Prediction of Glass Transition Temperatures from Monomer and Repeat Unit Structure Using Computational Neural Networks

Brian E. Mattioni and Peter C. Jurs*

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

Received June 24, 2001

Quantitative structure−property relationships (QSPR) are developed to correlate glass transition temperatures and chemical structure. Both monomer and repeat unit structures are used to build several QSPR models for Parts 1 and 2 of this study, respectively. Models are developed using numerical descriptors, which encode important information about chemical structure (topological, electronic, and geometric). Multiple linear regression analysis (MLRA) and computational neural networks (CNNs) are used to generate the models after descriptor generation. Optimization routines (simulated annealing and genetic algorithm) are utilized to find information-rich subsets of descriptors for prediction. A 10-descriptor CNN model was found to be optimal in predicting T_g values using the monomer structure (Part 1) for 165 polymers. A committee of 10 CNNs produced a training set rms error of 10.1K ($r^2 = 0.98$) and a prediction set rms error of 21.7K ($r^2 = 0.92$). An 11-descriptor CNN model was developed for 251 polymers using the repeat unit structure (Part 2). A committee of CNNs produced a training set rms error of 21.1K ($r^2 = 0.96$) and a prediction set rms error of 21.9K ($r^2 = 0.96$).

INTRODUCTION

The glass transition temperature (T_g) of amorphous polymers is the most important and widely studied polymeric property.[1−6] Below the T_g, molecules can oscillate and vibrate around a fixed position creating a certain amount of free volume, which is dependent upon the temperature of the system. Higher temperatures produce more free volume due to increased oscillations and vibrations of the molecules. The T_g occurs at the point where there is sufficient free volume to allow molecules in the polymer backbone to move relative to one another.[6] At this point, the once rigid backbone relaxes causing a transition from a solid polymer material to a quasi-liquid state.

Many other polymer properties such as heat capacity, coefficient of thermal expansion, and viscosity, are affected by this transition. It would be advantageous to produce robust quantitative structure−property relationship (QSPR) models that could predict T_g values for new polymeric materials. QSPR models use descriptors, which numerically encode a molecule's structure, in an attempt to find relationships between structure and a property of interest. Robust, accurate models have been difficult to develop because of the many other factors that affect the value for the T_g that cannot be captured by these simple numerical descriptors. A few examples include rate of measurement, average molecular weight, thermal history, and pressure.[2]

Materials research is constantly being pushed to the extreme so new materials can be developed with enhanced physical properties to be used in various applications. Testing these materials through experimentation can be a costly and time-consuming process. QSPR models can aid in experimental design by excluding some polymers as possible candidates for use in an application of interest. Such models have the ability to survey a list of possible candidates and exclude ones that do not fall into the desired property range for the application. By concentrating on only those compounds that fall into a desired property range, a substantial savings in time and money can be achieved. In particular, no time is wasted on synthesis and testing of new materials that are deemed inappropriate by QSPR models.

Numerous models for predicting the T_g of amorphous polymeric materials have been reported. Many empirical/semiempirical[5,7−10] and group contribution[11,12] methods have been used to build models for prediction of the T_g. Generally, these studies involved the use of small data sets, limiting the range of T_g values. In some cases, subsets were taken from the main data set to build models that only pertained to a particular class of polymers (i.e., polyacrylates, polymethacrylates). Waegell and co-workers approached modeling by using an Energy, Volume, Mass (EVM) QSPR model.[13,14] For linear and branched aliphatic acrylate and methacrylate polymers, the standard deviation from linear regression was 12K with an r^2 value of 0.96. This model also allowed for the calculation of polymer T_g values not used in the original multiple linear regression with an average absolute error of 10%.

Katritzky and co-workers used a different approach when working with a set of 88 compounds.[15] Structural descriptors were calculated for the middle repeat unit of a three-repeat unit string. Instead of using T_g values as the dependent variable, molar T_g ($T_g/M_{\text{repeat unit}}$, where $M_{\text{repeat unit}}$ represents the molar mass of the repeat unit) values produced better correlations for the data set. A five-parameter model was found that produced a correlation coefficient $r^2 = 0.946$. Additional experiments were performed to prove the gen-

* Corresponding author phone: (814)865-3739; fax: (814)865-3314; e-mail: pcj@psu.edu.

10.1021/ci010062o CCC: $22.00 © 2002 American Chemical Society
Published on Web 02/16/2002

eralization capability of the model. The external validation procedure produced average correlation coefficients of 0.935 for the test set.

Joyce and co-workers used computational neural networks to build models for T_g prediction based on monomer structure of the polymers.[16] Monomer structures for 360 polymers in a training set were represented in SMILES format, and a series of indicator variable descriptors were calculated based on the SMILES representation. Neural networks used in the study "... were able to predict the T_g values for a testing set of polymers with a wide range of structures with an rms error of ca.. 35K...".[16]

Another neural network approach was used by Sumpter and Noid using the repeat unit structure as representative of the polymeric material.[17,18] A hybrid model was developed which combined important attributes of Bicerano,[2] Porter,[19] and Askadskii[7,8] models. Using the PropNet[17] technology, models were built for a data set of 320 compounds (T_g values ranging from 50 K to 700 K), which could predict T_g values with a standard deviation of approximately 8K ($r = 0.992$).

Perhaps the most widely referenced model is one produced by Bicerano.[2] Bicerano used a data set that consisted of 320 polymer compounds. A model was built that combined a weighted sum of structural parameters along with the solubility parameter of each polymer. A linear regression procedure was used to produce a model with a standard deviation of 24.65K and a correlation coefficient of 0.9749. However, no external data set compounds were withheld to validate this model.

The goal of the present study was to produce robust QSPR models that could predict T_g values for a diverse set of polymers. This was attempted using two approaches: (1) using the monomer structure as representative of the polymeric material and (2) using the repeat unit as representative of the polymeric material. Descriptors of all types were used for the first part, but only topological descriptors were used for the second part of the study. Models produced with only topological descriptors have advantages over some other models discussed above due to the relative ease in calculating descriptors, thus lowering computational cost and time. This study also utilizes an external prediction set which validates models based on their ability to predict properties of materials that were not used in training. This essential aspect of QSPR is lacking in most studies discussed above.

EXPERIMENTAL SECTION

Part 1: Monomer Structure. The first part of this study involved the prediction of T_g based on numerical descriptors calculated from the structure of the monomer compound used in the polymerization. For example, the structure used to calculate descriptors for polystyrene was the styrene molecule. The 165 polymer compounds used in Part 1 of this study came from Bicerano.[2] The compounds and their respective T_g values appear in Table 1 as compounds **1**−**165**. The size of the monomers ranged from 28 to 254 atomic mass units (amu). The reported T_g values ranged from 188 K to 475 K. A 17-member prediction set was randomly removed from the data set leaving a 148-member training set. Care was taken to ensure that the prediction set dependent variables covered the dependent variable range of the entire data set.

Part 2: Repeat Unit Structure. The second part of this study involved the prediction of T_g based on the structure of the polymeric repeat unit. One repeat unit was used to represent the corresponding polymer. The structures were end-capped with two hydrogens. Only topological descriptors were calculated for the structures. This was done because geometric and electronic descriptors would not have encoded structural information correctly based on how the structures are drawn. For example, if an oxygen molecule was located at the terminal end of the repeat unit, the hydrogen end-cap would make an alcohol group, which is not present in the chains of the polymer as a whole. It was believed that electronic and geometric descriptors would falsely describe the entire polymeric material. By using the repeat unit, 87 more compounds were added to the data set in Part 1 of the study. The additional compounds are listed as **166**−**252** in Table 1. The only compound that was excluded from the data set in Part 2 was polypentadiene (**18**) because the exact positions of the double bonds in the structure were not known. This left a 251-member data set. This data set produced a larger molecular size range (28−679 amu) than Part 1. Furthermore, the range (188−673 K) of the dependent variable is now larger when using the extended data set. From the data set, 25 compounds were removed to generate a prediction set leaving a 226-member training set. Once again, the dependent variables of the prediction set covered the entire data set range.

Model Development. Models were formed using descriptors that were generated from the structures of the compounds. Descriptors are numerical values that encode various aspects of molecular structure. Models consist of information-rich subsets of descriptors which are examined for statistical soundness and predictive ability. The ultimate goal was to find models that minimized the root-mean-square (rms) error for the polymer training set. The best models were then validated with the prediction set compounds, which were set aside during the model development stages.

Computational neural networks (CNNs) were used to build nonlinear models. For Part 1 of the study, a 17-member cross-validation set was randomly removed from the training set leaving a new 131-member training set along with the original 17-member prediction set. The same was done for Part 2 except the cross-validation set contained 25 members randomly removed from the training set leaving only 201 members. The same 25 member prediction set was used. The cross-validation sets used in the two parts were employed to prevent over-training of the CNN, which will be discussed in more detail later.

All computations were performed on a DEC 3000 AXP Model 500 workstation using the Unix operating system. Quantitative structure property relationships (QSPRs) were developed using the **A**utomated **D**ata **A**nalysis and **P**attern recognition **T**oolkit (ADAPT) software,[20,21] simulated annealing,[22] genetic algorithm,[23] and CNN[24] routines that were developed in our laboratory at Penn State. The development of QSAR models involves four major steps: (1) structure entry and optimization, (2) descriptor generation with objective feature selection, (3) model formation and validation by linear means, and (4) model formation and validation by nonlinear means.

Structure Entry and Optimization. Compounds were sketched on a PC using HyperChem (HyperCube, Inc.

Table 1. Polymers Used in This Study along with Their Corresponding T_g

compd no.	polymer name	T_g (K)	compd no.	polymer name	T_g (K)
1	poly(vinyl chloride)	348	81	poly(*tert*-butyl acrylate)	315
2	poly(2-*n*-butyl-1,4-butadiene)	192	82	poly(2,3,3,3-tetrafluoropropylene)	315
3	poly(vinyl *n*-octyl ether)	194	83	poly(10-aminodecanoic acid)	316
4	poly(3-hexoxypropylene oxide)	188	84	poly(3,3-dimethylbutyl methacrylate)	318
5	poly(3-butoxypropylene oxide)	194	85	poly(N-butyl acrylamide)	319
6	poly(ethylene)	195	86	poly(vinyl trifluoroacetate)	319
7	poly(2-*n*-propyl-1,4-butadiene)	196	87	poly(*p*-*n*-butoxy styrene)	320
8	poly(vinyl *n*-decyl ether)	197	88	poly(isobutyl methacrylate)	321
9	poly(2-ethyl-1,4-butadiene)	197	89	poly(3-methyl-1-butene)	323
10	poly(isobutylene)	199	90	poly(9-aminononanoic acid)	324
11	poly(isoprene)	203	91	poly(8-aminocaprylic acid)	324
12	poly(propylene oxide)	206	92	poly(hexafluoropropylene)	425
13	poly(vinyl *n*-pentyl ether)	207	93	poly(ethyl methacrylate)	324
14	poly(vinyl 2-ethylhexyl ether)	207	94	poly(isopropyl methacrylate)	327
15	poly(*n*-octyl acrylate)	208	95	poly(*n*-butyl α-chloroacrylate)	330
16	poly(vinyl *n*-hexyl ether)	209	96	poly(7-aminoheptanoic acid)	330
17	poly(3-methoxypropylene oxide)	211	97	poly(*sec*-butyl methacrylate)	330
18	poly(pentadiene)	213	98	poly(*p*-isopentoxy styrene)	330
19	poly(*n*-heptyl acrylate)	213	99	poly(heptafluoropropyl ethylene)	331
20	poly(ϵ-caprolactone)	213	100	poly(3-cyclopentyl-1-propene)	333
21	poly(*n*-nonyl acrylate)	215	101	poly(3-phenyl-1-propene)	333
22	poly(*n*-hexyl acrylate)	216	102	poly(ϵ-caprolactam)	335
23	poly(dodecyl methacrylate)	218	103	poly(*p*-*n*-propoxy styrene)	343
24	poly(*n*-butyl acrylate)	219	104	poly(*n*-propyl α-chloroacrylate)	344
25	poly(1-heptene)	220	105	poly(N-vinyl carbazole)	423
26	poly(vinyl *n*-butyl ether)	221	106	poly(*sec*-butyl α-chloroacrylate)	347
27	poly(2-isopropyl-1,4-butadiene)	221	107	poly(3-cyclohexyl-1-propene)	348
28	poly(1-hexene)	223	108	poly(vinyl cyclopentane)	348
29	poly(1-pentene)	223	109	poly(2-hydroxypropyl methacrylate)	349
30	poly(chloroprene)	225	110	poly(*p*-methoxymethyl styrene)	350
31	poly(propylene sulfide)	226	111	poly(cyclohexyl methacrylate)	356
32	poly(1-butene)	228	112	poly(vinyl alcohol)	358
33	poly(2-octyl acrylate)	228	113	poly(*p*-*sec*-butyl styrene)	359
34	poly(*n*-propyl acrylate)	229	114	poly(*p*-ethoxy styrene)	359
35	poly(propylene)	233	115	poly(2-hydroxyethyl methacrylate)	359
36	poly(vinylidene fluoride)	233	116	poly(*p*-isopropyl styrene)	360
37	poly(2-heptyl acrylate)	235	117	poly(2-methyl-5-*tert*-butyl styrene)	360
38	poly(6-methyl-1-heptene)	239	118	poly(*p*-methoxy styrene)	362
39	poly(2-bromo-1,4-butadiene)	241	119	poly(isopropyl α-chloroacrylate)	363
40	poly(isobutyl acrylate)	249	120	poly(4-methoxy-2-methyl styrene)	363
41	poly(vinyl isobutyl ether)	251	121	poly(vinyl cyclohexane)	363
42	poly(ethyl acrylate)	251	122	poly(*m*-chloro styrene)	363
43	poly(*n*-octyl methacrylate)	253	123	poly(2-chloroethyl methacrylate)	365
44	poly(vinyl *sec*-butyl ether)	253	124	poly(ethyl α-chloroacrylate)	366
45	poly(*sec*-butyl acrylate)	253	125	poly(*m*-methyl styrene)	370
46	poly(vinyl ethyl ether)	254	126	poly(chlorotrifluoroethylene)	373
47	poly(vinylidene chloride)	256	127	poly(styrene)	373
48	poly(3-pentyl acrylate)	257	128	poly(*p*-methyl styrene)	374
49	poly(5-methyl-1-hexene)	259	129	poly(2,5-difluoro styrene)	374
50	poly(*n*-hexyl methacrylate)	268	130	poly(*o*-ethyl styrene)	376
51	poly(vinyl isopropyl ether)	270	131	poly(3,5-dimethyl styrene)	377
52	poly(α-vinyl naphthalene)	432	132	poly(*o*-vinyl pyridine)	377
53	poly(*β*-vinyl naphthalene)	424	133	poly(methyl methacrylate)	378
54	poly(3-phenoxypropylene oxide)	315	134	poly(acrylonitrile)	378
55	poly(α, *β*, *β*-trifluoro styrene)	475	135	poly(*o*-fluoro styrene)	378
56	poly(methyl α-cyano acrylate)	433	136	poly(2,3,4,5,6-pentafluoro styrene)	378
57	poly(*o*-hydroxymethyl styrene)	433	137	poly(acrylic acid)	379
58	poly(vinyl methyl sulfide)	272	138	poly(*p*-fluoro styrene)	379
59	poly(vinyl butyrate)	278	139	poly(*tert*-butyl methacrylate)	380
60	poly(*p*-*n*-hexoxymethyl styrene)	278	140	poly(3,4-dimethyl styrene)	384
61	poly(*p*-*n*-butyl styrene)	279	141	poly(2-fluoro-5-methyl styrene)	384
62	poly(methyl acrylate)	281	142	poly(2,4-dimethyl styrene)	385
63	poly(vinyl propionate)	283	143	poly(*p*-methoxycarbonyl styrene)	386
64	poly(2-ethylbutyl methacrylate)	284	144	poly(3-methyl-4-chloro styrene)	387
65	poly(*o*-*n*-octoxy styrene)	286	145	poly(cyclohexyl α-chloroacrylate)	387
66	poly(2-*tert*-butyl-1,4-butadiene)	293	146	poly(*p*-chloro styrene)	389
67	poly(*n*-butyl methacrylate)	293	147	poly(*o*-chloro styrene)	392
68	poly(2-methoxyethyl methacrylate)	293	148	poly(2,5-dichloro styrene)	393
69	poly(*p*-*n*-propoxymethyl styrene)	295	149	poly(phenyl methacrylate)	393
70	poly(3,3,3-trifluoropropylene)	300	150	poly(methacrylonitrile)	393
71	poly(vinyl acetate)	301	151	poly(α-*p*-dimethyl styrene)	394
72	poly(4-methyl-1-pentene)	302	152	poly(3-fluoro-4-chloro styrene)	395
73	poly(vinyl formate)	304	153	poly(*m*-hydroxymethyl styrene)	398
74	poly(vinyl chloroacetate)	304	154	poly(3,4-dichlorostyrene)	401
75	poly(neopentyl methacrylate)	306	155	poly(*p*-*tert*-butyl styrene)	402
76	poly(*n*-propyl methacrylate)	308	156	poly(2,4-dichloro styrene)	406
77	poly(12-aminododecanoic acid)	310	157	poly(*o*-methyl styrene)	409
78	poly(4-cyclohexyl-1-butene)	313	158	poly(α-methyl styrene)	409
79	poly(pentafluoroethyl ethylene)	314	159	poly(*p*-phenyl styrene)	411
80	poly(11-aminoundecanoic acid)	315	160	poly(*p*-hydroxymethyl styrene)	413

Table 1 (Continued)

compd no.	polymer name	T_g (K)	compd no.	polymer name	T_g (K)
161	poly(*p*-vinyl pyridine)	415	208	poly(1,1-cyclopentane bis[4-phenyl] carbonate)	440
162	poly(2,5-dimethyl styrene)	416	209	poly(1,1-cyclohexane bis[4-phenyl] carbonate)	444
163	poly(*p*-bromo styrene)	417	210	poly(2,2-hexafluoropropane bis[4-phenyl] carbonate)	449
164	poly(2-methyl-4-chloro styrene)	418	211	poly(1,1-[1-phenyl ethane] bis[4-phenyl] carbonate)	449
165	poly(n-vinyl pyrrolidone)	418	212	poly(2,2-[1,3-dichloro-1,1,3,3-tetrafluoro propane]-bis[4-phenyl carbonate)	457
166	poly(oxytetramethylene)	190	213	poly(perfluoro styrene)	467
167	poly(oxytrimethylene)	195	214	poly(2,2-propane bis[4-[2,6-dimethylphenyl]]-carbonate	473
168	poly(oxyoctamethylene)	203	215	poly(oxy[2,6-dimethyl-1,4-phenylene])	482
169	poly(oxyhexamethylene)	204	216	polyetherimide 2	482
170	poly(tetramethylene adipate)	205	217	poly(oxy[2,6-diphenyl-1,4-phenylene])	493
171	poly(oxyethylene)	206	218	polyetherimide 3	500
172	poly(decamethylene adipate)	217	219	polycarbonate 2	505
173	poly(oxymethylene)	218	220	polyetherimide 4	512
174	poly(ethylene azelate)	228	221	polycarbonate 3	520
175	poly(ethylene adipate)	233	222	polyetherimide 5	520
176	poly(ethylene sebacate)	243	223	polyquinoline 1	541
177	perfluoropolymer 2	255	224	polyquinoline 2	546
178	perfluoropolymer 1	260	225	polyquinoline 9	573
179	poly(1,2-butadiene)	269	226	polyphenolphthalein 2	580
180	poly(ethylene succinate)	272	227	polyphenolphthalein 3	583
181	poly(hexamethylene sebacamide)	313	228	polyphenolphthalein 1	593
182	poly(decamethylene sebacamide)	319	229	polyimide 11	606
183	poly(vinyl butyral)	324	230	polyetherimide 12	615
184	poly(ethyl-*p*-xylylene)	298	231	polyquinoline 10	618
185	poly(methyl-*p*-xylylene)	328	232	polyphenolphthalein 4	658
186	poly(hexamethylene adipamide)	330	233	polytricyclic 3	668
187	poly(*p*-xylylene)	333	234	polytricyclic 2	668
188	poly(ethylene terephthalate)	345	235	polyphenolphthalein 5	673
189	poly(chloro-*p*-xylylene)	353	236	poly(ethylene isophthalate)	324
190	poly(bromo-*p*-xylylene)	353	237	bisphenol a polycarbonate	423
191	poly(cyano-*p*-xylylene)	363	238	poly(oxycarbonyl-3-methylpentamethylene)	220
192	poly(α,α,α′,α′-tetrafluoro-*p*-xylylene)	363	239	poly(oxycarbonyl-1,5-dimethylpentamethylene)	240
193	poly(oxy-2,2-dichloromethyltrimethylene)	265	240	poly(ethylene-1,4-naphthalenedicarboxylate)	337
194	poly(2,5-dimethyl-*p*-xylylene)	373	241	poly(ethylene-1,5-naphthalenedicarboxylate)	344
195	poly(vinyl formal)	378	242	poly(ethylene oxybenzoate)	355
196	perfluoropolymer 3	390	243	poly(thio [*p*-phenylene])	360
197	polyetherimide 1	401	244	poly(ethylene-2,6-naphthalenedicarboxylate)	397
198	poly(1,1-dichloroethylene bis[4-phenyl] carbonate)	430	245	poly(hexamethylene isophthalamide)	403
199	poly(thio bis[4-phenyl] carbonate)	388	246	poly(*m*-phenylene isophthalate)	411
200	poly(1,1-ethane bis[4-phenyl] carbonate)	403	247	poly(*m*-phenylene isophthalamide)	545
201	poly(2,2-butane bis[4-phenyl] carbonate)	407	248	poly(*p*-hydroxybenzoate)	420
202	poly(2,2-pentane bis[4-phenyl] carbonate)	410	249	phenoxy resin	373
203	poly(methane bis[4-phenyl] carbonate)	420	250	resin F	384
204	poly(1,1-butane bis[4-phenyl] carbonate)	396	251	ultem	493
205	poly(oxy-1,4-phenylene-oxy-1,4-phenylene-carbonyl-1,4-phenylene)	419	252	torlon	550
206	poly(4,4-heptane bis[4-phenyl] carbonate)	421			
207	poly(1,1-[2-methyl propane] bis[4-phenyl] carbonate)	422			

Waterloo, ON). HyperChem provided reasonable starting geometries for all structures. The structures were then transferred to a Unix workstation where they were further optimized by the semiempirical molecular orbital package MOPAC.[25] The PM3 Hamiltonian[26] was used to ensure that low energy conformations were obtained for each structure.

Descriptor Generation and Objective Feature Selection. Topological, geometric, and electronic features of each molecule were calculated by the descriptor generation routines. Topological descriptors provide information about molecular shape[27] and connectivity.[28,29] Examples of these include the electrotopological state indices[30] and molecular distance edge information.[31] Geometric descriptors encode features such as shapes,[32] shadow projections,[33] and solvent accessible surface areas.[34] Electronic descriptors describe the electronic environment that exists within the molecule such as partial charges on each atom.[35] There are also several hybrid descriptors, which combine several aspects of molecules. One example is the charged partial surface area (CPSA) descriptors,[36] which combine electronic and geometric data about the molecules. More than 200 descriptors were calculated for Part 1 of this study, and 128 topological descriptors were calculated for Part 2.

Objective feature selection is done to remove descriptors than contain identical information or that are highly correlated with other descriptors. Descriptors that contain identical information for over 90% of the training set compounds are removed. Pairwise correlations are examined to remove descriptors that are highly correlated with other descriptors. If two descriptors are highly correlated, one is randomly removed from the descriptor pool. The reduction of the descriptor pool is also done to ensure that the ratio of descriptors to training set observations does not exceed 0.6, thereby reducing the risk of chance correlations during model development.[37] The descriptor pools were reduced to 62 and 35 members, respectively, for Parts 1 and 2.

Linear Feature Selection and Linear Modeling. Multiple linear regression analysis (MLRA)[38] accompanied by a simulated annealing[22] optimization algorithm was used to build linear models, termed Type I models. Subsets of information-rich descriptors were examined to see which

subsets minimize the root-mean-square (rms) error given by the following equation:

$$\text{rms} = \sqrt{\frac{\sum (f_i - x_i)^2}{N}}$$

In this expression, f_i is the predicted value for the ith compound, x_i is the observed value for the ith compound, and N represents the total number of compounds in the data set. Subset size was increased sequentially until addition of another descriptor did not significantly improve the training set rms error. The smallest subset of descriptors that did not compromise rms error was identified as optimal.

After model formation, statistical tests were performed using the best subset of descriptors to find any outliers that existed in the data set. Furthermore, a variance inflation factor (VIF) was calculated to see if multicollinearities existed between the descriptors in the model. A VIF (VIF = $1/(1-r^2)$) was calculated for each descriptor by regressing it against all other descriptors in the model. Models were not accepted if they contained descriptors with VIFs over a value of 10. This ensured that the squared multicollinearity coefficient for each descriptor in the model did not exceed 0.90. Finally, the model was validated using the external prediction set.

Linear Feature Selection and Nonlinear Modeling. The best subset of descriptors found in Type I modeling was then fed to a three-layer, fully connected, feed-forward CNN. These models are termed Type II models. The goal of neural networks is to take a set of input values and perform nonlinear transformations on the data. If a nonlinear relationship exists between the T_g and molecular features, then CNNs should be able to identify this relationship.

Each layer in the CNN consists of neurons connected by links having adjustable weights. The input layer neurons accept the values for each descriptor in the model. The number of CNN hidden layer neurons is increased sequentially until no improvement is seen for that model. A hidden layer neuron restriction, though, keeps the ratio of the number of compounds in the training set to adjustable parameters greater than two. This is done to reduce the possibility of chance correlation predictions by the neural networks.[39] The single output neuron represents the output value to be compared to the experimental value.

The cross-validation set is used to monitor the CNN during training. Periodically predicting the rms error of the cross-validation set during training allows the progress of the training to be measured. When the cross-validation set error reaches a minimum, training is stopped. This ensures that the model is able to generalize and not simply memorize structural idiosyncracies of the training set. The weights and biases that produce the lowest cross-validation set rms error over all iterations are considered optimal and passed to the CNN for final training. Network training was done with the BFGS (Broyden-Fletcher-Goldfarb-Shanno) quasi-Newton optimization algorithm.[40] Validation of the model was performed using the external prediction set.

Nonlinear Feature Selection and Nonlinear Modeling. Type III models perform nonlinear feature selection with the aid of simulated annealing[22] or a genetic algorithm[23] using a computational neural network fitness evaluator. Once optimal subsets were found, models were developed in an analogous fashion to Type II modeling. Final validation was performed using the external prediction set.

Table 2. Descriptors Used in Linear Type I Model for Polymers Represented by Their Monomers

descriptor[a]	type	coefficient	error	range
V6CH-18	topological	1781.0	138.0	0−0.136
N6PC-19	topological	−1.639	0.240	0−183
MOLC-8	topological	46.71	11.28	0−1.69
MOLC-9	topological	57.39	6.62	1.66−4.97
NSB-12	topological	−23.23	1.53	0−15
1SP2−1	topological	−68.15	6.76	0−2
MDE−44	topological	16.57	3.51	0−11.52
GRAV-4	geometric	4.93e-02	9.70e-03	83−2921
QSUM-1	electronic	50.78	6.60	0.57−4.28
RPCS-1	hybrid	6.10	1.02	3e-14−19.04
constant		126.1	19.27	

[a] V6CH-18, atom valence-corrected sixth-order chain values;[41] N6PC-19, number of sixth-order path clusters;[41] MOLC-8, heteroatom and aromatic ring corrected fourth-order path cluster;[45] MOLC-9, heteroatom and aromatic ring corrected average distance sum connectivity;[45] NSB-12, number of single bonds; 1SP2-1, number of sp^2 hybridized carbons bound to 1 other carbon; MDE-44, molecular distance edge between quaternary carbons;[31] GRAV-4, gravitational index using all pairs of atoms ($\Sigma(m_1*m_2/r^2)$);[42] QSUM-1, sum of absolute charges;[35] RPCS-1, relative positive charged surface area (SA_{MPOS}/RPCG where RPCG = (charge of most positive atom)/(sum total positive charge)).[36]

RESULTS AND DISCUSSION

Part 1. Subsets of 3−10 descriptors were examined for Type I modeling. Only subsets of descriptors with absolute T-values greater than four were considered. This ensured that the standard error of a coefficient did not exceed 25% of the coefficient value. The best model found contained 10 descriptors. Pairwise correlations for the 10 descriptors in the model ranged from 0.517 to 0.905 with an average of 0.773. Table 2 lists the descriptors in the best Type I model. The training set rms error was 25.87K ($r^2 = 0.869$). The prediction set rms error was 26.58K ($r^2 = 0.870$), demonstrating the ability of this model to generalize.

Several topological descriptors appeared in the Type I model. The descriptors V6CH-18, N6PC-19, MOLC-8, and MOLC-9 encode information about molecular size and degree of branching.[41] Molecular size and substituent branching of the polymer backbone influence the amount of free volume available to the chains of the polymer. The descriptor NSB-12 counts the number of single bonds in the molecule, which may indirectly describe the degree of flexibility of the polymer. The descriptor 1SP2-1 counts the number of sp^2 hybridized carbons attached to only one other carbon. This molecular configuration occurs at the double bond moiety of the initiator attack site to start propagation of the chains. The molecular distance edge descriptor,[31] MDE-44, encodes the distance between quaternary carbons.

The geometric descriptor, GRAV-4,[42] is the gravitational index over all atoms in a molecule. This descriptor is intended to encode information about the bulk size properties of a molecule. Molecular size can affect the T_g due to steric considerations within the polymer chains. The electronic environment of the molecule can be estimated by the electronic descriptor QSUM-1,[35] which is the summation of the absolute values of atomic charges. The electronic environment in polymers becomes important when considering the intermolecular cohesion forces that occur between

Table 3. Descriptors Used in Nonlinear Type III Model for Polymers Represented by Their Monomers

descriptor[a]	type	range
V5C-10	topological	0−0.306
V6CH-18	topological	0−0.136
MOLC-8	topological	0−1.69
NSB-12	topological	0−15
1SP2−1	topological	0−2
ELOW-1	topological	0−6.6
SHDW-4	geometric	0.428−0.585
DPOL-1	electronic	0.23e-03−4.98
CARB-1	electronic	−0.74e-02−0.32
DPSA-3	hybrid	17.9−72.1

[a] V6CH-18, MOLC-8, NSB-12, 1SP2-1, see Table 2 caption; V5C-10, atom valence-corrected fifth-order cluster;[41] ELOW-1, through space distance between EMIN and EMAX;[30] SHDW-4, standardized area projected onto the XY-plane;[33] DPOL-1, dipole moment; CARB-1, average charge on carbonyl carbons; DPSA-3, difference in charged weighted partial surface areas $[\Sigma(+SA_i)(Q^+_i) - \Sigma(-SA_i)(Q^-_i)]$ where $(+SA_i)$ and $(-SA_i)$ are the surface area contributions of the *i*th positive and or negative atom in the molecule, while Q^+_i and Q^-_i are the partial atomic charges for the *i*th positive and negative atoms.[36]

different segments of the polymer chain. The hybrid descriptor RPCS-1[36] measures the relative positively charged surface area of a molecule, which can also affect T_g due to intermolecular interactions.

The descriptors from the best Type I model were then passed as inputs to a CNN to form nonlinear Type II models. CNN architectures ranging from 10−3−1 to 10−5−1 were explored. A 10−5−1 CNN proved to be the best model in accordance with rms error and predictive generalization. This model generated rms errors of 15.67K ($r^2 = 0.952$), 15.08K ($r^2 = 0.959$), and 21.76K ($r^2 = 0.919$) for the training set, cross-validation set, and prediction set, respectively. This is a 39% improvement for the training set and an 18% improvement for the prediction set over the linear Type I results. Results were obtained by averaging the output values for each compound over 10 individual network trainings. Averaging minimizes the dependence of network training in the selection of initial random weights and biases.[43] This generally produces predictions much closer to experimental values than any single training run.

Due to the improvement of using CNNs in Type II modeling, Type III models were developed. The best model found using Type III feature selection included the 10 descriptors shown in Table 3. The descriptors V6CH-18, MOLC-8, NSB-12, and 1SP2-1 appear again in the Type III model which were described in detail previously. This clearly shows the important molecular contribution of these four descriptors and their effect on the T_g.

The descriptor V5C-10 is similar to V6CH-18 in that it encodes molecular features such as size and branching. The hybrid descriptor ELOW-1 combines electronic and topological aspects of a molecule to form atom-level topological indexes.[30] This descriptor encodes the through-space distance between atoms with the maximum and minimum E-state values. Electrotopological indexes, such as ELOW-1, describe the probability of interaction between atoms in a compound with other nearby atoms. The geometric descriptor SHDW-4 encodes the standardized area projected onto the XY-plane, which encodes information about the size and shape of the 3-D structure. DPOL-1 represents the dipole moment of the molecule. The descriptor CARB-1 represents the average charge of carbonyl carbons in the molecular structure. The hybrid descriptor DPSA-3 combines information about partial charges on atoms and their solvent accessible surface areas.[36] This descriptor attempts to describe the extent of interactions between molecules based on the solvent accessible surface areas of individual atoms.

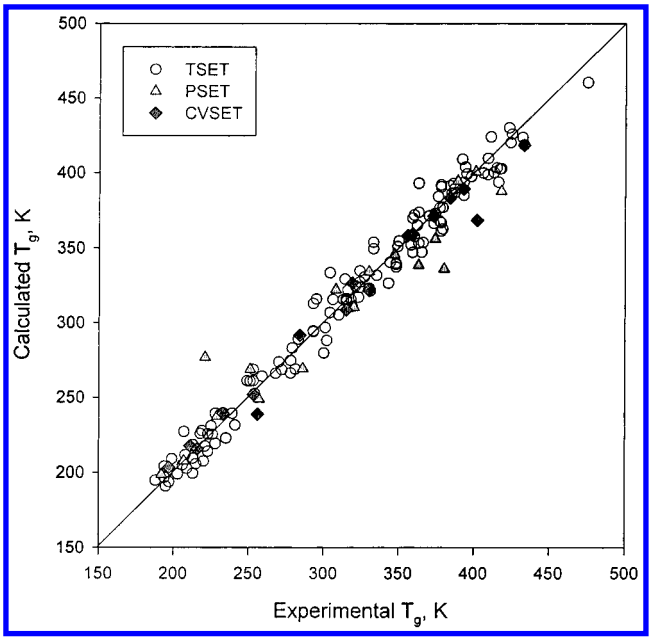

Figure 1. Plot of calculated vs experimental T_g values for monomer compounds used in Part 1 of this study using CNN models. Training set members include 131 compounds, prediction and cross-validation set members − 17 compounds each.

The best Type III model for Part 1 of this study is shown in Figure 1. A model architecture of 10−5−1 produced a training set rms error of 10.10K ($r^2 = 0.980$), cross-validation set rms error of 10.89K ($r^2 = 0.982$), and prediction set rms error of 21.69K ($r^2 = 0.917$). This is a 36% improvement for the training set, 28% for the cross-validation set, but negligible improvement for the prediction set over the Type II results. The model had difficulty in accurately predicting T_g values for two molecules in the prediction set: poly(2-isopropyl-1,4-butadiene) and poly(*tert*-butyl methacrylate). Both molecules are structurally similar to others in the data set so removal of them was not warranted.

Many different factors (rate of measurement, average molecular weight, etc., ...) have not been captured by molecular descriptors, which affect the T_g of polymeric materials. Despite this, accurate, robust models have been found, which show a relationship between monomer structure and T_g. Since the monomer structure does not completely represent the polymeric material as a whole, Part 2 employed the repeat unit of the polymer as a structural basis for generating the descriptors.

Part 2. The procedures outlined for Type I modeling in Part 1 were also used in Part 2. The *T*-value cutoff had to be reduced to three, however, to find good models. The best model contained the 10 descriptors shown in Table 4. This model had a training set rms error of 40.06K ($r^2 = 0.846$) and a prediction set rms error of 43.16K ($r^2 = 0.831$). The VIF did not exceed 10 for this model with pairwise correlations ranging from 0.523 to 0.896 (mean = 0.741). The high rms error is likely attributed to several statistical outlying compounds. Clearly, a linear function cannot capture

Table 4. Topological Descriptors Used in Linear Type I Model for Polymers Represented by Their Repeat Unit Structures

descriptor[a]	coefficient	error	range
KAPA-6	−6.23	1.48	0−17.13
V3C-8	79.7	12.1	0−1.9
V5C-10	181.1	53.9	0−0.5
S4C-9	−212.4	28.6	0−1.78
NN-4	43.5	4.7	0−4
NSB-12	−5.50	1.42	1−36
WTPT-2	398.2	31.4	1.5−2.1
EMIN-1	−16.1	3.1	−8.9−2
MDE−12	−11.5	2.8	0−5.4
MDE−13	−3.4	1.1	0−37.9
constant	−371.7	57.9	

[a] KAPA-6, atom valence-corrected third-order kappa index;[27] V3C-8, V5C-10, S4C-9 third- and fifth-order atom valence-corrected cluster, simple fourth-order cluster respectively;[41] NN-4, number of nitrogen molecules; NSB-12, number of single bonds; WTPT-2, ratio of molecular ID for molecule to number of atoms in molecule;[44] EMIN-1, minimum atomic E-state value;[30] MDE-12, MDE-13, molecular distance edge between primary/secondary and primary/tertiary carbons respectively.[31]

an accurate relationship between polymer repeat units and the T_g.

The descriptor KAPA-6[27] encodes information about the size and branching of the molecule. Furthermore, the descriptors V3C-8, V5C-10, and S4C-9[41] also contribute information about a molecule's size and branching. The number of nitrogen atoms (NN-4) and number of single bonds (NSB-12) are simple atom and bond counts. The descriptors WTPT-2,[44] MDE-12, and MDE-13[31] provide information about weighted paths and molecular distance edges in the molecule. Finally, the descriptor EMIN-1[30] describes the minimum atomic E-state value.

Type II modeling for Part 2 of this study used the same procedure as in Part 1. A 10−5−1 CNN architecture was determined to be optimal, producing a training set rms error of 27.33K ($r^2 = 0.929$), a cross-validation set rms error of 32.96K ($r^2 = 0.891$), and a prediction set rms error of 28.40K ($r^2 = 0.929$). This is a 32% improvement for the training set and a 34% improvement for the prediction set over the linear Type I model results. This shows that the nonlinearity of CNNs is useful for predicting T_g; therefore, Type III models were explored.

The size of the descriptor subsets examined for Type III modeling was increased to 11 to find better models than the Type II procedure. For this experiment, a simulated annealing optimization afforded better results than the genetic algorithm. The best model was found by using a CNN architecture of 11−7−1. The rms errors for the training set, cross-validation set, and prediction set were 21.14K ($r^2 = 0.958$), 25.24K ($r^2 = 0.934$), and 21.94K ($r^2 = 0.962$), respectively. This is a 23% improvement for all three sets, training set, cross-validation set, and prediction set over the Type II results. A plot of predicted versus experimental T_g values is shown in Figure 2. Note that the range of temperatures covered by this figure is significantly greater than the plot for Part 1 of this study. The 11-descriptor Type III model is shown in Table 5.

The descriptors KAPA-6, NN-4, WTPT-2, and MDE-12 (detailed description above) appear in the best subset of descriptors found by simulated annealing for Type I and Type III modeling. The χ index, V6C-11, provides information about the size and branching of the molecules. The number

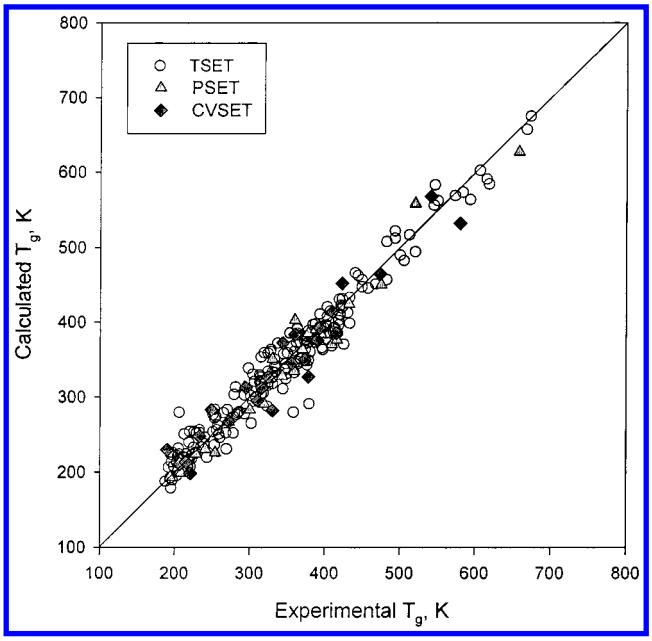

Figure 2. Plot of calculated vs experimental T_g values for repeat unit end-capped with hydrogen in Part 2 of this study using CNN models. Training set members include 201 compounds, prediction and cross-validation set members − 25 compounds each.

Table 5. Topological Descriptors Used in Nonlinear Type III Model for Polymers Represented by Their Repeat Unit Structures

descriptor[a]	range
KAPA-6	0−17.13
V6C-11	0−0.14
NN-4	0−4
NDB-13	0−5
NLP-19	0−16
WTPT-2	1.5−2.1
3SP2−1	0−11
2SP3−1	0−12
EDIF-1	0−23.4
MDE−11	0−2.89
MDE−12	0−4.70

[a] KAPA-6, NN-4, WTPT-2, MDE-12, see Table 4 caption; V6C-11, sixth-order valence cluster;[41] NDB-13, number of double bonds; NLP-19, number of lone pairs; 3SP2-1, number of sp^2 hybridized carbons bound to three other carbons; 2SP3-1, number of sp^3 hybridized carbons bound to two other carbons; EDIF-1, difference between max E-state value and minimum E-state value;[30] MDE-11, molecular distance edge between primary carbons.[31]

of double bonds that appear in the molecule is accounted for by NDB-13. One or more double bonds appear in 100 out of the 201 compounds in the training set so the incorporation of this descriptor in the model is important. The number of lone pairs of electrons (NLP-19) helps to measure the potential of intermolecular attraction forces that can occur between polymer segments in the material. The number of sp^2-hybridized carbons attached to three other carbons (3SP2-1) is accounted for by the model. This is most likely describing the different carbon types that can occur with alkyl-substituted styrene molecules and alkyl-substituted butadiene molecules. A count of sp^3-hybridized carbons attached to two other carbon atoms (2SP3-1) is included in the model. This descriptor can indirectly describe polymer backbone flexibility. If many methylene units appear in the backbone, then more movement is allowed due to rotations around the single bonds then backbones with fewer meth-

ylene groups. The electrotopological descriptor EDIF-1 can measure the extent of intermolecular interactions that can occur between polymer units in the material. Molecular distance edge information between primary carbons is provided by MDE-11.

Calculating molecular descriptors from the repeat unit structure produced good quantitative models. In this case, it was also significant that only topological descriptors were used. Topological descriptors are simple and relatively easy to calculate computationally thus saving time. Also, producing geometry-optimized structures which are not needed for topological descriptors, can be avoided saving additional time. Because of this, models utilizing only topological descriptors may be preferred over their geometric and electronic counterparts for some applications. The rms errors shown above prove that predictive models can be produced taking into account only topological considerations. Even though many other experimental factors have a profound affect on the value of the T_g, glass transition temperatures can be predicted within ±22 K.

Randomization Experiments. Several precautionary steps were taken during model building to reduce the odds of chance correlations between independent variables and the T_g. Randomization experiments were also performed to prove that models were capturing the structural significance of molecules and their relation to the property of interest. The dependent variable was randomly scrambled and used in the experiment. Models were then investigated, analogously to Type I, II, and III methods, to find the most predictive models. The rms errors and correlation coefficients found using random dependent variables should be very poor if the original models did accurately represent the relationship between chemical structure and the T_g.

The same model sizes and architectures that produced the best models for the standard experiment were tested with the randomized dependent variable. For Part 1 of this study, using the random dependent variables with the same architecture as the best Type III model produced rms errors of 41.67K ($r^2 = 0.677$), 53.17K ($r^2 = 0.642$), and 80.79K ($r^2 = 0.010$) for the training set, cross-validation set, and prediction set, respectively. The prediction set rms error and r^2-value indicates that a poor correlation was found between structure and T_g, which proves the validity of the real models. The same random experiment was done for Part 2 using the size and architecture of the best Type III model. The CNN output produced training set, cross-validation set, and prediction set rms errors of 78.87K ($r^2 = 0.400$), 94.72K ($r^2 = 0.292$), and 112.39K ($r^2 = 0.001$), respectively. It is clear again that the real models have found valid relationships between chemical structure and T_g due to the poor prediction set rms error and r^2-value of the random model.

CONCLUSION

For Part 1 of this study, models were developed which related monomer structure to the property of interest (T_g). The best model found produced a training set rms error of 10.10K, a cross-validation set rms error of 10.89K, and a prediction set rms error of 21.69K. It was thought that the repeat unit structure end-capped with hydrogen atoms would represent the polymeric material more realistically and produce more accurate models. Since geometric and electronic descriptors would misrepresent the polymer (as a whole), only topological descriptors were calculated for the structures. The best model found generated a training set rms error of 21.14K, a cross-validation set rms error of 25.24K, and a prediction set rms error of 21.94K. Even with the larger range and diversity of the data set, models containing simple topological descriptors could still accurately predict to nearly ±20K.

REFERENCES AND NOTES

(1) Bicerano, J. *Computational Modeling of Polymers*; Hudgin, D. E., Ed.; Marcel Dekker: New York, 1992; Plastics Engineering Series, Vol. 25.

(2) Bicerano, J. *Prediction of Polymer Properties*; Hudgin, D. E., Ed.; Marcel Dekker: New York, 1993; Plastics Engineering Series, Vol. 27.

(3) Seymour, R. B.; Carraher, C. E., Jr. *Structure−Property Relationships in Polymers*; Plenum Publishing Corporation: New York, 1984.

(4) Meier, D. J. *Molecular Basis of Transitions and Relaxations*; Elias, H.-G., Ed.; Gordon & Breach: New York, 1978; Midland Macromolecular Monographs Series, Vol. 4.

(5) Krevelen, D. W. v. *Properties of Polymers − Their Estimation and Correlation with Chemical Structure*, 2nd ed.; Elsevier: New York, 1976.

(6) Painter, P. C.; Coleman, M. M. *Fundamentals of Polymer Science − An Introductory Text*, 2nd ed.; Technomic Publishing Company, Inc.: Lancaster, PA, 1997.

(7) Askadskii, A. A. Some Rules Concerning the Effect of Chemical and Supermolecular Structure on the Softening Point of Amorphous Polymers. *Polym. Sci. U.S.S.R.* **1966**, *9*, 471−487.

(8) Askadskii, A. A.; Slonimskii, G. L. Universal System of Calculation for Determining the Glass Transition Temperature of Polymers. *Polym. Sci. U.S.S.R.* **1971**, *13*, 2158−2160.

(9) Gao, H.; Harmon, J. P. An Empirical Correlation Between Glass Transition Temperatures and Structural Parameters for Polymers with Linear and Branched Alkyl Substituents. *J. Appl. Polym. Sci.* **1997**, *64*, 507−517.

(10) Wiff, D. R.; Altieri, M. S.; Goldfarb, I. J. Predicting Glass Transition Temperatures of Linear Polymers, Random Copolymers, and Cured Reactive Oligomers from Chemical Structure. *J. Polym. Sci., Polym. Phys. Ed.* **1985**, *23*, 1165−1176.

(11) Hopfinger, A. J.; Koehler, M. G.; Pearlstein, R. A. Molecular Modeling of Polymers. IV. Estimation of Glass Transition Temperatures. *J. Polym. Sci., Part B, Polym. Phys.* **1988**, *26*, 2007−2028.

(12) Boudouris, D.; Constantinou, L.; Panayiotou, C. Prediction of Volumetric Behavior and Glass Transition Temperature of Polymers: A Group Contribution Approach. *Fluid Phase Equilibria* **2000**, *167*, 1−19.

(13) Camelio, P.; Cypcar, C. C.; Lazzeri, V.; Waegell, B. A Novel Approach Toward the Prediction of the Glass Transition Temperature: Application of the EVM Model, a Designer QSPR Equation for the Prediction of Acrylate and Methacrylate Polymers. *J. Polym. Sci., Part A, Polym. Chem.* **1997**, *35*, 2579−2590.

(14) Cypcar, C. C.; Camelio, P.; Lazzeri, V.; Mathias, L. J.; Waegell, B. Prediction of the Glass Transition Temperature of Multicyclic and Bulky Substituted Acrylate and Methacrylate Polymers Using the Energy, Volume, Mass (EVM) QSPR Model. *Macromolecules* **1996**, *29*, 8954−8959.

(15) Katritzky, A. R.; Sild, S.; Lobanov, V.; Karelson, M. Quantitative Structure−Property Relationship (QSPR) Correlations of Glass Transition Temperatures of High Molecular Weight Polymers. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 300−304.

(16) Joyce, S. J.; Osguthorpe, D. J.; Padgett, J. A.; Price, G. J. Neural Network Prediction of Glass-transition Temperatures from Monomer Structure. *J. Chem. Soc., Faraday Trans.* **1995**, *91*, 2491−2496.

(17) Ulmer, C. W., II.; Smith, D. A.; Sumpter, B. G.; Noid, D. I. Computational Neural Networks and the Rational Design of Polymeric Materials: The Next Generation Polycarbonates. *Comput. Theor. Polym. Sci.* **1998**, *8*, 311.

(18) Sumpter, B. G.; Noid, D. W. On the Use of Computational Neural Networks for the Prediction of Polymer Properties. *J. Thermal Anal.* **1996**, *46*, 833−851.

(19) Porter, D. *Group Interaction Modeling of Polymer Properties*; Marcel Dekker: New York, 1995.

(20) Jurs, P. C.; Chou, J. T.; Yuan, M. Studies of Chemical Structure-Biological Activity Relations Using Pattern Recognition. In *Computer-Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979.

(21) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
(22) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulating Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(23) Wessel, M. D. Ph.D. Thesis, Department of Chemistry, The Pennsylvania State University: University Park, PA, 1997.
(24) Lu, X.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure−Activity Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.
(25) *MOPAC*, v. 6.0; Quantum Chemistry Program Exchange, Program 455; Indiana University: Bloomington, IN.
(26) Stewart, J. P. P. Mopac: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.
(27) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat.* **1986**, *5*, 7−12.
(28) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17.
(29) Kier, L. B.; Hall, L. H. Intermolecular Accessibility: The Meaning of Molecular Connectivity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 792−795.
(30) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.
(31) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, λ. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
(32) Rohrbaugh, R. H.; Jurs, P. C. Molecular Shape and the Prediction of High-Performance Liquid Chromatographic Retention Indexes of Polycyclic Aromatic Hydrocarbons. *Anal. Chem.* **1987**, *59*, 1048−1054.
(33) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4−12.
(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) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure−Property Relationships. *J. Comput. Chem.* **1992**, *13*, 492−504.
(36) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(37) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative-Structure Property Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(38) Jurs, P. C. *Computer Software Applications in Chemistry*, 2nd ed.; Wiley: New York, 1996.
(39) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.
(40) Wessel, M. D.; Jurs, P. C. Prediction of Reduced Ion Mobility Constants from Structural Information Using Multiple Linear Regression Analysis and Computational Neural Networks. *Anal. Chem.* **1994**, *66*, 2480−2487.
(41) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press Ltd.: Wiley: 1986.
(42) Wessel, M. D.; Jurs, P. C. Prediction of Human Intestinal Absorption of Drug Compounds From Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.
(43) Kauffman, G. W.; Jurs, P. C. Prediction of Inhibition of the Sodium Ion-Proton Antiporter by Benzoylguanidine Derivatives from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 753−761.
(44) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.
(45) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

CI010062O