Prediction of Surface Tension, Viscosity, and Thermal Conductivity for Common Organic Solvents Using Quantitative Structure−Property Relationships

Gregory W. Kauffman and Peter C. Jurs*

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

Received October 2, 2000

Predictive models for the surface tension, viscosity, and thermal conductivity of 213 common organic solvents are reported. The models are derived from numerical descriptors which encode information about the topology, geometry, and electronics of each compound in the data set. Multiple linear regression and computational neural networks are used to train and evaluate models based on statistical indices and overall root-mean-square error. Eight-descriptor models were developed for both surface tension and viscosity, while a nine-descriptor model was developed for thermal conductivity. In addition, a single nine-descriptor model was developed for prediction of all three properties. The results of this study compare favorably to previously reported prediction methods for these three properties.

INTRODUCTION

Solvent properties are important in many areas of scientific research. A recent book by Ash et al. categorized over 1700 organic liquids deemed "solvents" into their respective areas of application.[1] The industries represented in this compilation include textile, agricultural, automotive, chemical processing, cosmetic, ink manufacturing, petroleum, and pharmaceutical. In addition, the selective use of solvents in academic environments is practiced on a daily basis by scientists of every discipline.

Surface tension (σ), viscosity (η), and thermal conductivity (λ) are classified as surface and transport properties of a liquid. The magnitudes of these properties are dependent upon intermolecular interactions between the solvent molecules. These interactions dictate the surface elasticity and the flow and the miscibility of two phases as well as the ability of the liquid to dissipate temperature fluctuations during a reaction process. In the current work, an attempt is made to use numerical descriptors to encode fundamental features of 213 common organic solvents based solely on molecular structure and to identify those features which best model the surface tension, viscosity, and thermal conductivity of each solvent at a single temperature. Our quantitative structure−property relationship (QSPR) approach to model building uses multiple linear regression (MLR) analysis and computational neural networks (CNNs). Each model generated is then validated using an external prediction set of compounds.

QSPRs have been reported for a variety of physical properties which have been recently and thoroughly reviewed.[2] In particular, many recent QSPRs for the prediction of surface tension and viscosity have appeared in the literature. Two papers have reported QSPR studies for surface tension. The first, by Stanton and Jurs,[3] employed the ADAPT software package[4,5] to find a 10-descriptor multiple linear regression (MLR) model for a limited set of 166 alkanes, aliphatic esters, and alcohols. The second, by Kavun and co-workers,[6] employed ESMA[7] to find a seven-descriptor MLR model for a diverse but a small data set of 81 compounds.

Since 1997, four papers reporting QSPR models for viscosity have been published. The first, reported by Suzuki and co-workers,[8] described a quantitative property−property relationship for a diverse set of 361 compounds using the software package *ChemProp*.[9] Nine-descriptor MLR and CNN models using four physicochemical properties and five indicator variables were developed. As pointed out in subsequent papers, the use of physical properties for the prediction of another property can have serious limitations.[10,11] The remaining three papers employed *CODESSA*[12] to develop either MLR models or PLS regression models. Ivanciuc and co-workers[10] developed a five-descriptor MLR model for a diverse set of 337 compounds using a leave 20% out cross-validation method. Similarly, Katritzky and co-workers[11] developed a five-descriptor model for a set of 361 compounds using a one-third data subset cross-validation. Another model was reported by Cocchi and co-workers.[13] Using a small set of 46 compounds, a 16-variable four-component PLS regression model was developed. The results of all three CODESSA-based models were consistent with one another and employed similar descriptors.

The common theme among each of the models presented in the literature for predicting surface tension and viscosity is the use of descriptors which encode information about hydrogen-bonding and polar surface interactions. This makes sense due to the influence of such electronic features in real-system intermolecular interactions. Therefore, it was expected that several such descriptors would appear in the models generated in the present work. No literature reports of QSPRs for thermal conductivity could be found; however, the close relationship to viscosity would allow one to confidently assume that similar descriptors would be useful for the prediction of both properties.

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

10.1021/ci000139t CCC: $20.00
Published on Web 03/02/2001

EXPERIMENTAL SECTION

Data for the 213 organic solvents used in this study were taken from a 1998 compilation by Marcus.[14] Experimental values were used whenever available; however, values from the DIPPR database[15] were also used when reliable experimental data was not provided. The surface tension data set contained 199 solvent compounds spanning the range 11.0−63.3 mN/m (milliNewton per meter). The 212 compounds used in the viscosity study spanned the range 0.215−944.06 mPa·s (milliPascal second); therefore, the corresponding log values, ranging from −0.668−2.975, were used as the dependent variable. Finally, thermal conductivity data ranged from 0.088 to 0.353 W/m·K (Watt per meter Kelvin) for the 185 compounds used. The names and respective experimental values of surface tension, viscosity, and thermal conductivity for each compound are presented in Table 1.

The corresponding experimental errors for the values used in this study varied greatly due to numerous primary sources being used for the compilation. Errors for surface tension varied the least with a range of only 3−5% error across the compounds in the data set. The experimental error for viscosity measurements had the largest range of fluctuation, with errors as low as 1% for some compounds and as high as 25% for others. An average experimental error of 5% over the entire data set though shows that very few of the compounds had extremely high errors. The experimental values for thermal conductivity had errors of 3−25% associated with them. The average error for these compounds though was 10%, again indicating that very few compounds had high errors associated with them. The accuracies cited here are a combination of reported experimental errors from the primary sources available to the authors and the quality factor reported for values used from the DIPPR database. These values will be compared to the accuracies of the predicted values generated in this study.

For each property, approximately 10% of the compounds were randomly chosen to serve as an external prediction set. These compounds were not used for descriptor selection or model training, rather they were used only for validation once the best models had been identified. For development of models using multiple linear regression (MLR) analysis, the remaining 90% of the compounds were used as the training set. For the development of nonlinear computational neural network (CNN) models, an additional 10% of the compounds were set aside to serve as a cross-validation set.

All computations for this study were performed on a DEC 3000 AXP model 500 workstation running the UNIX operating system. The Automated Data Analysis and Pattern recognition Toolkit (ADAPT) software package,[4,5] simulated annealing[16] and CNN[17] routines used to develop the QSPR models described were written independently at Penn State. The stages of QSPR development in this study include the following: structure entry and geometry optimization, descriptor generation, descriptor feature selection, MLR model formation and validation, and CNN model formation and validation.

Structure Entry and Optimization. All compounds were sketched using HyperChem (Hypercube, Inc., Waterloo, ON) on a Pentium PC. HyperChem connection tables provide information about atom types and bonding arrangements which can be used for geometry optimization. Three-dimensional representations for each of the compounds were obtained using a PM3 Hamiltonian[18] in the semiempirical molecular orbital package MOPAC.[19] Optimization using an AM1 Hamiltonian[20] was used to calculate geometry-dependent charge information. The suitability of both Hamiltonians for these purposes is described in the literature.[21]

Descriptor Generation. The ADAPT software package calculates numerical descriptors which encode information about the topology, geometry, and electronic nature of a compound. The topological class of descriptors describes the constitution and connectivity of a compound. These include specific atom and bond counts, shape[22,23] and branching indices,[24,25] and distance-edge measures.[26] These descriptors require no geometry optimization prior to calculation. Geometric descriptors capture information about spatial orientation of atoms in a molecule, therefore requiring good low-energy conformations for accurate calculation. Examples of these descriptors include moments of inertia, solvent-accessible surface areas,[27] and shadow projections.[28] Electronic descriptors provide information about partial atomic charges, dipole moment, and HOMO and LUMO energies. Three of the electronic descriptors which capture atomic charge information are calculated by an empirical scheme which is topological in nature.[29−30] The remaining seven are geometry dependent. Finally, topological, geometric, and electronic information is hybridized into a final class termed the charged-partial surface area (CPSA) descriptors.[31] These combine information about atomic charges and solvent-accessible surface areas thereby describing polar surface interactions of the molecules. Analogously, descriptors which encode CPSA information around sites of probable hydrogen-bonding interactions can be calculated.[32−33] A total of 138 topological, 43 geometric, 10 electronic, and 48 hybrid descriptors were calculated for the entire data set. A complete list of descriptors calculated is available upon request from the authors.

Descriptor Feature Selection. Objective feature selection is first used to remove descriptors with little or redundant information from the initial 239-member pool without making use of the dependent variable. First, descriptors with identical values for greater than 90% of the compounds in the training set were removed. Second, any one of two descriptors with a pairwise correlation above 0.93 was randomly eliminated. This reduced the descriptor pools to 105 members for surface tension, 105 members for viscosity, and 98 members for thermal conductivity.

Subjective feature selection uses the dependent variable to select information-rich subsets of descriptors for model development using only the training set compounds. For each property, simulated annealing coupled with an MLR fitness evaluator[16] was used to search the descriptor space for models comprised of 4−12 descriptors. The initial selection criterion for a descriptor subset as a potential model is that each descriptor must have a T-value magnitude greater than four, where a T-value is defined as the quotient of a model coefficient and its standard error. Final ranking of descriptor subsets is based on minimization of the rms error.

MLR Model Formation and Validation. The best three descriptor subsets for each property were first submitted to regression diagnostics to test for the presence of compound outliers in the training set and multicollinearlities among the descriptors within each subset. The presence of compound

Table 1. Solvents and Property Experimental Values

no.	solvent	surface tension ($mN \cdot m^{-1}$)	viscosity[a] (mPa·s)	thermal conductivity ($W \cdot m^{-1} K^{-1}$)	no.	solvent	surface tension ($mN \cdot m^{-1}$)	viscosity[a] (mPa·s)	thermal conductivity ($W \cdot m^{-1} K^{-1}$)
1	*n*-pentane	15.5[b]	−0.648	0.1125	76	phetole	32.9[c]	0.056[c]	0.1397
2	2-methylbutane	14.5	−0.668	0.1096	77	diphenyl ether	39.4	0.415[c]	
3	*n*-hexane	17.9	−0.532	0.1196	78	dibenzyl ether	38.2	0.668[b]	0.1249
4	cyclohexane	24.6[c]	−0.047	0.1234	79	1,2-dimethoxybenzene		0.516	
5	*n*-heptane	19.7	−0.401	0.1247	82	propionaldehyde		−0.498[b]	0.1601
6	*n*-octane	21.2	−0.288	0.1277	81	butyraldehyde	29.9	−0.367	0.1451
7	2,2,4-trimethylpentane	18.3[b]	−0.323[b]	0.0982	82	benzaldehyde	38.3[b]	0.121	0.1525
8	*n*-decane	23.4	−0.065[c]	0.1318	83	*p*-methoxybenzaldehyde		0.625[c]	
9	*n*-dodecane	24.9	0.139[c]	0.1354	84	cinnamaldehyde		0.732	
10	*n*-hexadecane	27.1	0.452	0.1421	85	acetone	22.7	−0.519	0.1605
11	benzene	28.2	−0.220	0.1433	86	2-butanone	23.7	−0.423	0.145[c]
12	toluene	27.9[c]	−0.257	0.1323	87	2-pentanone	24.5	−0.334	0.1420
13	*o*-xylene	29.5	−0.121	0.1313	88	methyl-*i*-propyl ketone	24.8	−0.368	0.1424
14	*m*-xylene	28.1	−0.236	0.1302	89	3-pentanone	24.8	−0.355	0.1439
15	*p*-xylene	27.8	−0.218	0.1297	90	cyclopentanone	33.2	0.116	0.1484
16	ethylbenzene	28.5	−0.196	0.1289	91	methyl-*i*-butyl ketone	23.2[b]	−0.263[c]	0.1439[c]
17	cumene	27.7[b]	−0.131	0.1232[c]	92	methyl-*tert*-butyl ketone		−0.147[b]	0.1384[c]
18	mesitylene	28.3	0.017	0.1351	93	cyclohexanone	35	0.302	0.1403
19	styrene	31.6	−0.157	0.1365	94	2-heptanone	26.1	−0.119[c]	0.1375
20	tetralin	34.5	0.330	0.1296	95	3-heptanone	25.5	−0.129	0.1360
21	*cis*-decalin	31.6	0.482	0.1130	96	acetophenone	38.8	0.220	0.1471[b]
22	methanol	22.3	−0.259	0.1999	97	benzophenone	45.1	1.134	
23	ethanol	21.9	0.035	0.1681	98	acetyl acetone	30.3[c]	−0.115[b]	0.1533
24	*n*-propanol	23.1	0.288	0.156[b]	99	formic acid	37[b]	0.294	0.2698
25	*i*-propanol	21.2	0.310	0.1350	100	acetic acid	26.9	0.053	0.1593
26	*n*-butanol	24.2	0.410	0.1530	101	propanoic acid	26.2	0.010	0.1465
27	*i*-butanol	22.5[b]	0.523	0.1318	102	*n*-butanoic acid	26.2	0.184	0.1466
28	2-butanol	23	0.477	0.1344	103	*n*-pentanoic acid	26.1	0.296[b]	0.1420
29	*t*-butanol	20.1	0.647	0.1158	104	*n*-hexanoic acid	27.5	0.451[c]	0.142[c]
30	*n*-pentanol	25.2	0.546	0.1528	105	*n*-heptanoic acid	27.8	0.584	0.1426
31	*i*-pentanol	23.9	0.573	0.1407	106	dichloroacetic acid	35.4	0.772	0.1869
32	*t*-pentanol	22.3[c]	0.550[c]	0.1213	107	trifluoroacetic acid	13.5[c]	−0.068	0.1621
33	*n*-hexanol	25.7	0.662[b]	0.1537[b]	108	acetic anhydride	31.9	−0.075	0.1640
34	cyclohexanol	33.8	1.613	0.1341	109	benzoyl chloride	38.7	0.056	0.1041[b]
35	*n*-octanol	26.9[b]	0.867	0.1598	110	methyl formate	23.9	−0.484	0.1851
36	*n*-decanol	28.4	1.054	0.1615[c]	111	ethyl formate	24	−0.424	0.1603
37	*n*-dodecanol	29.4[b]	1.196[b]	0.1496[b]	112	methyl acetate	24.1	−0.194	0.1534
38	benzyl alcohol	39.5	0.816	0.1603	113	ethyl acetate	23.1	−0.371	0.1439
39	2-phenylethanol	40.6	0.155[c]	0.1627	114	propyl acetate	23.7	−0.259	0.1409[b]
40	allyl alcohol	25.3	0.125	0.1546	115	butyl acetate	24.5	−0.162	0.1367
41	2-chloroethanol	38.9[c]	0.484	0.1332	116	*i*-pentyl acetate	24.2[b]	−0.103[c]	0.1304
42	2,2,2-trifluoroethanol		0.244		117	methyl propanoate	24.4	−0.366	0.1453
43	hexafluoro-2-propanol	16.1	0.198		118	ethyl propanoate	23.7	−0.299[b]	0.1387
44	2-methoxyethanol	30.8	0.207	0.1880	119	diethyl malonate	31.1	0.288	0.1503
45	2-ethoxyethanol	28.2	0.267	0.1757[c]	120	methyl benzoate	37.5[c]	0.269	0.1536[c]
46	1,2-propanediol	36.5	1.625	0.2004[b]	121	ethyl benzoate	34.8	0.289	0.1443
47	1,3-propanediol	45.2	1.668	0.2226	122	dimethyl phthalate	40.4	1.157	0.1487
48	1,2-butanediol	35.3	1.699[b]	0.1730	123	dibutyl phthalate	33.4	1.188[c]	0.1361
49	2,3-butanediol	30.6	1.818		124	ethyl chloroacetate	31.3[b]	0.045	0.1362[b]
50	1,4-butanediol	44.2[b]	1.854	0.2059[c]	125	ethyl trichloroacetate		−0.363	
51	1,5-pentanediol	43.4	2.059	0.2000	126	ethyl acetoacetate	31.3	0.178[b]	0.1515
52	diethyleneglycol	48.5[c]	1.477	0.2037	127	4-butyrolactone	38.5	0.235	0.1613
53	triethyleneglycol	45.2	1.690	0.1931	128	perfluoro-*n*-hexane	11[c]	−0.179	
54	glycerol	63.3	2.975	0.2918	129	perfluoro-*n*-heptane	11.9	−0.050	
55	phenol	38.8	0.544[c]	0.1565[c]	130	perfluoro-methylcyclohexane	14	−0.059	
56	2-methylphenol	35	0.881[c]	0.1517	131	perfluoro-decalin	15	0.711	
57	3-methylphenol	37[b]	0.992	0.1493	132	fluorobenzene	27.1	−0.260	0.1260
58	4-methylphenol	34.6[b]	0.973	0.1426	133	hexafluorobenzene	21.6[b]	−0.066	0.0882
59	2,4-dimethylphenol	31.2	1.836	0.1612[b]	134	1-chlorobutane	23.4	−0.371	0.1187
60	3-chlorophenol		1.063		135	chlorobenzene	32.5	−0.120[b]	0.1269
61	diethyl ether	16.5	−0.616	0.1282	136	dichloromethane	27.2	−0.386	0.1390
62	di-*n*-propyl ether	19.9[c]	−0.470	0.1266	137	1,1-dichloroethane	24.2	−0.297	0.1110
63	di-*i*-propyl ether	17.2	−0.421[b]	0.1093	138	1,2-dichloroethane	31.5	−0.108	0.1347
64	di-*n*-butyl ether	22.5	−0.190	0.1279[c]	139	*trans*−1,2-dichloroethylene	27.8[b]	−0.415	0.1120
65	di(2-chloroethyl) ether	37	0.330		140	*o*-dichlorobenzene	36.2	0.122	0.1211
66	1,2-dimethoxyethane	24.6	−0.342[c]	0.1405	141	*m*-dichlorobenzene	35.5	0.012[b]	0.1172
67	bis(methoxyethyl) ether	30.4	−0.005		142	chloroform	26.5	−0.271	0.1175
68	furan	23.4	−0.442	0.1262	143	1,1,1-trichloroethane	24.9	−0.100	0.1012[c]
69	tetrahydrofuran	26.4[b]	−0.335	0.1200	144	1,1,2-trichloroethane	33	0.042	0.1328
70	2-methyl tetrahydrofuran		−0.325		145	trichloroethylene	28.8[c]	−0.274	0.1150
71	tetrahydropyran		−0.117		146	1,2,4-trichlorobenzene	44.7[c]	0.426	0.1117
72	1,4-dioxane	32.8	0.077	0.1588	147	tetrachloromethane	26.1	−0.045	0.0997
73	1,3-dioxolane		−0.222		148	tetrachloroethylene	31.3[c]	−0.075	0.1100
74	1,8-cineole	31.1[b]	0.362		149	1,1,2,2-tetrachloroethane	35.4	0.197[c]	0.1127[b]
75	anisole	34.6	−0.007[c]	0.1560	150	pentachloroethane	34.2[b]	0.357[b]	0.094[c]

Table 1 (Continued)

no.	solvent	surface tension ($mN \cdot m^{-1}$)	viscosity[a] (mPa·s)	thermal conductivity ($W \cdot m^{-1} K^{-1}$)	no.	solvent	surface tension ($mN \cdot m^{-1}$)	viscosity[a] (mPa·s)	thermal conductivity ($W \cdot m^{-1} K^{-1}$)
151	1-bromobutane	24.8[b]	−0.224	0.1037	183	2,4,6-trimethylpyridine	31.8	−0.094	0.1463[b]
152	bromobenzene	35.5	0.029	0.1108[c]	184	pyrimidine	30.3		0.1521
153	dibromomethane	40.1	−0.008[b]	0.1085	185	quinoline	45.2	0.498	0.1492
154	1,2-dibromoethane	38.3	0.207	0.1011	186	acetonitrile	28.3	−0.467	0.1877
155	bromoform	45[c]	0.271	0.0994	187	propionitrile	26.7[c]	−0.393	0.1677[b]
156	1-iodobutane	28.7	−0.083		188	butyronitrile	26.8	−0.260	0.1673
157	iodobenzene	38.8	0.181	0.1000	189	valeronitrile	27	−0.160	0.165[b]
158	diiodomethane	50	0.414	0.098[b]	190	acrylonitrile	26.7	−0.470	0.165[b]
159	*n*-butylamine	23.5	−0.238[c]	0.1607	191	benzyl cyanide	40.8	0.292	0.1245
160	benzylamine	39.5	0.201	0.1669	192	benzonitrile	38.5	0.092	0.1485
161	1,2-diaminoethane	40.1	0.188	0.2322[c]	193	nitromethane	36.3	−0.212	0.2068
162	diethylamine	19.4	−0.539	0.1341[b]	194	nitroethane	32.1	−0.195[c]	0.1631
163	di-*n*-butylamine	24.1	−0.024	0.1334	195	1-nitropropane	30.1	−0.102	0.1542
164	pyrrole	37.1	0.091	0.1641	196	2-nitropropane	29.3	−0.142	0.1408
165	pyrrolidine	29.2	−0.154[b]	0.1592	197	nitrobenzene	42.4[c]	0.251	0.1480
166	piperidine	29.4	0.134	0.1789[b]	198	formamide	58.2	0.519	0.3529
167	morpholine	36.9	0.303[b]	0.1643	199	*N*-methylformamide	39.5	0.217	0.2127[b]
168	triethylamine	20.1	−0.440[c]	0.1187	200	*N*,*N*-dimethylformamide	36.4	−0.096	0.1840
169	tri-*n*-butylamine	24.3[c]	0.118	0.1204[c]	201	*N*,*N*-dimethylthioformamide	45.4	0.297	
170	aniline	42.8	0.576	0.1722	202	*N*,*N*-diethylformamide		0.098	
171	*o*-chloroaniline	43.1	0.465	0.1523[b]	203	*N*-methylacetamide	32.9	0.562[c]	
172	*N*-methylaniline	39.7	0.303	0.1577	204	*N*,*N*-dimethyl acetamide	31.7	−0.033	0.1672
173	*N*,*N*-dimethylaniline	25.6	0.110	0.1419	205	2-pyrrolidinone	46.3	1.124	0.1943
174	ethanolamine	48.3	1.287	0.2366	206	*N*-methylpyrrolidinone	40.7[c]	0.222	0.1340
175	diethanolamine	49	2.546		207	*N*-methylthiopyrrolidinone		0.628[b]	0.1344
176	triethanolamine	45.2	2.788	0.1964[c]	208	tetramethylurea		0.145	
177	pyridine	36.3	−0.054	0.1624	209	dimethyl sulfide	23.8	−0.554	0.1407[c]
178	2-methylpyridine	32.8	−0.123	0.1465	210	diethyl sulfide	24.5	−0.380	0.1324
179	3-methylpyridine	34.5	−0.059	0.1368	211	di-*n*-butyl sulfide	26.8	−0.010[b]	
180	4-methylpyridine	35.5	−0.062	0.1368	212	tetrahydrothiophene	35	−0.013	0.1409
181	2,4-dimethylpyridine	33.2	−0.052		213	thiobis(2-ethanol)	52.9[c]	−0.076	0.1187
182	2,6-dimethylpyridine	31	−0.061	0.1302					

[a] Log(η) values. [b] Member of cross-validation set. [c] Member of prediction set.

outliers was evaluated by six tests which measure the ability of individual data points to influence the regression prediction and model coefficients. Variance inflation factors (VIF = $1/(1-R^2)$) were used to determine if multicollinearities existed among the descriptors in a single subset. A VIF less than 10 indicated that no multicollinearities were present. With this information at hand, the final model selections were then made based on the following sequential criterion: (1) the smallest descriptor subset sizes which did not compromise the overall rms errors, (2) contained no multicollinearities among its descriptors, and (3) minimal statistical compound outliers. Finally, the models identified as best were validated using the external prediction sets for each property.

CNN Model Formation and Validation. The final models from MLR analysis were then submitted to three-layer, fully connected, feed-forward CNNs. The number of input neurons was determined by the number of descriptors in each MLR model. Each descriptor input value was scaled and shifted onto the range of 0.05−0.95 using a linear transformation. The number of hidden neurons was adjusted until suitable network architectures to adequately model the properties were found. One limitation of adjusting hidden neurons is that the ratio of training set observations to adjustable parameters[34] must be greater than two to reduce the risk of chance correlations and network overtraining. Finally, a single output neuron was used to represent the learned response or solvent property of interest.

Network training was achieved using the quasi-Newton BFGS (Broyden-Fletcher-Goldfarb-Shanno) method.[35] Training was directed by optimization of weights and biases and ceased when minimization of the cross-validation set rms error was achieved. Beyond this minimum, the network fails to generalize trends in the training data and begins to memorize structural idiosyncrasies, which results in poor predictive ability. For each individual experiment in this study, all possible neural network architectures were examined within the limits of the training set observations to adjustable parameters ratio criterion. Final selection of the optimal network architecture was based on the lowest and most consistent training and cross-validation set rms errors.

RESULTS AND DISCUSSION

Surface Tension. The eight-descriptor model shown in Table 2 was identified as statistically optimal for predicting surface tension. All descriptor absolute T-values were greater than four and variance inflation factors were all below ten, indicating that no multicollinearities were present among the eight descriptors. The eight descriptors had pairwise correlations ranging from 0.687 to 0.928 with an average value of 0.805. Four statistical compound outliers were identified (**43**, **54**, **133**, **161**); however, removal of these observations showed a minimal effect on model coefficients and overall regression. The MLR model had a training set rms error of 3.37 mN/m ($R = 0.914$) and a prediction set rms error of 5.31 mN/m ($R = 0.915$). Model coefficients, the standard error of the coefficients, and descriptor value ranges for this model are included in Table 2.

Table 2. Descriptors Used for the MLR and CNN Models of Surface Tension

descriptor	type	coefficient	error	range	explanation[a]
constant		−46.80	3.315		
KAPA-4	top.	−3.301	0.158	2.0 to 25.94	κ shape index
QPOS-1	elec.	−19.46	3.112	0.05 to 0.99	charge of most positive atom
SADH-3	hyb.	87.40	4.665	0 to 0.46	(total SA of all donatable H)/(total molecular SA)
PPSA-1	hyb.	5.95×10^{-2}	4.49×10^{-3}	0 to 624	total SA of all partial positively charged atoms.
FNSA-3	hyb.	−140.6	15.93	−0.28 to 0	(atomic charge weighted partial negative SA)/(total molecular SA)[b]
GRAV-3	geom.	9.501	0.409	3.93 to 14.24	cube root of the gravitational index
RNCS-1	hyb.	0.200	3.504×10^{-2}	0 to 51.14	relative negative charged SA[c]
SAAA-3	hyb.	−13.96	3.285	0 to 0.89	(sum of SA of acceptor atoms)/(total molecular SA)[d]

[a] SA = surface area. [b] FNSA-3 is expressed by $[\Sigma(-SA_i)(Q^-_i)]/(SA_{TOT})$, where $-SA_i$ and Q^-_i are the surface area and charge contributions of the *i*th negatively charged atom, respectively, and SA_{TOT} is the total molecular surface area. [c] RNCS-1 is expressed by $SA_{MNEG} \times [(Q^-_{MNEG})/(Q^-_{TOT})]$, where SA_{MNEG} is the surface area of the most negatively charged atom, Q^-_{MNEG} is the partial charge of the most negatively charged atom, and Q^-_{TOT} is the total negative charge of the molecule. [d] Acceptor atoms are N, O, S, and F.

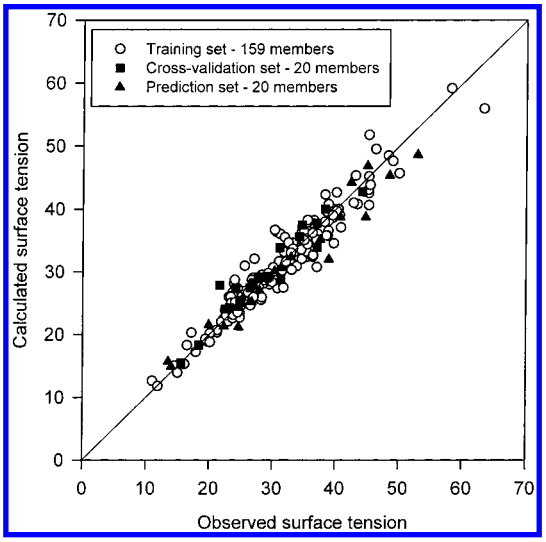

Figure 1. Plot of calculated versus observed surface tension for the training, cross-validation, and prediction set compounds used to develop the CNN model described in Table 2.

The eight descriptors from the MLR model were submitted to a CNN to see if nonlinear model formation would significantly improve the quality of the results. Architectures ranging from 8-2-1 to 8-7-1 were explored with an 8-6-1 network architecture (61 adjustable parameters) offering the best results. The training set rms error improved to 2.22 mN/m ($R = 0.965$), and the cross-validation set rms error was consistent with training having an rms error of 2.22 mN/m ($R = 0.960$). The rms error of the prediction set was 2.76 mN/m ($R = 0.976$), a 48% improvement over the MLR model. The average percent error of the predictions using this model was 5.3% for the training set, 6.1% for the cross-validation set, and 6.4% for the prediction set. Figure 1 shows a calculated versus observed plot for the CNN model of surface tension.

Examination of the descriptors used in this model shows that five of the eight are hybrid descriptors. The three CPSA descriptors, PPSA-1, FNSA-3, and RNCS-1, are encoding both positive and negative partial atomic charge information. The two hydrogen bonding-specific descriptors, SADH-3 and SAAA-3, encode complementary information about the partial surface areas of potentially hydrogen-bonding donator and acceptor atoms, respectively. In addition, one electronic descriptor, QPOS-1, is the charge on the most positive atom. The geometric descriptor, GRAV-3, is the cube-root of the gravitational index,[36] while the topological KAPA-4 descriptor is the atom type corrected first-order kappa index.[23] These two descriptors are encoding the size and shape properties of each solvent molecule.

A minor concern with the surface tension data set was the presence of two compounds in the training set (**54**, **198**) whose experimental values were noticeably larger than the bulk of the data. To determine whether the presence of these compounds artificially overrated the results, they were removed from the study and a new model was developed using a new 177-member training set. Based on all statistical requirements outlined previously, a 10-descriptor model was chosen as best for this experiment. The MLR model showed improved results with a training set rms error of 2.86 mN/m ($R = 0.929$) and a prediction set rms error of 4.58 mN/m ($R = 0.894$). The 10-7-1 CNN model also showed improved results with a training set rms error of 1.52 mN/m ($R = 0.980$), a cross-validation set rms error of 1.52 mN/m ($R = 0.982$), and a prediction set rms error of 2.57 mN/m ($R = 0.970$). These results indicate that compounds **54** (glycerol) and **198** (formamide) had an effect on model training. It should be noted that glycerol is the only triol in the data set, while formamide is the only non-*N*-methylated amide compound. The apparent contributions of the additional polar functionality to the surface tension makes modeling the property for these compounds difficult.

Viscosity. An eight-descriptor model was also deemed most statistically sound for modeling viscosity. These descriptors are shown in Table 3. Again, all descriptor absolute T-values were above four, and variance inflation factors were below ten. Pairwise correlations between the descriptors in this model range from 0.497 to 0.946 with an average of 0.828. Of the training set compounds, only two were identified as statistical outliers (**161**, **213**). Removal of these compounds demonstrated minimal influence on the model regression predictions or on the model coefficients. The MLR model produced rms errors of 0.257 mPa·s ($R = 0.913$) for the training set and 0.278 mPa·s ($R = 0.787$) for the prediction set. Model coefficients, standard error of the coefficients, and descriptor value ranges are included in Table 3.

The eight descriptors were submitted to a CNN and architectures from 8-2-1 to 8-8-1 were examined. An 8-7-1 architecture (71 adjustable parameters) was determined to be optimal giving rms errors for the training set of 0.147 mPa·s ($R = 0.974$) and for the cross-validation set of 0.148 mPa·s ($R = 0.965$). The prediction set rms error was 0.242 mPa·s ($R = 0.887$) which is only a 13% improvement over

Table 3. Descriptors Used for the MLR and CNN Models of Viscosity

descriptor	type	coefficient	error	range	explanation[a]
constant		−1.475	0.115		
V0-1	top.	0.263	2.629×10^{-2}	1.43 to 12.19	valence corrected zero-order χ-index
DPOL-1	elec.	9.831×10^{-2}	2.108×10^{-2}	0 to 4.95	dipole moment
SADH-3	comb.	−3.032	0.684	0 to 0.46	(total SA of all donatable H)/(total molecular SA)
SCDH-1	comb.	0.168	1.139×10^{-2}	0 to 26.81	sum of (SA*Q) for all donatable H
NRA-18	top.	7.096×10^{-2}	7.061×10^{-3}	0 to 12	number of ring atoms
FNSA-2	comb.	1.065	0.166	−1.95 to −0.02	(total charge weighted partial negative SA)/(total molecular SA)[b]
FNSA-3	comb.	−4.053	0.948	−0.28 to 0	(atomic charge weighted partial negative SA)/(total molecular SA)[c]
WPSA-3	comb.	-6.808×10^{-2}	1.442×10^{-2}	0 to 20.71	surface weighted charged partial positive SA[d]

[a] SA = surface area, Q = charge. [b] FNSA-2 is expressed by $[(Q^-_{TOT}) \times (-SA_{TOT})]/(SA_{TOT})$, where Q^-_{TOT} is the total negative charge for the molecule, $-SA_{TOT}$ is the total partial negative surface area of the molecule, and SA_{TOT} is the total molecular surface area. [c] FNSA-3 is expressed by $[\Sigma(-SA_i)(Q^-_i)]/(SA_{TOT})$, where $-SA_i$ and Q^-_i are the surface area and charge contributions of the ith negatively charged atom, respectively, and SA_{TOT} is the total molecular surface area. [d] WPSA-3 is expressed by $[(+SA_{TOT}) \times (SA_{TOT})]/1000$, where $+SA_{TOT}$ is the total partial positive surface area of the molecule and SA_{TOT} is the total molecular surface area.

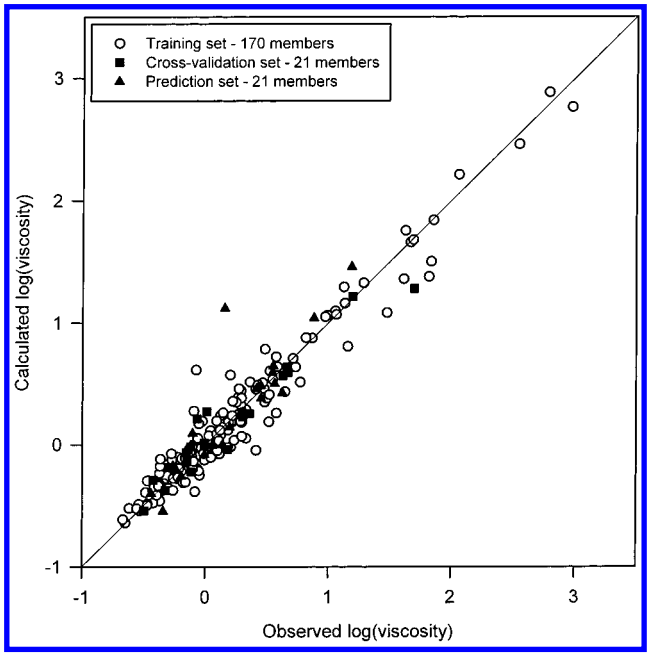

Figure 2. Plot of calculated versus observed log(viscosity) for the training, cross-validation, and prediction set compounds used to develop the CNN model described in Table 3.

the MLR model validation. However, one compound outlier in the prediction set (**40**) accounted for 75.6% of the variance in the rms error. Removal of this compound from the prediction set lowered the rms error to 0.122 mPa·s (R = 0.970), a 53% improvement over the MLR model. The average percent error of the predictions using this model was 24.9% for the training set, 24.9% for the cross-validation set, 62.7% for the prediction set with the outlier, and 24.8% for the prediction set without the outlier. Figure 2 shows a plot of calculated versus observed viscosity including the prediction set outlier.

As with the surface tension model, five of the eight descriptors incorporated into the viscosity model were hybrid descriptors; two of these being the same descriptor. Three CPSA descriptors, FNSA-2, FNSA-3, and WPSA-3, encode the total and atomic negative charge weighted surface areas as well as the total positive charge weighted surface area. Similarly, the two hydrogen-bonding specific descriptors, SADH-3 and SCDH-1, are encoding hydrogen-bonding interactions that may contribute to the viscosity of a substance. NRA-18 is a constitutional descriptor which encodes the number of ring atoms, and V0-1 is the valence corrected zero-order topological χ index.[24] These descriptors may offer information about how steric bulk of a substance effects trends in viscosity. Finally, the dipole moment, DIPO-1, is the sole electronic descriptor.

The calculated versus observed plot of viscosity (Figure 2) illustrates that the bulk of the data lies below approximately 1.5 log units. Further breakdown led to the investigation of models employing two new training sets, one with three training set compounds removed (**54**, **175**, **176**), and another with 13 training set compounds removed (**34**, **46**, **47**, **48**, **49**, **50**, **51**, **52**, **53**, **54**, **59**, **175**, **176**). Again, the purpose of this experiment was to examine the effect of these training set compounds on model development.

Removal of three training set compounds left a 188-member training set. For this experiment, a nine-descriptor MLR model was found with a training set rms error of 0.258 mPa·s (R = 0.881) and a prediction set rms error of 0.258 mPa·s (R = 0.844). The corresponding CNN model, using a 9-8-1 architecture, produced rms errors of 0.091 mPa·s (R = 0.986) for the training set, 0.087 mPa·s (R = 0.989) for the cross-validation set, and 0.258 mPa·s (R = 0.947) for the prediction set. The training results are better, while the validation results are comparable to the initial model containing the highly viscous compounds. However, no clear prediction set outlier is present which contributes more significantly to the rms error as before with the 191-member training set. Thus, one can conclude that the model developed without these three compounds is slightly less predictive than the original model which included them.

Removal of the 13 training set compounds left a 178-member training set. Again, a nine-descriptor MLR model was chosen as the most statistically sound. The training set rms error was 0.183 mPa·s (R = 0.893), and the prediction set rms error was 0.245 mPa·s (R = 0.854). The 9-8-1 CNN model showed improvement in training with rms errors of 0.095 mPa·s for both the training set (R = 0.972) and cross-validation set (R = 0.976). The prediction set rms error was relatively consistent at 0.255 mPa·s (R = 0.860); however, removal of one observation (**40**) which accounted for 53% of the variance in the rms error reduced this value to 0.177 mPa·s (R = 0.927). Graphically more convincing than rms errors may indicate, this model is clearly better than the original model using a 191-member training set. To account for this result, the structural composition of the compounds removed from training was examined. With the exception of compounds **34** (cyclohexanol) and **59** (2,4-dimethyl-

Table 4. Descriptors Used for the MLR and CNN Models of Thermal Conductivity

descriptor	type	coefficient	error	range	explanation[a]
constant		6.607×10^{-2}	5.368×10^{-3}		
KAPA-2	top.	5.260×10^{-3}	9.664×10^{-4}	0 to 15	κ shape index
SADH-3	hyb.	0.128	2.700×10^{-2}	0 to 0.46	(total SA of all donatable H)/(total molecular SA)
DPSA-3	hyb.	-2.118×10^{-3}	3.050×10^{-3}	11.33 to 76.28	difference in atomic charge weighted partial SA[b]
FPSA-3	hyb.	1.860	0.138	0 to 0.11	(atomic charge weighted partial positive SA)/(total molecular SA)[c]
FNSA-3	hyb.	−0.657	7.964×10^{-2}	−0.28 to 0	(atomic charge weighted partial negative SA)/(total molecular SA)[d]
1SP3−1	top.	-9.478×10^{-3}	1.566×10^{-3}	0 to 6.0	number of primary sp^3 carbons
MDE−14	top.	8.887×10^{-3}	1.678×10^{-3}	0 to 9.91	distance edge between 1^{o}and 4^{o}carbons
RPCS-1	hyb.	-3.955×10^{-3}	4.864×10^{-4}	0 to 24.79	relative positive charged SA[e]
RNCS-1	hyb.	6.846×10^{-4}	1.709×10^{-4}	0 to 51.1	relative negative charged SA[f]

[a] SA = surface area. [b] DPSA-3 is expressed by $\Sigma(Q^+_i)(+SA_i) - \Sigma(Q^-_i)(-SA_i)$, where $+SA_i$, $-SA_i$, Q^+_i, and Q^-_i are the partial positive surface area, partial negative surface area, partial positive charge, and partial negative charge, respectively, on the ith atom. [c] FPSA-3 is expressed by $[\Sigma(+SA_i)(Q^+_i)]/(SA_{TOT})$, where $+SA_i$ and Q^+_i are the surface area and charge contributions of the ith positively charged atom, respectively, and SA_{TOT} is the total molecular surface area. [d] FNSA-3 is expressed by $[\Sigma(-SA_i)(Q^-_i)]/(SA_{TOT})$, where $-SA_i$ and Q^-_i are the surface area and charge contributions of the ith negatively charged atom, respectively, and SA_{TOT} is the total molecular surface area. [e] RPCS-1 is expressed by $SA_{MPOS} \times [(Q^+_{MPOS})/(Q^+_{TOT})]$, where SA_{MPOS} is the surface area of the most positively charged atom, Q^+_{MPOS} is the partial charge of the most positively charged atom, and Q_{TOT} is the total positive charge of the molecule. [f] RNCS-1 is expressed by $SA_{MNEG} \times [(Q^-_{MNEG})/(Q^-_{TOT})]$, where SA_{MNEG} is the surface area of the most negatively charged atom, Q^-_{MNEG} is the partial charge of the most negatively charged atom, and Q^-_{TOT} is the total negative charge of the molecule.

phenol), all compounds removed contained two or three hydroxyl groups. This may infer that while descriptors used in ADAPT do model the viscosity of diols and triols sufficiently well, better models are possible with their exclusion or with inclusion of additional information in the form of new descriptors.

Thermal Conductivity. The best model found to predict thermal conductivity contained nine descriptors. These are listed and explained in Table 4. Descriptor T-values and multiple correlation coefficients fulfilled the specified requirements set forth above. The pairwise correlation coefficients for this model ranged from 0.592 to 0.947 with an average of 0.804. Three compounds were flagged as statistical outliers (**54**, **198**, **213**); however, minimal effects on the model coefficients and regression predictions were observed with their removal. Rms errors for the MLR model were 0.0143 W/m·K ($R = 0.904$) for the training set and 0.0136 W/m·K ($R = 0.953$) for the prediction set. Table 4 contains the model coefficients, the standard error of the coefficients, and the value ranges for the nine descriptors included in this model.

9-2-1 to 9-6-1 architectures were examined in the development of a CNN model with a 9-4-1 architecture (45 adjustable parameters) providing the best results. The training set rms error was 0.0106 W/m·K ($R = 0.951$), the cross-validation set rms error was 0.0116 W/m·K ($R = 0.919$), and the prediction set rms error was 0.0083 W/m·K ($R = 0.972$). Clearly, significant improvement over the MLR model was achieved with a 40% reduction in prediction set rms error. The average prediction error across the range of experimental values in this dataset was 5.6% for the training set, 6.2% for the cross-validation set, and 4.6% for the prediction set. A calculated versus observed plot for this CNN model is shown in Figure 3.

The nine descriptors used for predicting thermal conductivity included five CPSA descriptors, one hydrogen-bonding descriptor, and three topological descriptors. The CPSA descriptors FPSA-3 and FNSA-3 encode information about the positively charged and negatively charged surface areas, respectively, using a weighted atomic charge. The RPCS-1 and RNCS-1 CPSA descriptors contain information about the charged surface area relative to the most positively charged and negatively charged atoms, respectively. The remaining CPSA descriptor, DPSA-3, is the difference in the positively charged and negatively charged surface areas using a weighted atomic charge. KAPA-2 is the second-order path κ shape index,[22] and MDE-14 is the molecular distance edge measure[26] between primary and quaternary carbon atoms. These two topological descriptors measure the degree of branching and provide information about the shape of the molecules. Finally, 1SP3-1 is a count of the number of sp^3 carbons attached to only one other carbon atom.

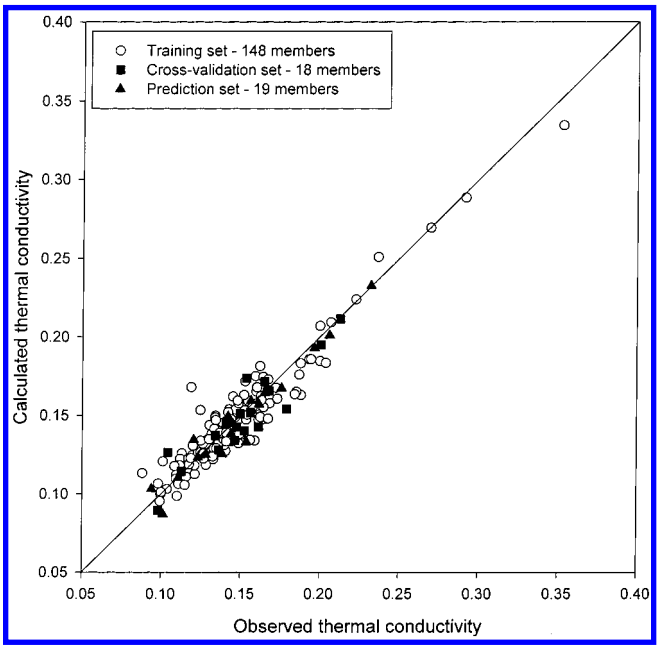

Figure 3. Plot of calculated versus observed thermal conductivity for the training, cross-validation, and prediction set compounds used to develop the CNN model described in Table 4.

As with surface tension and viscosity, attention is drawn to training set compounds which are clearly dependent variable extrema. Inspection of the plot in Figure 3 indicates that three outliers (**54**, **99**, **198**) are present; therefore, these compounds were removed and a new model was generated. Starting with a 163-member training set, a nine-descriptor model was selected as best. The MLR model had a training set rms error of 0.0123 W/m·K ($R = 0.877$) and a prediction

Table 5. Descriptors Used in the General CNN Model for Predicting Surface Tension, Viscosity, and Thermal Conductivity

descriptor	type	range	explanation[a]
V2-3	top.	0 to 6.6	valence-corrected second-order χ-index
QSUM-1	elec.	0.25 to 3.91	sum of the absolute values of all atomic charges
SADH-3	hyb.	0 to 0.458	(total SA of all donable H)/(total molecular SA)
PPSA-1	hyb.	0 to 624	total partial positive SA
DPSA-3	hyb.	11.3 to 76.3	difference in atomic charge weighted partial SA[b]
NLP-19	top.	0 to 8	number of lone pairs of electrons
WTPT-2	top.	1.5 to 2.07	molecular ID/total number of atoms
FNSA-3	hyb.	−0.279 to 0	(atomic charge weighted partial negative SA)/(total molecular SA)[c]
SCAA-3	hyb.	−0.176 to 0	(sum of (Q*SA) for all acceptor atoms)/(total molecular SA)[d]

[a] SA = surface area, Q = charge. [b] DPSA-3 is expressed by $\Sigma(Q^+_i)(+SA_i) - \Sigma(Q^-_i)(-SA_i)$, where $+SA_i$, $-SA_i$, Q^+_i, and Q^-_i are the partial positive surface area, partial negative surface area, partial positive charge, and partial negative charge, respectively, on the ith atom. [c] FNSA-3 is expressed by $[\Sigma(-SA_i)(Q^-_i)]/(SA_{TOT})$, where $-SA_i$ and Q^-_i are the surface area and charge contributions of the ith negatively charged atom, respectively, and SA_{TOT} is the total molecular surface area. [d] Acceptor atoms are N, O, S, and F.

set rms error of 0.0100 W/m·K (R = 0.959). The CNN model, using a 9-6-1 architecture, showed improved results over the MLR model as well as the original model which contained the three outliers. For this model, the training set rms error was 0.0073 W/m·K (R = 0.959), the cross-validation set rms error was 0.0074 W/m·K (R = 0.966), and for the prediction set the rms error was 0.0083 W/m·K (R = 0.972). This demonstrates that removal of the three compounds with the highest thermal conductivity can allow the development of a better model for the remaining members of the data set.

General CNN Model. The similarity among the descriptors included in all three physical property models, especially the large number of CPSA descriptors, prompted an experiment to find a general predictive model for all three properties using a common set of descriptors. The goal of this experiment was to modify our feature selection routine which utilizes simulated annealing with an MLR fitness evaluator to search the descriptor space for subsets which were information-rich across all three solvent properties. This introduces a new method capability of the ADAPT software package. A modified cost function was used to ensure that all three properties contributed equally to the overall assessment of the descriptor subsets evaluated. Then, the descriptors from the subset which possessed the lowest cost function were submitted to CNNs and individually trained for each property using a 9-X-1 architecture. As with the independent property studies, all architectures for the CNNs were evaluated within the limit of the training set observations to adjustable parameters ratio.

Experimental data for 179 of the solvents from ref 15 was available for all three properties. This was split into a 145-member training set, an 18-member cross-validation set, and a 17-member prediction set. The descriptor pool was reduced to 94-members using objective feature selection. The reduced pool was then submitted to the modified linear feature selection routine to find the best subsets of descriptors. The cost function (eq 1) used to direct the search was a scaled combination of rms errors for the three properties.

$$\text{cost} = 0.9*\frac{\sigma_{rms}}{\sigma_{range}} + 0.9*\frac{\eta_{rms}}{\eta_{range}} + 1.28*\frac{\lambda_{rms}}{\lambda_{range}} \quad (1)$$

The dependent variable ranges for each property were given in the Experimental Section. The coefficients were determined experimentally by averaging the (property rms error/property dependent variable range) for each property over three search trials. The coefficients chosen gave reproducibly consistent results which ensured that the contribution of each of the three properties to the overall cost function was essentially one-third. Therefore, no bias toward one particular property would be observed.

Searches for the best eight- and nine-descriptor models were performed. The best subset from each model size was then trained for each property with a CNN and evaluated based on the training set and cross-validation set rms errors. A nine-descriptor model produced the best overall results for the three properties. Table 5 lists these nine descriptors along with the range of values for each and an explanation. A 9-4-1 architecture was chosen for modeling surface tension. The training set rms error was 2.48 mN/m (R = 0.958), the cross-validation set rms was 2.48 mN/m (R = 0.897), and the prediction set rms error was 2.89 mN/m (R = 0.877). For viscosity, a 9-5-1 architecture was determined to be optimal with a training set rms error of 0.123 mPa·s (R = 0.982), a cross-validation set rms error of 0.125 mPa·s (R = 0.971), and a prediction set rms error of 0.150 mPa·s (R = 0.860). Finally, thermal conductivity required a 9-4-1 architecture to achieve the best results. The rms errors for this network were 0.0117 W/m·K (R = 0.945) for the training set, 0.0112 W/m·K (R = 0.935) for the cross-validation set, and 0.0236 W/m·K (R = 0.860) for the prediction set. Removal of one prediction set outlier (**193**) which accounted for 57% of the variance in the rms error lowered its value to 0.0160 W/m·K (R = 0.776). Figures 4−6 show the calculated versus observed plots for each of the three properties using the general model.

As expected, a large number of hybrid descriptors were present in the model. In particular, the hydrogen-bonding specific descriptor, SADH-3, and the CPSA descriptor, FNSA-3, were present in all three individual models as well as the general model. In addition, the CPSA descriptors, PPSA-1 and DPSA-3, and the hydrogen bonding descriptor, SCAA-3, were present in the general model. The observation that CPSA and hydrogen-bonding descriptors comprise a majority of the descriptors used in these models lends evidence to the chemical nature of the properties studied. It may be possible to infer from these results that the surface and transport phenomena of solvents are largely dependent upon accessible surface area and the localized partial charges of the molecules. In addition to the CPSA descriptors, one electronic descriptor and three topological descriptors were present. QSUM-1 is the absolute sum of all partial charges

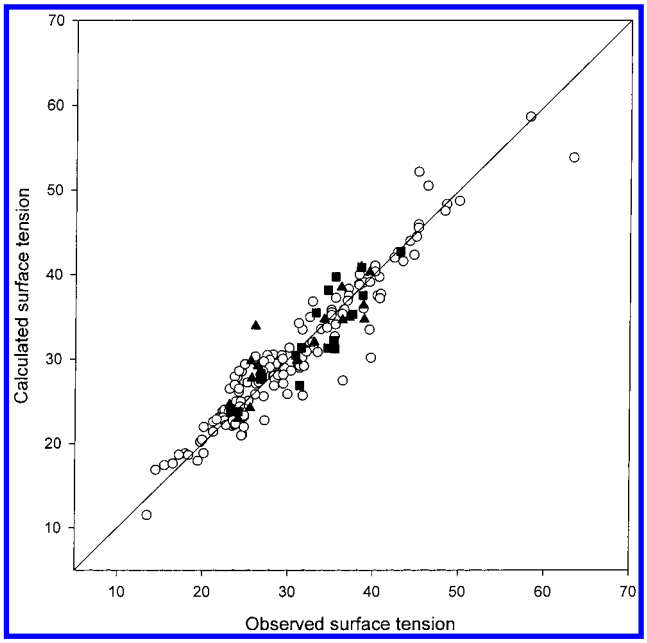

Figure 4. Plot of calculated versus observed surface tension for the training, cross-validation, and prediction set compounds used to develop the general CNN model described in Table 5.

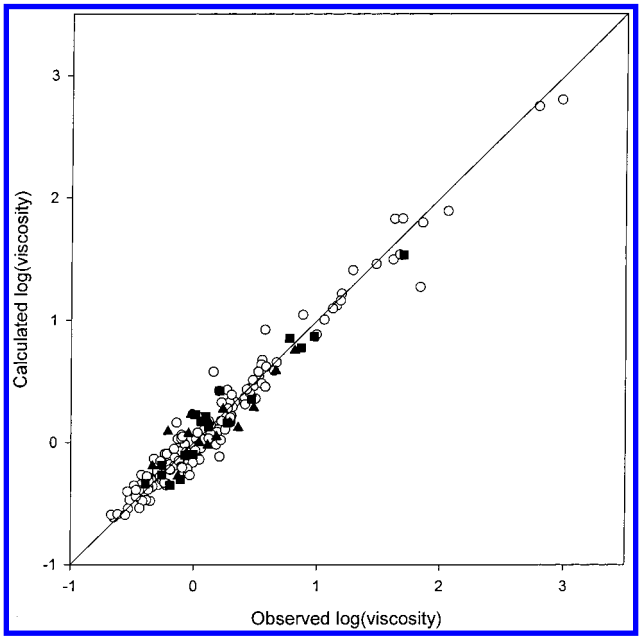

Figure 5. Plot of calculated versus observed log(viscosity) for the training, cross-validation, and prediction set compounds used to develop the general CNN model described in Table 5.

in each molecule, again implying that charge information is important for accurate predictions. V2-3 is the valence-corrected second-order χ index[24] which encodes branching information. NLP-19 is a constitutional descriptor which is a simple count of lone pairs of electrons. Finally, WTPT-2 is a weighted path descriptor which encodes information about the molecular connectivity of the solvent compounds.[25]

Overall, the rms errors of the three solvent properties using the general model compare quite well to the models developed for the individual properties. Table 6 compares the results of the individual models to those of the general model. The results for surface tension show that comparable rms errors were obtained by both models, with a slight advantage for the individual property model. All viscosity

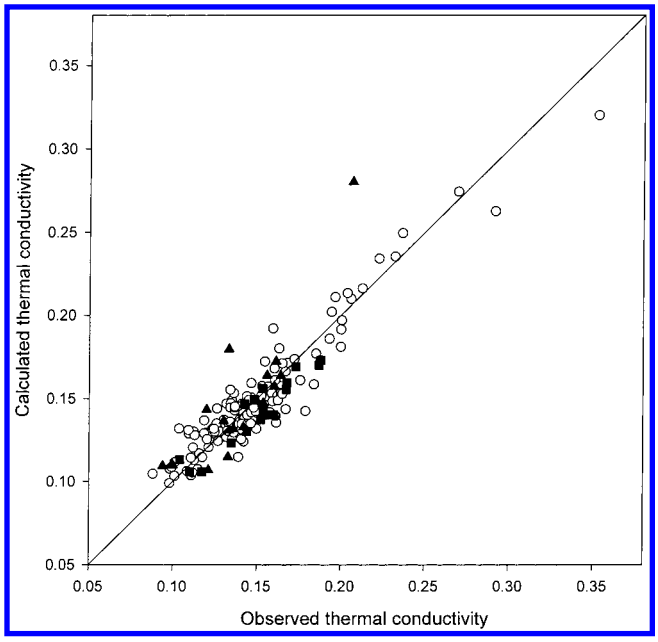

Figure 6. Plot of calculated versus observed thermal conductivity for the training, cross-validation, and prediction set compounds used to develop the general CNN model described in Table 5.

results clearly improved using the general model. This is most likely explained by a simplification in the data set diversity after reducing the original 212-solvent data set to the present 179-solvent dataset. Finally, for thermal conductivity the results are mixed. The training set and cross-validation rms errors were generally consistent with the individual property model; however, the prediction set rms error was substantially higher than the corresponding error from the individual property model. This was mainly due to the compound outlier (**193**).

Monte Carlo Models. One potential shortcoming of QSPR methods is the risk of chance correlations between the structures and the property being studied. To ensure that the models developed in this QSPR study were not due to chance effects, randomization experiments were conducted. For these experiments, the dependent variable was scrambled for each property such that each structure was assigned an experimental value of another structure in the data set. Feature selection and model development were then performed as described in the Experimental Section for the MLR and CNN models. In all cases, the number of descriptors employed in the standard QSPR models was used for the Monte Carlo models. The network architectures used for the CNN models were also preserved. In addition, the rms errors reported for the CNN models in this experiment were averaged over 10 trainings to minimize the effect of runs producing extreme errors.

For the MLR model of surface tension, the training set rms error was 7.624 mN/m ($R = 0.473$) and the prediction set rms error was 9.085 mN/m ($R = 0.276$). The CNN model gave an average training set rms error of 6.303 mN/m ($R = 0.703$), cross-validation set rms error of 7.645 mN/m ($R = 0.366$), and prediction set rms error of 9.204 mN/m ($R = 0.251$). For viscosity, the MLR model gave rms errors of 0.557 mPa·s ($R = 0.391$) for the training set and 0.656 mPa·s ($R = 0.272$) for the prediction set. The CNN model gave average rms errors of 0.431 mPa·s ($R = 0.685$) for the training set, 0.520 mPa·s ($R = 0.772$) for the cross-validation

Table 6. Rms Error Comparison between Individual Models and General Model

	surface tension			viscosity			thermal conductivity		
model	tset	cvset	pset	tset	cvset	pset	tset	cvset	pset
individual	2.222	2.224	2.757	0.147	0.148	0.242	0.0106	0.0116	0.0083
general	2.475	2.476	2.886	0.123	0.125	0.150	0.0117	0.0112	0.0236
% diff.[a]	10.2	10.2	4.5	−16.3	−15.5	−38.0	9.4	−3.4	64.8

[a] Calculated as (individual model rms error − general model rms error)/(individual model rms error).

set, and 0.782 mPa·s ($R = 0.100$) for the prediction set. Finally, rms errors from the prediction of thermal conductivity by an MLR model were 0.0306 W/m·K ($R = 0.437$) for the training set and 0.0354 W/m·K ($R = 0.259$) for the prediction set. The CNN model generated a training set with an rms error of 0.0192 W/m·K ($R = 0.775$), a cross-validation set with an rms error of 0.0555 W/m·K ($R = 0.310$), and a prediction set with an rms error of 0.0406 W/m·K ($R = 0.045$). The rms errors and correlation coefficients presented here clearly indicate that randomly scrambling the dependent variable has an adverse effect on the quality of QSPR that can be developed. This adds strength to the models presented in this paper as sound links between the molecular structure of common organic solvents and their experimental surface tension, viscosity, and thermal conductivity values.

CONCLUSIONS

In this paper, models using MLR analysis and CNNs are reported for the prediction of three related surface and transport properties of 213 common organic solvents. First, models for the individual properties were developed followed by the investigation of removing experimental value extrema from training and its effect on model quality. Then, as a new feature in the ADAPT software package, the algorithm for descriptor feature selection using simulated annealing with an MLR fitness evaluator was modified to find a general model for predicting all three properties from one subset of descriptors. These results were comparable in quality to those of the individual models for each property. Generally, the models developed in this study were rich in CPSA and hydrogen-bonding descriptors. This demonstrates that a common chemical theme exists among the three properties.

The results for surface tension obtained in this study are difficult to compare to the two previous QSPRs reported. While the present data set is equivalent in size to that used by Stanton and Jurs, the diversity is much greater. On the other hand, the model published by Kavun and co-workers was developed using a more diverse, yet smaller data set. It can be confidently stated, however that the prediction accuracy for the solvents investigated in this study were within the limits of experimental error. The model shown here clearly presents the most comprehensive and diverse QSPR model for surface tension in the literature.

The results for viscosity compare favorably to the four models discussed earlier. While the CNN model developed in this work produced average percent errors at the higher end of experimental errors, it showed improved rms errors and correlation coefficients over previously reported linear models. It should be noted, however, that the data sets employed for those studies were larger thus leading to potentially greater diversity requiring more general models. The results from the CNN model developed by Suzuki et al. are very comparable to the CNN model results obtained in our work. The only distinct advantage is that the descriptors employed in the present work are theoretically based on molecular structure alone and not on physical properties of the compounds involved. The results obtained for thermal conductivity mark the first QSPR reported for this physical property and were also within the limits of error of the experimental data.

REFERENCES AND NOTES

(1) Ash, I.; Ash, M. *The Index of Solvents*; Gower Publishing: England, 1996.

(2) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.

(3) Stanton, D. T.; Jurs, P. C. Computer-Assisted Study of the Relationship between Molecular Structure and Surface Tension of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109−115.

(4) Jurs, P. C.; Chou, J. T.; Yuan, M. In Olsen, E. C., Christoffersen, R. E., Eds.; Americal Chemical Society: Washington, DC, 1979.

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

(6) Kavun, S. M.; Chalykh, A. E.; Palyulin, V. A. Prediction of Surface Tension of Organic Liquids. *Colloid J.* **1995**, *57*, 767−771.

(7) Gupta, S. P. QSAR Studies on Drugs Acting at the Central Nervous System. *Chem. Rev.* **1989**, *89*, 1765−1800.

(8) Suzuki, T.; Ebert, R.; Shuurmann, G. Development of Both Linear and Nonlinear Methods to Predict the Liquid Viscosity at 20 °C of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1122−1128.

(9) Schuurmann, G.; Kuhne, R.; Kleint, F.; Ebert, R.-U.; Rothenbacher, C.; Herth, P. A. In *A Software System for Automatic Chemical Property Estimation from Molecular Structure*; Chen, F., Schuurmann, G., Eds.; SETAC Press: Pensacola, 1997.

(10) Ivanciuc, O.; Ivanciuc, T.; Filip, P.; Cabrol-Bass, D. Estimation of the Liquid Viscosity of Organic Compounds with a Quantitative Structure−Property Model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 515−524.

(11) Katritzky, A. R.; Chen, K.; Wang, Y.; Karelson, M.; Lucic, B.; Trinajstic, N.; Suzuki, T.; Schuurmann, G. Prediction of Liquid Viscosity for Organic Compounds by a Quantitative Structure−Property Relationship. *J. Phys. Org. Chem.* **2000**, *13*, 80−86.

(12) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA Version 2.0 Reference Manual*; 1994.

(13) Cocchi, M.; De Benedetti, P. G.; Seeber, R.; Tassi, L.; Ulrici, A. Development of Quantitative Structure−Property Relationships Using Calculated Descriptors for the Prediction of the Physiochemical Properties of a Series of Organic Solvents. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1190−1203.

(14) Marcus, Y. *The Properties of Solvents*; John Wiley and Sons: West Sussex, 1998.

(15) DIPPR. *The DIPPR Pure Component Data Compilation*, version 12.4 (for Windows); Technical Database Services, Inc.: New York, 1997.

(16) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.

(17) Xu, L.; 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.

(18) Stewart, J. P. P. Mopac: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.

(19) Stewart, J. P. P., MOPAC 6.0, Quantum Chemistry Program Exchange; Indiana University: Bloomington, IN, Program 455.
(20) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(21) Aleman, C.; Luque, F. J.; Orozco, M. Suitability of the PM3-Derived Molecular Electrostatic Potentials. *J. Comput. Chem.* **1993**, *14*, 799−808.
(22) Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.
(23) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat.* **1986**, *5*, 7−12.
(24) Hall, L. H.; Kier, L. B. *The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Modeling*; VCH Publishers: New York, 1991.
(25) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for All Self-Avoiding Paths for Molecular Graphs. *Computers Chem.* **1979**, *3*, 5−13.
(26) 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*, 8, 387−394.
(27) Pearlman, R. S. In Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980.
(28) 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.
(29) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure−Property Relationships. *J. Comput. Chem.* **1991**, *13*, 492−504.
(30) Dixon, S. L.; Jurs, P. C. Estimation of pK_a for Organic Oxyacids Using Calculated Atomic Charges. *J. Comput. Chem.* **1993**, *14*, 1460−1467.
(31) 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.
(32) Vinogradov, S. N.; Linnell, R. H. *Hydrogen Bonding*; van Nostrand Reinhold: New York, 1971.
(33) Pimentel, G. I.; McClellan, A. L. *The Hydrogen Bond*; Freeman: San Francisco, 1960.
(34) Adjustable parameters are calculated based on the number of neurons used in each layer of the architecture as (input × hidden) + (hidden × output) + hidden + output.
(35) 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.
(36) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.

CI000139T