molecular Structure and Henry's Law Constant. *Anal. Chem.* **1992**, *64*, 1350.

(15) Belsey, D. A.; Kuh, E.; Welsch R. E. *Regression Diagnostics*; Wiley: New York, 1980.

(16) Stanton D. T.; Jurs, P. C. Computer Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans, and Thiophenes. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 301.

(17) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure-roperty Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxical. Chem.* **1994**, *13*, 841.

CI9800054