Prediction of Infinite Dilution Activity Coefficients of Organic Compounds in Aqueous Solution from Molecular Structure

Brooke E. Mitchell and Peter C. Jurs*

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

Received October 24, 1997

A quantitative structure−property relationship study is performed to develop models that relate the structures of a heterogeneous group of organic compounds to their infinite dilution activity coefficients, γ^{∞}. The molecular structures are represented by calculated descriptors that encode their topological, electronic, and geometric features. The descriptors are used to develop multiple linear regression and computational neural network models to predict the γ^{∞}. Genetic algorithm and simulated annealing routines are used to select subsets of descriptors that form the best models. The models that are developed have predictive ability in the range of the experimental error of infinite dilution activity coefficient measurements.

INTRODUCTION

Physical property and thermodynamic data for organic chemicals are needed by industrial chemical engineers.[1,2] The infinite dilution activity coefficient (γ^{∞}) is an important parameter related to the study of the thermodynamic behavior of dilute solutions. Because the most common industrial solvent is water, the study of γ^{∞} of aqueous solutions is of great interest. The parameter γ^{∞} is related to the phase equilibrium of aqueous solutions and gives a measure of how different the solvent environment of water is from a pure solute environment. Because of the unique characteristics of water, including hydrogen bonding, aqueous solutions exhibit strong nonideality.[3] Because it is measured at infinite dilution, γ^{∞} provides a measure of intermolecular solute−solvent interactions without the interference of solute−solute interactions. The magnitude of γ^{∞} provides insight into the chemical and physical forces that exist between solute and solvent molecules.[4]

Knowledge of phase equilibrium behavior is industrially important. It plays a role in all separation processes. Some of the areas affected by phase equilibria are in petroleum processing, distillation apparatus, chromatographic systems, waste treatment, and environmental cleanup.[1,2,4−6] All of these areas often involve aqueous solutions.

Many of the techniques used to measure γ^{∞} are applicable in only a small volatility range, so different techniques are necessary to make measurements for samples that cover a wide range of solution volatility. Some of the direct measurement techniques that are used include dynamic gas chromatography,[7,8] differential ebulliometry,[9,10] gas stripping,[6,11,12] differential static equilibrium cell,[13] and headspace chromatography.[14] Values of γ^{∞} are also calculated indirectly from measurements of aqueous solubility, liquid−liquid chromatography, and partition coefficients.[15−18] The most common of these indirect methods is the use of the aqueous solubility value. When the solute is at its solubility limit in water, $\gamma^{\infty} = 1/X_2$, where X_2 is the mole fraction of the solute in an aqueous solution.

For reasons of cost, time, safety, and availability of chemical samples, it is useful to be able to predict γ^{∞} rather than measure it.[5] There are a variety of predictive tools available, including group contribution methods, such as ASOG[19,20] and UNIFAC,[21,22] linear solvation energy relationships (LSER),[23] free energy perturbation simulations,[24] and quantitative structure−property relationships (QSPR). The free energy perturbation method with Monte Carlo simulations is very computationally expensive and time-consuming. LSER have only fairly recently been applied to γ^{∞} and depend on the availability of experimental values for the solvatochromic descriptors used. Predictions from group contribution methods are highly dependent on the quality of the structural parameters that are developed from experimental data. Higher quality parameters are developed from a large, diverse data set with accurate experimentally measured values.

Some correlations between structural features and γ^{∞} have been reported, including relationships with the molecular surface area,[25,26] molar refraction,[27] number of carbons,[26,28] molecular connectivity,[26] electronic energy,[26] and the dipole moment.[26] In all of these cases, the correlations were studied between one or a small number of structural feature descriptors and γ^{∞}.

In this work, the best subset of structural descriptors is chosen from a large pool of calculated descriptors for the development of a QSPR model for a diverse set of organic compounds. Both multiple linear regression and computational neural networks (CNN) have been used to develop the models.

EXPERIMENTAL SECTION

This QSPR study was performed using the Automated Data Analysis and Pattern Recognition Toolkit (ADAPT)[29,30] software as well as genetic algorithm,[31] simulated annealing,[32,33] and computational neural network[34] routines. All computations were performed on a DEC 3000AXP model 500 workstation at Penn State University.

The set of compounds used in this study was taken from a compilation of data by Sherman et al.[23] The data in this

S0095-2338(97)00092-9 CCC: $15.00
Published on Web 02/11/1998

Table 1. List of Compounds, Their Infinite Dilution Activity Coefficients (γ^{∞}), the Experimental and Calculated ln (γ^{∞}),[a] and References

cmpd. no.	compound name	expt γ^{∞}	expt ln (γ^{∞})	calc ln (γ^{∞})	ref	cmpd. no.	compound name	expt γ^{∞}	expt ln (γ^{∞})	calc ln (γ^{∞})	ref
1	tetrachloromethane	1.54E+04	9.64	8.42	15	**73**	*n*-propyl acetate[b]	2.42E+02	5.49	5.65	17
2	trichloromethane	9.03E+02	6.81	6.63	6	**74**	*n*-butyl acetate	8.14E+02	6.70	6.99	17
3	tribromomethane	3.38E+03	8.13	8.09	13	**75**	*n*-pentyl acetate	3.23E+03	8.08	8.30	17
4	dichloromethane	2.53E+02	5.53	5.76	6	**76**	*n*-hexyl acetate	1.25E+04	9.43	9.51	17
5	dibromomethane	8.50E+02	6.75	6.82	13	**77**	isopropyl acetate	1.96E+02	5.28	5.38	17
6	tetrachloroethene	3.60E+04	10.5	10.5	6	**78**	isobutyl acetate	8.44E+02	6.74	6.77	17
7	1,1,1,2-tetrachloroethane	8.91E+03	9.10	9.08	13	**79**	isopentyl acetate	2.98E+03	8.00	8.13	17
8	1,1,2,2-tetrachloroethane	3.46E+03	8.15	8.83	15	**80**	methyl propanoate[b]	8.71E+01	4.47	4.47	23
9	trichloroethene	8.75E+03	9.08	8.58	6	**81**	ethyl propanoate	2.56E+02	5.55	5.40	17
10	1,1,1-trichloroethane	5.90E+03	8.68	7.71	29	**82**	*n*-propyl propanoate	1.09E+03	6.99	6.61	32
11	1,1,2-trichloroethane	1.50E+03	7.31	7.68	6	**83**	methyl butyrate	3.31E+02	5.80	5.84	23
12	*cis*-1,2-dichloroethene[c]	8.70E+02	6.77	7.39	13	**84**	ethyl butyrate	7.30E+02	6.59	6.68	17
13	*trans*-1,2-dichloroethene	1.26E+03	7.14	7.25	13	**85**	methyl pentanoate	1.26E+03	7.14	7.25	23
14	1,1-dichloroethane	1.08E+03	6.99	6.70	29	**86**	methyl hexanoate	3.98E+03	8.29	8.59	23
15	1,2-dichloroethane	6.41E+02	6.46	6.61	6	**87**	nitromethane	3.16E+01	3.45	3.58	17
16	iodoethane	2.19E+03	7.69	7.68	6	**88**	nitroethane	8.86E+01	4.48	4.56	17
17	bromoethane	6.79E+02	6.52	7.04	6	**89**	1-nitropropane	2.99E+02	5.70	5.82	17
18	1,3-dichloropropene[c]	1.40E+03	7.24	8.49	13	**90**	3-nitrotoluene	7.09E+03	8.87	9.10	15
19	1-chloropropane[c]	1.75E+03	7.47	7.29	6	**91**	diethylamine	5.40E+00	1.69	1.48	17
20	1-bromopropane	2.86E+03	7.96	8.03	6	**92**	triethylamine	6.75E+01	4.21	4.52	30
21	1-iodopropane	8.55E+03	9.05	8.71	6	**93**	*N*,*N*-dimethylformamide	8.30E-01	−0.186	−0.251	17
22	2-chloropropane	1.48E+03	7.30	7.25	6	**94**	*N*,*N*-dimethylacetamide	1.04E+00	0.0392	0.736	17
23	2-bromopropane	2.09E+03	7.65	8.25	6	**95**	acetonitrile[b]	1.11E+01	2.41	3.27	17
24	1,2-dichloropropane	2.33E+03	7.75	7.87	13	**96**	propionitrile	3.53E+01	3.56	3.72	17
25	1-chlorobutane	7.61E+03	8.94	8.42	6	**97**	butyronitrile	1.18E+02	4.77	4.62	17
26	1-bromobutane[b]	1.22E+04	9.41	9.16	6	**98**	isobutyronitrile	1.17E+02	4.76	4.21	17
27	2-bromobutane	8.32E+03	9.03	9.55	6	**99**	pentanenitrile	4.02E+02	6.00	5.77	17
28	1-chloropentane[c]	3.21E+04	10.4	9.72	6	**100**	hexanenitrile	1.40E+03	7.24	7.14	17
29	1-chlorohexane	1.41E+05	11.9	11.1	6	**101**	benzonitrile[d]	1.74E+03	7.46		17
30	fluorobenzene[b]	4.80E+03	8.48	9.22	15	**102**	pyridine	1.99E+01	2.99	3.64	17
31	chlorobenzene	1.40E+04	9.55	9.33	15	**103**	4-methylpyridine[c]	4.23E+01	3.75	4.04	23
32	bromobenzene[c]	2.25E+04	10.0	9.53	15	**104**	3-methylpyridine	4.91E+01	3.89	4.17	23
33	iodobenzene	5.41E+04	10.9	11.0	15	**105**	benzene	2.48E+03	7.82	8.04	6
34	1,2-dichlorobenzene	6.82E+04	11.1	11.0	15	**106**	toluene	9.19E+03	9.13	9.27	6
35	benzyl chloride	3.20E+04	10.4	10.3	15	**107**	ethylbenzene	3.27E+04	10.4	10.6	6
36	methanol	1.58E+00	0.457	0.357	17	**108**	*n*-propylbenzene	1.36E+05	11.8	11.9	6
37	ethanol	3.74E+00	1.32	1.13	17	**109**	*n*-butylbenzene	5.66E+05	13.3	13.3	6
38	1-propanol	1.34E+01	2.60	2.39	17	**110**	*o*-xylene	3.05E+04	10.3	10.3	6
39	2-propanol[c]	7.60E+00	2.03	2.01	17	**111**	*m*-xylene	3.32E+04	10.4	10.5	6
40	2-methyl-1-propanol	4.89E+01	3.89	3.67	17	**112**	*p*-xylene	3.33E+04	10.4	10.5	6
41	1-butanol	5.02E+01	3.92	3.94	17	**113**	isopropylbenzene	1.02E+05	11.5	11.7	6
42	2-butanol	2.62E+01	3.27	2.95	23	**114**	1,3,5-trimethylbenzene[c]	1.17E+05	11.7	11.8	6
43	*tert*-butanol	1.19E+01	2.48	2.99	17	**115**	tetrahydrofuran	1.70E+01	2.83	2.93	17
44	1-pentanol	1.98E+02	5.29	5.58	17	**116**	tetrahydropyran	7.86E+01	4.36	4.35	17
45	2-pentanol	9.69E+01	4.57	4.45	23	**117**	1,4-dioxane	5.42E+00	1.69	1.25	17
46	3-methyl-1-butanol	2.08E+02	5.34	5.06	17	**118**	methoxybenzene	3.65E+03	8.20	7.68	15
47	1-hexanol[c]	7.99E+02	6.68	7.20	17	**119**	ethoxybenzene	1.57E+04	9.66	9.05	15
48	2-hexanol	2.82E+02	5.64	5.97	23	**120**	diethyl ether	6.88E+01	4.23	4.41	17
49	cyclohexanol	1.57E+02	5.06	5.02	17	**121**	di-*n*-propyl ether[b]	2.31E+03	7.75	7.47	6
50	1-heptanol	3.27E+03	8.09	8.67	23	**122**	di-*n*-butyl ether	4.72E+04	10.8	10.4	6
51	formaldehyde	2.80E+00	1.03	1.24	23	**123**	*tert*-butyl methyl ether	1.13E+02	4.73	4.94	17
52	acetaldehyde	3.94E+00	1.37	1.54	23	**124**	diisopropyl ether	6.28E+02	6.44	6.62	6
53	propionaldehyde[c]	1.30E+01	2.57	2.47	17	**125**	2,2,2-trifluoroethanol	8.65E+00	2.16	2.09	17
54	butyraldehyde	4.86E+01	3.88	3.72	17	**126**	ethylene oxide	6.23E+00	1.83	2.24	31
55	pentanal	2.20E+02	5.39	5.13	17	**127**	formic acid	7.20E-01	−0.329	0.0200	23
56	hexanal	8.13E+02	6.70	6.60	17	**128**	acetic acid[c]	9.20E-01	−0.0834	0.593	17
57	octanal	8.24E+03	9.02	9.46	15	**129**	methyl isobutyrate	3.09E+02	5.73	5.51	23
58	acetone[c]	7.01E+00	1.95	2.25	17	**130**	dimethyl sulfoxide	9.00E-02	−2.41	−2.58	23
59	2-butanone	2.56E+01	3.24	3.33	17	**131**	*n*-methyl-2-pyrrolidone	3.70E-01	−0.994	−0.232	23
60	2-pentanone	9.38E+01	4.54	4.71	17	**132**	trichlorofluoromethane	7.06E+03	8.86	8.64	1
61	cyclopentanone	2.92E+01	3.37	2.70	17	**133**	dichlorodifluoromethane	2.24E+04	10.0	9.77	1
62	3-pentanone	1.07E+02	4.67	4.49	17	**134**	chlorotrifluoromethane	6.44E+04	11.1	11.2	1
63	3-methyl-2-butanone	8.40E+01	4.43	4.38	17	**135**	tetrafluoromethane	3.05E+05	12.6	12.6	1
64	2-hexanone	3.56E+02	5.88	6.15	17	**136**	nitrotrichloromethane	5.63E+03	8.64	8.32	32
65	3-hexanone[b]	4.12E+02	6.02	5.90	23	**137**	dichlorofluoromethane	3.05E+02	5.72	6.91	1
66	2-heptanone	1.40E+03	7.24	7.62	17	**138**	chlorodifluoromethane[c]	1.73E+03	7.46	7.69	1
67	2-nonanone	1.63E+04	9.70	10.4	15	**139**	trifluoromethane	4.32E+03	8.37	8.34	1
68	methyl formate	1.55E+01	2.74	2.86	17	**140**	triiodomethane[d]	2.19E+05	12.3		1
69	ethyl formate	4.73E+01	3.86	3.94	17	**141**	difluoromethane	6.59E+02	6.49	6.01	1
70	*n*-propyl formate	1.69E+02	5.13	5.34	17	**142**	diiodomethane	1.20E+04	9.39	9.11	1
71	methyl acetate	2.26E+01	3.12	3.33	17	**143**	bromomethane	3.94E+02	5.98	5.93	1
72	ethyl acetate	6.53E+01	4.18	4.33	17	**144**	chloromethane	4.76E+02	6.17	5.75	1

Table 1 (Continued)

cmpd. no.	compound name	expt γ^{∞}	expt ln (γ^{∞})	calc ln (γ^{∞})	ref	cmpd. no.	compound name	expt γ^{∞}	expt ln (γ^{∞})	calc ln (γ^{∞})	ref
145	fluoromethane	7.91E+02	6.67	6.64	1	**217**	3,3-dimethyl-2-butanone	2.88E+02	5.66	5.42	32
146	hexachloroethane	1.64E+06	14.3	14.0	1	**218**	3-methyl-2-pentanone	2.61E+02	5.57	5.64	32
147	1,1,2-trichlorotrifluoroethane	6.12E+04	11.0	11.5	1	**219**	4-methyl-2-pentanone	2.92E+02	5.68	5.88	32
148	1,2-dichlorotetrafluoroethane[b]	6.93E+04	11.2	11.3	1	**220**	2-methyl-3-pentanone	3.61E+02	5.89	5.58	32
149	chloropentafluoroethane	1.48E+05	11.9	11.5	1	**221**	4-heptanone	1.65E+03	7.41	7.27	32
150	tetrafluoroethene	3.51E+04	10.5	10.6	1	**222**	2,4-dimethyl-3-pentanone[c]	1.11E+03	7.01	6.55	32
151	hexafluoroethane[d]	9.70E+05	13.8		1	**223**	5-methyl-2-hexanone	1.52E+03	7.33	7.32	33
152	pentachloroethane	2.39E+04	10.1	10.9	32	**224**	acetophenone	9.76E+02	6.88	7.01	32
153	1,1,2,2-tetrabromoethane	2.95E+04	10.3	10.4	32	**225**	2,6-dimethyl-4-heptanone[b]	8.76E+03	9.08	9.31	32
154	chloroethene	1.29E+03	7.16	6.19	1	**226**	5-nonanone	2.17E+04	9.99	9.96	32
155	1-bromo-2-chloroethane	1.16E+03	7.06	7.50	32	**227**	hypochlorous acid *tert*-butyl ester	1.88E+03	7.54	7.31	32
156	1,2-dibromoethane	2.53E+03	7.84	7.71	32	**228**	ethyl propenoate	2.71E+02	5.60	5.12	32
157	chloroethane	3.97E+02	5.98	6.47	1	**229**	1-ethenyl ethyl acetate	4.74E+02	6.16	5.66	32
158	3-bromo-1-propene[b]	1.76E+03	7.47	8.05	1	**230**	isopropyl butyrate	3.08E+03	8.03	7.52	32
159	3-chloro-1-propene	1.06E+03	6.97	6.95	1	**231**	ethyl pentanoate	2.88E+03	7.97	7.91	32
160	1,2,3-trichloropropane[b]	4.31E+03	8.37	8.66	1	**232**	cyclohexyl acetate[c]	2.72E+03	7.91	7.53	32
161	1,2-dibromopropane	7.84E+03	8.97	9.14	1	**233**	butyl pentanoate	1.94E+04	9.87	9.97	32
162	1,3-dibromopropane	6.67E+03	8.81	8.67	32	**234**	methyl propyl ether	1.32E+02	4.88	4.61	32
163	1,3-dichloropropane	2.30E+03	7.74	7.60	32	**235**	butyl methyl ether[c]	5.46E+02	6.30	6.25	32
164	2-iodopropane	6.74E+03	8.82	9.20	1	**236**	*sec*-butyl methyl ether	3.02E+02	5.71	5.54	32
165	octafluorocyclobutane	2.22E+05	12.3	12.3	1	**237**	ethyl isopropyl ether	2.00E+02	5.30	5.59	32
166	2-chlorobutane[c]	5.14E+03	8.55	8.53	1	**238**	ethyl propyl ether	2.57E+02	5.55	5.96	32
167	1-chloro-2-methylpropane	5.56E+03	8.62	8.27	1	**239**	isobutyl methyl ether	4.42E+02	6.09	5.68	32
168	1-bromopentane	6.60E+04	11.1	10.3	1	**240**	2-methyl *sec*-butyl methyl ether	4.49E+02	6.11	6.12	32
169	2-chloro-2-methylbutane	1.78E+03	7.48	9.56	1	**241**	isopropyl propyl ether	1.20E+03	7.09	7.23	32
170	hexachlorobenzene	3.36E+09	21.9	21.8	1	**242**	pentanoic acid	1.27E+02	4.84	4.51	32
171	1,3-dichlorobenzene	6.63E+04	11.1	11.6	1	**243**	hexanoic acid	6.02E+02	6.40	5.94	32
172	1,4-dichlorobenzene	1.02E+05	11.5	11.6	1	**244**	benzoic acid[d]	2.00E+03	7.60		1
173	2-methyl-1-butanol	1.61E+02	5.08	5.11	32	**245**	2-nitropropane	2.97E+02	5.69	5.44	1
174	2,2-dimethyl-1-propanol	1.36E+02	4.91	4.83	32	**246**	nitrobenzene	3.53E+03	8.17	8.31	32
175	1-hexene-3-ol[b]	2.16E+02	5.38	4.85	32	**247**	2-nitrotoluene	1.17E+04	9.37	8.93	32
176	4-hexene-1-ol	1.41E+02	4.95	4.57	32	**248**	2-nitro-1-methoxybenzene	5.03E+03	8.52	8.55	32
177	2-methyl-4-pentene-3-ol	1.77E+02	5.18	4.97	32	**249**	aniline[b]	1.47E+02	4.99	5.43	32
178	2,2-dimethyl-1-butanol	7.41E+02	6.61	6.14	32	**250**	butylethylamine	1.30E+02	4.87	4.92	32
179	2,3-dimethyl-2-butanol	1.31E+02	4.88	5.10	32	**251**	dipropylamine[c]	1.34E+02	4.90	4.82	32
180	3,3-dimethyl-2-butanol	2.29E+02	5.43	5.01	32	**252**	2-aminotoluene	3.67E+02	5.91	6.22	32
181	3-hexanol	3.48E+02	5.85	5.88	32	**253**	1-ethylpiperidine	1.27E+02	4.84	4.67	32
182	*m*-cresol	2.76E+02	5.62	5.83	1	**254**	1-propylpiperidine	9.43E+02	6.85	6.56	32
183	*o*-cresol	2.46E+02	5.51	5.68	1	**255**	methanethiol	1.12E+02	4.72	4.82	1
184	*p*-cresol	3.10E+02	5.74	5.82	1	**256**	dimethyl sulfide[c]	1.77E+02	5.18	5.11	1
185	2-methyl-2-pentanol	1.70E+02	5.14	5.41	32	**257**	ethanethiol	2.34E+02	5.46	5.26	1
186	2-methyl-3-pentanol[b]	2.78E+02	5.63	5.37	32	**258**	thiophene	1.55E+03	7.35	6.96	1
187	4-methyl-2-pentanol	3.50E+02	5.86	5.35	32	**259**	diethyl sulfide	1.61E+03	7.38	7.54	1
188	3-methyl-2-pentanol	2.88E+02	5.66	5.39	32	**260**	1-butanethiol[b]	8.34E+03	9.03	8.11	1
189	3-methyl-3-pentanol	1.28E+02	4.85	5.24	32	**261**	styrene	1.80E+04	9.80	9.86	32
190	2-methyl-2-hexanol	6.60E+02	6.49	6.86	32	**262**	indan[b]	6.02E+04	11.0	11.0	32
191	3-methyl-3-hexanol	5.37E+02	6.29	6.57	32	**263**	*m*-methylstyrene	7.37E+04	11.2	11.0	2
192	2,3-dimethyl-2-pentanol	4.13E+02	6.02	6.40	32	**264**	*p*-methylstyrene[b]	7.37E+04	11.2	11.0	2
193	2,4-dimethyl-2-pentanol	4.76E+02	6.17	6.52	32	**265**	1,2,3-trimethylbenzene	8.87E+04	11.4	11.6	32
194	2,2-dimethyl-3-pentanol	7.81E+02	6.66	6.21	32	**266**	1,2,4-trimethylbenzene	1.17E+05	11.7	11.6	32
195	2,3-dimethyl-3-pentanol	3.88E+02	5.96	6.01	32	**267**	*o*-ethyltoluene	7.17E+04	11.2	11.6	2
196	2,4-dimethyl-3-pentanol	9.16E+02	6.82	6.46	32	**268**	*p*-ethyltoluene	7.03E+04	11.2	12.0	2
197	3-ethyl-3-pentanol	3.79E+02	5.94	5.85	32	**269**	*sec*-butylbenzene	4.24E+05	13.0	12.8	32
198	1-octanol	1.42E+04	9.56	10.0	32	**270**	*tert*-butylbenzene	2.53E+05	12.4	12.5	32
199	2,2,3-trimethyl-3-pentanol	1.04E+03	6.95	6.78	32	**271**	1,2,4,5-tetramethylbenzene	2.14E+06	14.6	12.8	2
200	1-nonanol	6.17E+04	11.0	11.3	32	**272**	1-methylnaphthalene	2.82E+05	12.6	12.8	32
201	1,3-nonanediol	6.10E+02	6.41	6.69	33	**273**	1,3-dimethylnaphthalene	1.09E+06	13.9	13.6	32
202	1-decanol	2.38E+05	12.4	12.5	32	**274**	1,4-dimethylnaphthalene[c]	7.63E+05	13.6	13.2	32
203	2-propyl-1,3-heptanediol[b]	9.40E+02	6.85	6.94	33	**275**	1-ethylnaphthalene	8.06E+05	13.6	13.5	32
204	2,4-dimethyl-2,4-octanediol	7.84E+02	6.66	6.68	33	**276**	*n*-butane	5.25E+04	10.9	10.6	2
205	2,4-dimethyl-2,4-nonanediol[c]	2.48E+03	7.82	7.71	33	**277**	2-methylpropane	6.60E+04	11.1	10.7	2
206	1-dodecanol	4.44E+06	15.3	14.7	32	**278**	2-methyl-1,3-butadiene	5.89E+03	8.68	8.34	2
207	2-butyl-1,3-octanediol	1.87E+04	9.84	10.1	33	**279**	cyclopentene	7.07E+03	8.86	8.61	32
208	1-tetradecanol[b]	3.97E+07	17.5	17.1	1	**280**	1,4-pentadiene	6.78E+03	8.82	8.84	32
209	1-pentadecanol	1.42E+08	18.8	18.4	1	**281**	1-pentyne	2.41E+03	7.79	7.94	32
210	1-hexadecanol	3.85E+08	19.8	19.9	1	**282**	2-methyl-2-butene	1.78E+04	9.79	9.11	23
211	1-heptadecanol	1.78E+09	21.3	21.5	1	**283**	3-methyl-1-butene	2.99E+04	10.3	9.77	2
212	1-octadecanol[c]	1.37E+10	23.3	22.8	1	**284**	cyclopentane	2.49E+04	10.1	10.6	32
213	5-methylfurfural	1.28E+02	4.85	4.91	32	**285**	1-pentene	2.63E+04	10.2	10.2	32
214	heptanal[b]	4.18E+03	8.34	8.04	32	**286**	2-pentene[c]	1.92E+04	9.86	10.1	32
215	nonanal	7.52E+04	11.2	10.8	32	**287**	2,2-dimethylpropane	1.21E+05	11.7	11.3	34
216	cyclohexanone	5.41E+01	3.99	3.77	32	**288**	2-methylbutane	8.25E+04	11.3	11.6	34

Table 1 (Continued)

cmpd. no.	compound name	expt γ^∞	expt ln (γ^∞)	calc ln (γ^∞)	ref	cmpd. no.	compound name	expt γ^∞	expt ln (γ^∞)	calc ln (γ^∞)	ref
289	*n*-pentane	9.44E+04	11.5	11.8	34	**308**	1-methylcyclohexene	1.03E+05	11.5	11.0	32
290	1,4-cyclohexadiene	5.45E+03	8.60	9.40	34	**309**	1,6-heptadiene	1.21E+05	11.7	11.7	32
291	cyclohexene	2.83E+04	10.3	10.4	34	**310**	1-heptyne	5.68E+04	11.0	10.6	32
292	1,5-hexadiene	2.70E+04	10.2	10.3	32	**311**	cycloheptane	1.82E+05	12.1	12.2	32
293	1-hexyne	1.27E+04	9.45	9.23	32	**312**	methylcyclohexane[b]	3.61E+05	12.8	12.7	34
294	cyclohexane	8.03E+04	11.3	11.2	34	**313**	2-heptene	3.63E+05	12.8	13.4	32
295	methylcyclopentane	1.09E+05	11.6	11.7	34	**314**	2,2-dimethylpentane	1.26E+06	14.1	13.9	34
296	2,3-dimethyl-1-butene	1.02E+04	9.23	10.3	34	**315**	2,3-dimethylpentane	1.06E+06	13.9	13.6	34
297	1-hexene[b]	8.86E+04	11.4	11.7	34	**316**	2,4-dimethylpentane	1.31E+06	14.1	14.0	34
298	2-methyl-1-pentene	5.98E+04	11.0	10.8	34	**317**	3,3-dimethylpentane[b]	9.39E+05	13.8	13.4	34
299	4-methyl-1-pentene[b]	9.74E+04	11.5	11.2	32	**318**	*n*-heptane	2.01E+06	14.5	14.7	34
300	2,2-dimethylbutane	2.27E+05	12.3	12.3	34	**319**	2-methylhexane	2.19E+06	14.6	14.5	34
301	2,3-dimethylbutane[c]	2.30E+05	12.4	12.2	34	**320**	4-ethenylcyclohexene	1.20E+05	11.7	11.6	32
302	*n*-hexane	4.23E+05	13.0	13.2	34	**321**	1-octyne	2.55E+05	12.5	12.0	32
303	2-methylpentane	3.49E+05	12.8	13.0	34	**322**	*cis*-1,2-dimethylcyclohexane	1.04E+06	13.9	13.5	32
304	3-methylpentane	3.71E+05	12.8	12.7	34	**323**	cyclooctane[c]	7.89E+05	13.6	13.2	32
305	1,3,5-cycloheptatriene[c]	7.92E+03	8.98	9.36	34	**324**	1-octene	2.31E+06	14.7	14.7	32
306	1,6-heptadiyne	3.10E+03	8.04	8.44	32	**325**	*n*-octane	9.09E+06	16.0	16.1	32
307	cycloheptene	8.09E+04	11.3	11.8	32						

[a] Calculated ln (γ^∞) are from Type 3 model developed for all compounds. [b] Cross-validation set compound. [c] Prediction set compound. [d] Compound removed as an outlier.

Table 2. Theoretical Linear Solvation Energy Relationship Descriptors Calculated from MOPAC Data

label	descriptor definition [method of calculation]
VOL	molecular van der Waals volume [use volume from ADAPT routine]
POLE	dipolarity/polarizability [(MOPAC output polarization volume)/VOL]
CHBB	covalent hydrogen bonding basicity [MOPAC LUMO(H_2O) − MOPAC HOMO(solute)]
CHBA	covalent hydrogen bonding acidity [MOPAC LUMO(solute) − MOPAC HOMO(H_2O)]
EHBB	electrostatic hydrogen bonding basicity [MOPAC most negative formal charge on a solute atom]
EHBA	electrostatic hydrogen bonding acidity [MOPAC most positive formal charge on a solute hydrogen]

set were gathered from many different literature sources,[1,2,6,13,15,17,23,35−40] and, because of errors in the published compilation, the experimental values used in this study were taken from the original source whenever possible. In the development of the models reported here, the ln (γ^∞) was used as the dependent variable to compress the range covered by the data. The compounds, their experimental γ^∞, experimental and calculated ln (γ^∞), and the reference from which the values were taken are listed in Table 1.

The data set of 325 organic compounds contained C, H, O, N, S, X, and spanned a molecular weight range of 30 to 394. The γ^∞ values ranged from 9.00×10^{-2} for dimethyl sulfoxide to 1.37×10^{10} for 1-octadecanol, giving a range of −2.41 to 23.3 for ln (γ^∞). The compounds were functionally diverse and included hydrocarbons, halohydrocarbons, alcohols, ethers, aldehydes, ketones, esters, carboxylic acids, amines, amides, nitro-compounds, nitriles, and sulfur-containing compounds. Of the 325 compounds, 40% of the experimental values were measured directly and 60% were determined indirectly from aqueous solubility values. The experimental error associated with the directly measured values was 10% for compounds with $\gamma^\infty < 1000$ and 20% for compounds with $\gamma^\infty > 1000$. There was no experimental error reported for the indirectly determined γ^∞ values. This experimental error would depend on the error associated with the measurement of the aqueous solubility values.

The 325 compound data set was divided into a 300-compound training set (t_{set}) that was used for developing the Type 1 linear regression models and a 25-compound prediction set (p_{set}) that was used for external validation of all of the models that were generated. In the generation of the Type 2 and Type 3 CNN models, the t_{set} was further subdivided into a 271-compound t_{set} and a 25-compound cross-validation set (cv_{set}). The cv_{set} was used to determine when to stop training the CNN so that it would have good, general predictive ability. The p_{set} used for CNN was retained from the multiple linear regression portion of the study for comparison purposes. The specific p_{set} and cv_{set} compounds were chosen randomly, but they were chosen to span the entire range of the data. These compounds are identified by superscripts in Table 1.

The compounds were sketched into HyperChem as two-dimensional (2D) representations, and initial modeling was performed to generate three-dimensional (3D) conformations. Further modeling was performed to refine the 3D conformations to their lowest energy states using MOPAC,[41] a semiempirical molecular orbital modeling routine, with the PM3 Hamiltonian. MOPAC was used to put the compounds into accurate 3D conformations for the generation of descriptors dependent on geometry.

Standard ADAPT routines were used to calculate descriptors that encode the topological, geometric, and electronic features of the compounds in the data set. The descriptors were calculated directly from the energy-minimized 3D conformations, with the goal of encoding the structural features that give rise to differences in γ^∞ for different compounds. The topological descriptors included counts of atom types, bond types, and functional groups, as well as molecular connectivity indices and path counts to represent the size and degree of branching of data set compounds. Geometric descriptors included molecular surface area, volume, and moments of inertia. The electronic environments of compounds were encoded with partial atomic charge descriptors, like the charge on the most positive or most negative atom. Descriptors that combined geometric and

electronic information, such as charged partial surface area (cpsa) descriptors and hydrogen bonding descriptors, were also included. Because the property being studied, γ^{∞}, was a solution property, the hydrogen bonding descriptors were calculated assuming a mixed solution of solute and water to take into account the effects that the solvent would have on the hydrogen bonding of the system. The heat of formation, electronic energy, and ionization potentials calculated during the MOPAC energy minimization with the PM3 Hamiltonian were also used.

A set of descriptors using information from MOPAC runs on all compounds using the MNDO Hamiltonian was also calculated. These descriptors were developed and used by Lowrey et al.[42] to develop a theoretical complement to LSERs, which has been termed the theoretical linear solvation energy relationship (TLSER) method. The TLSER descriptors were intended to describe the chemical and physical interactions of molecules with solvents. Lowrey et al.[42] generated six theoretical descriptors, defined in Table 2, from information taken from the semiempirical molecular orbital calculations—a steric term represented by the van der Waals volume, a volume-independent molecular polarizability term, covalent hydrogen bonding acidity and basicity, and electrostatic hydrogen bonding acidity and basicity. In this study, five of these descriptors were calculated, and the steric term was represented by the volume descriptor generated by ADAPT programs.

Finally, a new set of topological descriptors developed by Cao[43] was calculated and found to be useful in this study. The molecular distance-edge (MDE) vector consisted of the descriptor values for 10 distance-edge terms (D_{11}, D_{12}, D_{13}, D_{14}, D_{22}, D_{23}, D_{24}, D_{33}, D_{34}, D_{44}) between four different types of carbon atoms. For example, D_{12} was the distance-edge term between C_1 and C_2 carbons. The four types of carbon atoms were classified simply as C_1 (3 bonded hydrogens), C_2 (2 bonded hydrogens), C_3 (1 bonded hydrogen), and C_4 (no bonds to hydrogens). The distance-edge d_{ij} was defined as follows:

$$d_{ij} = \prod_{k<l}^{n_{ij}} (d_{ik,jl})^{1/2n_{ij}} \; (i = 1, 2, 3, 4; \; j \geq i) \qquad (1)$$

where $d_{ik,jl}$ was the number of bonds between carbons C^k_i and C^l_j with k and l denoting the carbon identity and i and j denoting the classification type already described. The identity of the carbon atoms was arbitrarily assigned. In eq1, n_{ij} was the number of times a particular interaction, for example between types C_1 and C_2, occurred in the compound of interest. The members of the MDE vector were then defined as follows:

$$D_{ij} = (n_{ij}/d_{ij}^2) \; (i = 1, 2, 3, 4; \; j \geq i) \qquad (2)$$

and represented 10 new descriptor values that were added to the descriptor pool. The new MDE descriptors did not correlate highly with other topological descriptors.

All of the descriptors were subjected to objective feature selection to remove those that did not contribute useful information to the pool. Pairwise correlations between descriptors were examined so that only one descriptor was retained from a pair contributing similar information (correlation coefficients ≥ 0.93), and descriptors with >90% identical values were dropped because those descriptors were not encoding the structural differences between compounds that accounts for their different γ^{∞} values. Objective feature selection left a reduced pool of 109 descriptors for the 300 compound t_{set}, well below the cutoff of 0.6 for the ratio of descriptors to observations used to limit the possibility of finding a relationship based on chance effects.[44]

The reduced pool of information-rich descriptors was then examined to find subsets of descriptors that accurately represented the relationship between molecular structure and ln (γ^{∞}). Simulated annealing[32] and genetic algorithm routines,[31] along with an interactive regression routine, were used to generate statistically valid linear models, termed Type 1 models. The genetic algorithm and simulated annealing routines each investigated a large number of descriptor subsets, with the quality of the model based on the root-mean-square (rms) error and statistical integrity. A variety of subset sizes was investigated to determine the optimum number of descriptors in a model. When adding another descriptor did not improve the statistics of a model, it was determined that the optimum subset size had been achieved.

The descriptors chosen for the Type 1 models were submitted to CNN for improvement as Type 2 models. A fully connected, feed-forward, three-layer neural network was used. The input layer consisted of as many neurons as there were descriptors in the Type 1 model. The number of neurons in the hidden layer was considered to be optimized when the t_{set} error did not significantly decrease with the addition of another neuron. A single neuron in the output layer provided the estimated ln (γ^{∞}). A linear transformation of the descriptor values restricted them to the interval [0,1], and these values were then used as the input for the network. A quasi-Newton BFGS (Broyden-Fletcher-Goldfarb-Shanno)[45−49] algorithm[34] was used to train the networks. A cv_{set} was used to prevent overtraining. The cv_{set} was a small subset of compounds, usually 10%, randomly drawn from the t_{set} that was not included during training but was tested periodically during training. When the cv_{set} rms error was minimized, training was stopped because beyond this point the network was fitting characteristics specific to individual t_{set} compounds rather than general characteristics of the entire data set that relate to the property of interest.

The CNN results were very dependent on the starting weights and biases, which were randomly selected. Two methods were used to sample a variety of starting points. A generalized simulated annealing algorithm was used to try to optimize the starting weights and biases, and an automated CNN program was run that selected a user-determined number of random starting weights and biases. The results from this large number of trials were then examined to find a good starting point for the weights and biases.

CNN were also used as the basis for feature selection to generate Type 3 models. A genetic algorithm routine was used to investigate subsets of descriptors. The quality of the models was based on the following fitness function:

$$\text{quality} = \text{TSET} + 0.4|\text{TSET} - \text{CVSET}| \qquad (3)$$

where TSET and CVSET refer to the t_{set} and cv_{set} rms errors.[33] Models chosen with this quality factor that had the lowest t_{set} errors, as well as cv_{set} and t_{set} errors that were similar, performed better than models chosen with just the

Table 3. Twelve Descriptors Chosen by Multiple Linear Regression for the Prediction of Infinite Dilution Activity Coefficients of Organic Compounds

label	descriptor definition	coeff.	SD of coeff.
NDB	no. of double bonds	0.606	0.103
ESTR	no. of ester groups	0.963	0.190
V6PC	valence corrected sixth order path-cluster molecular connectivity	1.07	0.14
PPSA 3	atomic charge weighted partial positive surface area	0.833	0.023
DPSA 3	difference in atomic charge weighted partial positive and negative surface areas	−0.303	0.019
FPSA 3	fractional atomic charge weighted partial positive surface area	−66.8	8.2
WNSA 1	surface weighted partial negative surface area	0.153	0.008
CHAA 1	sum of charges on hydrogen bonding acceptor atoms	16.5	0.6
SCAA 1	sum of (surface area x charge) of hydrogen bonding acceptor atoms	−0.240	0.015
HoF	heat of formation	−0.0234	0.0022
CHBB	covalent hydrogen bonding basicity	0.868	0.075
EHBB	electrostatic hydrogen bonding basicity	−10.5	0.6
	constant	−7.68	1.38

R = 0.978 t_{set} rms error = 0.753 ln units n = 296 compounds

t_{set} error as the quality factor. Including the error of the cv_{set} in the quality factor of the CNN gave models with good external predictive ability. The coefficient of 0.4 in the quality factor was empirically determined. Unlike the development of Type 1 models, it was extremely computationally intensive to investigate a large number of different Type 3 models. The architecture for the CNN was retained from the Type 2 models, providing a basis for comparison. Type 3 models were found that had lower rms errors than the corresponding Type 2 models, proving that the descriptors chosen based on linear criteria were not the best subset that could be found for use with CNN.

All of the models that were developed in the QSPR study were validated with an external prediction set. The models were able to accurately predict the ln (γ^{∞}) of these compounds, which had not been used in the development of the models, with p_{set} rms errors of the same magnitude as the t_{set} and cv_{set} errors. The validity of Type 1 models was also determined with the use of standard statistical parameters such as the correlation coefficient, the overall F value of the model, and the T values of the individual descriptors. The presence of outliers was detected by looking at regression diagnostics that measured the effect of individual data points on the model, such as standardized and studentized residuals, leverage, and Cook's distance.[50] Compounds were often flagged as outliers if they were not well represented in the data set. An additional method of validation for all three model types was the visual inspection of both calculated versus observed plots of ln (γ^{∞}) and plots of residuals.

A second experiment was performed to compare the models developed for the entire data set with models developed for a smaller subset of compounds with small γ^{∞} values. The 214 compounds in this subset had $\gamma^{\infty} \leq 1000$. These compounds have very large aqueous solubility values that are difficult to measure experimentally. The 214 compound subset used for further model development was broken down into a tset with 197 members and a pset with 17 members. The 17 pset members were those compounds from the initial pset with $\gamma^{\infty} \leq 1000$. For the CNN portion of this experiment, a cv_{set} of 17 compounds was removed from the t_{set}. Descriptor pool reduction, model generation, and model validation were then performed on this set of compounds.

RESULTS AND DISCUSSION

All Compounds. The best Type 1 model for the entire data set using the combination of multiple linear regression routines[31,32] consisted of 12 structural descriptors defined in Table 3. The model contained three topological descriptors (NDB, ESTR, and V6PC), four cpsa descriptors (PPSA 3, DPSA 3, FPSA 3, and WNSA 1), two hydrogen bonding descriptors from ADAPT routines (CHAA 1 and SCAA 1), the heat of formation calculated during the MOPAC geometry optimization (HoF), and two of the TLSER descriptors (CHBB and EHBB), both of which described hydrogen bonding basicity. The combination of these 12 descriptors was able to accurately encode intermolecular interactions taking place between the solute and water that determine the γ^{∞} values. It was not surprising that four of the 12 descriptors chosen contained information characterizing the hydrogen bonding ability of the system, considering the substantial hydrogen bonding capability of water. Dipole interactions could be interpreted as being represented by the descriptor indicating the presence of ester groups and the four cpsa descriptors that included electronic information about the positive and negative partial surface areas of the solutes. The descriptors that encoded size and shape, namely the number of double bonds and the valence corrected path cluster descriptor, could be encoding London dispersion forces.

The Type 1 model for ln (γ^{∞}) had a t_{set} rms error of 0.753 ln units. After diagnostic testing for outliers, four compounds were flagged and removed from the tset, leaving 296 of the original 300 t_{set} compounds for the linear regression. The outliers were benzonitrile, hexafluoroethane, triiodomethane, and benzoic acid. Hexafluoroethane and triiodomethane were both halogenated compounds whose behavior could be substantially affected by the high degree of halogenation. Benzoic acid was one of five carboxylic acids in the data set and was the only aromatic acid. Similarly, benzonitrile was the only aromatic nitrile of the seven in the data set. The fact that these compounds were removed as outliers could be explained by the fact that they were unrepresentative of the structures of the data set compounds. In validation of this Type 1 model, the external prediction set had an rms error of 0.624 ln units. Figure 1 is a plot of calculated and predicted ln (γ^{∞}) versus observed ln (γ^{∞}) for the Type 1 linear

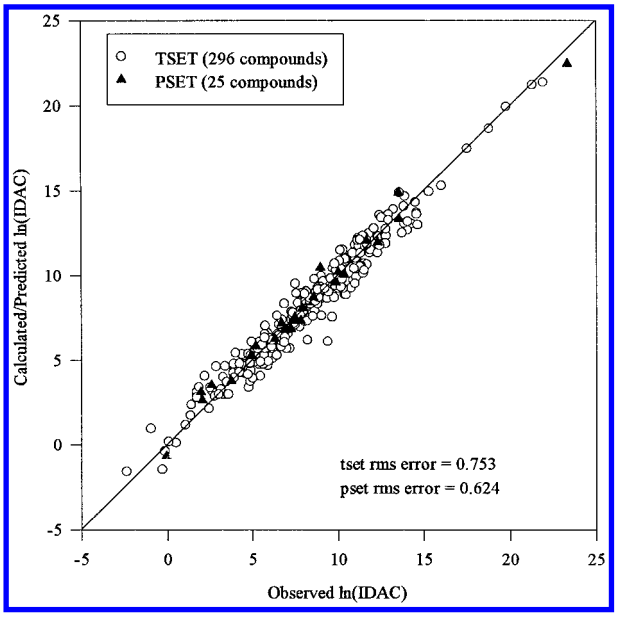

Figure 1. Calculated/predicted versus observed ln (γ^{∞}) for Type 1 model for all compounds.

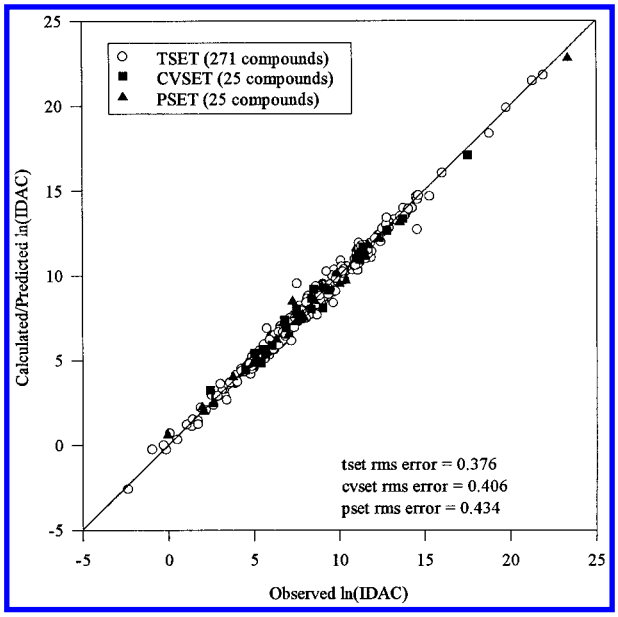

Figure 2. Calculated/predicted versus observed ln (γ^{∞}) for Type 3 model for all compounds.

regression model for the entire data set with the four outliers removed.

The descriptors chosen by the linear regression feature selection routines were then submitted to CNN to develop a nonlinear model. The Type 2 model resulted in t_{set}, cv_{set}, and p_{set} rms errors of 0.472, 0.538, and 0.484 ln units, respectively. These values represented a 37% improvement over the Type 1 t_{set} error and a 14% improvement over the Type 1 p_{set} error. The CNN consisted of 12 input neurons corresponding to the 12 descriptors chosen by linear feature selection, six hidden neurons, and one output neuron corresponding to the calculated ln (γ^{∞}). The number of hidden layer neurons was optimized by adding neurons until there was no further improvement in the cost function of the neural network. The 12:6:1 architecture of the CNN had 85 adjustable parameters for 271 t_{set} compounds, which is well above the cutoff of 2.0 for the ratio of observations to adjustable parameters.

The entire data set was also submitted to the CNN genetic algorithm feature selection routine[31] to search for a nonlinear relationship between structural descriptors and ln (γ^{∞}). A high quality CNN model with 12:6:1 architecture was found. The t_{set} rms error for the Type 3 model was 0.376 ln units, an improvement of 20% over the Type 2 CNN model; the cv_{set} rms error was 0.406 ln units, a 25% improvement; and the p_{set} rms error was 0.434 ln units, a 10% improvement. These results are shown in Figure 2. The corresponding mean absolute errors are 0.275 ln units for the t_{set}, 0.344 ln units for the cv_{set}, and 0.327 ln units for the p_{set}. This Type 3 neural network model is the best model found in the course of this work.

The descriptors chosen for the Type 3 model are listed and defined in Table 4. Of the 12 descriptors chosen by the nonlinear feature selection routine, six (PPSA 3, DPSA 3, WNSA 1, HoF, CHBB, and EHBB) were identical, with descriptors chosen in the linear model. The 12 descriptors consisted of three topological descriptors (NBR, ALLP 4, and WTPT 3), one geometric descriptor (GRVH 3), four cpsa descriptors (PPSA 3, DPSA 3, FPSA 1, and WNSA 1), a hydrogen bonding descriptor (CHAA 2), the heat of formation from MOPAC (HoF), and the same TLSER descriptors (CHBB and EHBB). As in the Type 1 model, hydrogen bonding and cpsa descriptors were highly represented. The size and shape of the solute molecules were encoded through the presence of the topological and geometric descriptors. The nonlinear relationships among these descriptors was able to more accurately predict ln (γ^{∞}) than the descriptors chosen based on a linear relationship.

Compounds with $\gamma^{\infty} \leq 1000$. Separate Type 1, 2, and 3 models were developed for the subset of 214 extremely

Table 4. Twelve Descriptors Chosen for a Type 3 Model for the Prediction of Infinite Dilution Activity Coefficients of Organic Compounds

label	descriptor definition
NBR	no. of basis rings
ALLP 4	total weighted no. of paths/no. of atoms
WTPT 3	sum of path weights starting from heteroatoms
GRVH 3	cube root of gravitation index including hydrogens
PPSA 3	atomic charge weighted partial positive surface area
DPSA 3	difference in atomic charge weighted partial positive and negative surface areas
FPSA 1	fractional partial positive surface area
WNSA 1	surface weighted partial negative surface area
CHAA 2	sum of charges on hydrogen bonding acceptor atoms/no. of hydrogen bonding acceptor atoms
HoF	heat of formation
CHBB	covalent hydrogen bonding basicity
EHBB	electrostatic hydrogen bonding basicity
t_{set} rms error = 0.376 ln units cv_{set} rms error = 0.406 ln units n = 271 compounds	

Table 5. Twelve Descriptors Chosen by Multiple Linear Regression for the Prediction of Infinite Dilution Activity Coefficients ≤ 1000

label	descriptor definition	coeff.	SD of coeff.
NC	no. of carbons	1.13	0.04
ESTR	no. of ester groups	1.80	0.18
MDE 14	product of distances between primary and quaternary carbons	−0.221	0.046
ALLP 4	total weighted no. of paths/no. of atoms	−3.18	0.32
WTPT 4	sum of path weights starting from oxygens	−0.766	0.071
WTPT 5	sum of path weights starting from nitrogens	−1.22	0.07
FNSA 2	fractional charge weighted partialnegative surface area	−2.76	0.56
CHAA 1	sum of charges on hydrogen bonding acceptor atoms	8.03	0.39
CTAA 0	no. of hydrogen bonding acceptor atoms	2.45	0.19
POLE	dipolarity/polarizability	217	23
CHBB	covalent hydrogen bonding basicity	0.745	0.057
EHBB	electrostatic hydrogen bonding basicity	−7.94	0.72
	constant	−9.95	1.15

$R = 0.967$ t_{set} rms error = 0.559 ln units $n = 193$ compounds

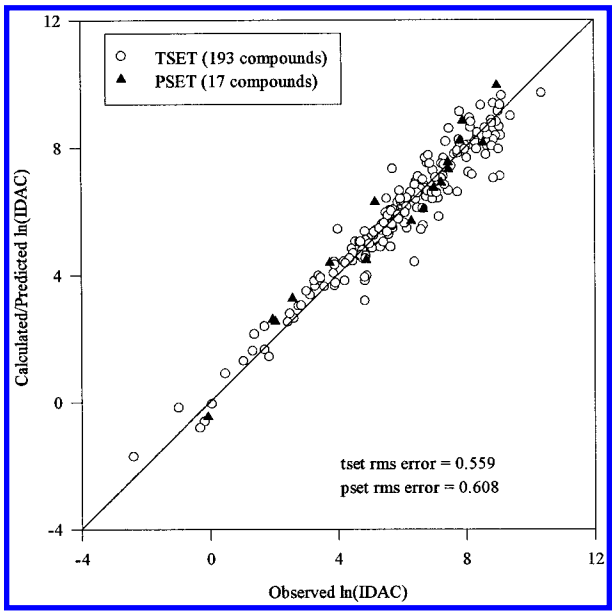

Figure 3. Calculated/predicted versus observed ln (γ^{∞}) for Type 1 model for compounds with $\gamma^{\infty} \leq 1000$.

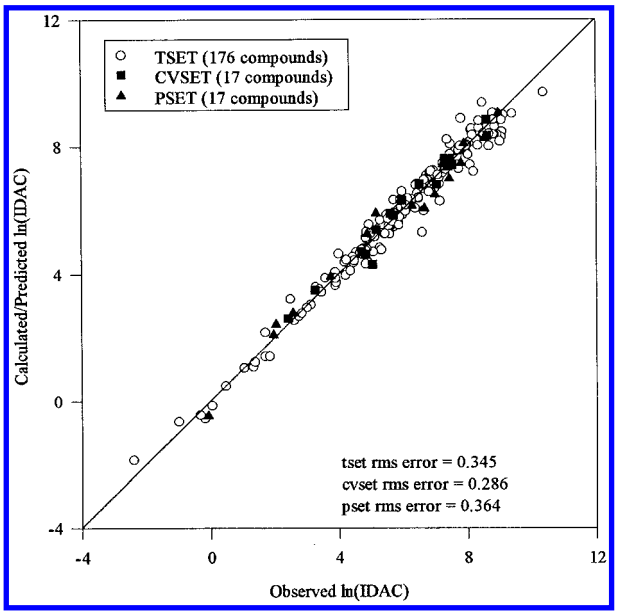

Figure 4. Calculated/predicted versus observed ln (γ^{∞}) for Type 2 model for compounds with $\gamma^{\infty} \leq 1000$.

soluble compounds for which $\gamma^{\infty} \leq 1000$. The best Type 1 model that was generated from a reduced pool of 113 descriptors using the simulated annealing,[32] genetic algorithm,[31] and interactive regression analysis routines, had 12 descriptors, with t_{set} and p_{set} rms errors of 0.559 and 0.608 ln units. Figure 3 shows the calculated and predicted versus observed ln (γ^{∞}) for the Type 1 model for this reduced data set. Note that the range of the ln (γ^{∞}) values is now from negative four to 12, a much narrower range than in the previous plots.

Four outliers were removed from the 197 compound tset. They were trifluoroethanol, benzoic acid, 1-propylpiperidine, and 1,6-heptadiyne. All of these compounds were severely underrepresented in the data set. Trifluoroethanol was the only halogenated alcohol, 1-propylpiperidine was one of only two piperidines, 1,6-heptadiyne was the only diyne, and, as mentioned previously, benzoic acid was one of five carboxylic acids and the only aromatic acid in the data set. Benzoic acid was the only compound that was flagged as an outlier and removed in both parts of this study.

Table 5 lists the descriptors in the Type 1 model with their definitions. There were more topological descriptors and fewer cpsa descriptors compared with the Type 1 model for the data set of all compounds. Six of the 12 descriptors were topological (NC, ESTR, MDE 14, ALLP 4, WTPT 4, and WTPT 5). Included in the subset of topological descriptors was one of the new MDE descriptors discussed previously. MDE 14 described a relationship between primary and quaternary carbons in the data set compounds. Of the remaining six descriptors, one was a cpsa descriptor (FNSA 2), two described hydrogen bonding (CHAA 1 and CTAA 0), and three were TLSER descriptors (POLE, CHBB, and EHBB). The combination of these descriptors could again be interpreted as encoding dipole interactions, London dispersion forces, and hydrogen bonding interactions. CHBB and EHBB were the only two descriptors present in all of the models, Type 1 and Type 3 for the whole data set and the reduced data set, developed in this study. ESTR, ALLP 4, WTPT 5, CHAA 1, and POLE also appeared in more than one model.

These 12 descriptors were then used as input for a 12:6:1 solution of CNN. After training, the resulting Type 2 model had a t_{set} rms error of 0.345 ln units, a cv_{set} rms error of 0.286 ln units, and a p_{set} rms error of 0.364 ln units. These rms errors represented large improvements, 38% and 40%, respectively, between the Type 1 and Type 2 t_{set} and p_{set} rms errors. The plot in Figure 4 shows the high degree of agreement between the calculated/predicted and observed

values of ln (γ^{∞}).

Once again, feature selection was performed with the genetic algorithm routine[31] using CNN as the cost function to develop a Type 3 model with 12:6:1 CNN architecture. The descriptors chosen in this case, however, did not show an improvement for the t_{set} and the cv_{set} over the descriptors chosen by linear feature selection in relating structure to ln (γ^{∞}). The t_{set} and cv_{set} rms errors were 0.381 and 0.400 ln units, both higher than the 0.345 and 0.286 ln units for the Type 2 CNN results. The p_{set} rms error did drop slightly, to 0.353 ln units for the Type 3 model from the Type 2 results of 0.364 ln units, but this was only a 3% decrease.

CONCLUSIONS

In this QSPR study, models have been developed using the ADAPT methodology that relate molecular structure to γ^{∞} in water. The models reported from this study are the first with the ability to predict γ^{∞} from structural information alone for such a diverse set of organic compounds. The structural environments of a set of industrially useful organic compounds are numerically encoded as molecular descriptors, including theoretical linear solvation energy relationship descriptors and a new set of topological descriptors, molecular distance-edge descriptors. Linear feature selection led to the development of separate Type 1 linear regression and improved Type 2 CNN models for both the entire set of compounds and the reduced set of compounds with $\gamma^{\infty} \leq 1000$. Nonlinear feature selection led to the development of Type 3 computational neural network models. The Type 3 model for the entire set of compounds had the lowest rms errors of all of the models that were generated.

An LSER has been reported by Sherman et al.[23] relating ln γ^{∞} to five experimentally determined solvatochromic parameters. From the same data compilation, only 225 of the solutes were used in developing the LSER relationship because the necessary solvatochromic parameters were not available for many of the compounds. An average absolute deviation of 0.294 ln units was reported for this correlation. The t_{set}, cv_{set}, and p_{set} mean absolute errors of 0.275, 0.344, and 0.327 ln units, respectively, for the Type 3 model for the entire data set are somewhat smaller than the error associated with the LSER model and a larger set of compounds was used. Sherman et al.[23] compare their LSER method to several UNIFAC variations and show that the UNIFAC methods have substantially larger errors overall and in addition show systematic deviations for higher γ^{∞} values.

This study has shown that a relationship can be developed between structure and γ^{∞} for a diverse set of compounds covering a wide range of γ^{∞} values. The γ^{∞} value of a compound is directly related to its aqueous solubility, but the relationships that have been developed can be applied to compounds whose aqueous solubility values are unmeasurable or measured only with great difficulty, extending the range of compounds for which phase equilibrium information can be studied and predicted.

The models that have been developed with the use of multiple linear regression and CNN techniques can be applied to the prediction of γ^{∞} of compounds not present in the data set used in this study. The predictive power of these models will be useful in cases where it is too costly, time-consuming, or impossible to measure the values experimentally. The models in this study can also provide insight into the structural features that contribute to the physical forces between solute and solvent that give rise to γ^{∞}.

ACKNOWLEDGMENT

This work was supported by the Design Institute for Physical Property Data (DIPPR)® Project 931: Data Prediction Methods.

REFERENCES AND NOTES

(1) Yaws, C. L.; Yang, H.; Hopper, J. R.; Hansen, K. C. Organic Chemicals: Water Solubility Data. *Chem. Eng.*, **1990**, *97*, 115−118.
(2) Yaws, C. L.; Yang, H.; Hopper, J. R.; Hansen, K. C. Hydrocarbons: Water Solubility Data. *Chem. Eng.* **1990**, *97*, 177−182.
(3) Grain, C. F. Activity Coefficient. In *Handbook of Chemical Property Estimation Methods, Environmental Behavior of Organic Compounds*; Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H., Eds.; McGraw-Hill: New York, 1982; Chapter 11.
(4) Bergmann, D. L.; Eckert, C. A. Measurement of Limiting Activity Coefficients for Aqueous Systems by Differential Ebulliometry. *Fluid Phase Equilib.* **1991**, *63*, 141−150.
(5) Lazaridis, T.; Paulaitis, M. E. Activity Coefficients in Dilute Aqueous Solutions from Free Energy Simulations. *AIChE J.* **1993**, *39*, 1051−1060.
(6) Li, J.; Dallas, A. J.; Eikens, D. I.; Carr, P. W.; Bergmann, D. L.; Hait, M. J.; Eckert, C. A. Measurement of Large Infinite Dilution Activity Coefficients of Nonelectrolytes in Water by Inert Gas Stripping and Gas Chromatography. *Anal. Chem.* **1993**, *65*, 3212−3218.
(7) Pescar, R. E.; Martin, J. J. Solution Thermodynamics from Gas−Liquid Chromatography. *Anal. Chem.* **1966**, *38*, 1661−1669.
(8) Shaffer, D. L.; Daubert, T. E. Gas−Liquid Chromatographic Determination of Solution Properties of Oxygenated Compounds in Water. *Anal. Chem.* **1969**, *41*, 1585−1589.
(9) Scott, L. S. Determination of Activity Coefficients by Accurate Measurement of Boiling Point Diagram. *Fluid Phase Equilib.* **1986**, *26*, 149−163.
(10) Trampe, D. M.; Eckert, C. A. Limiting Activity Coefficients from an Improved Differential Boiling Point Technique. *J. Chem. Eng. Data* **1990**, *35*, 156−162.
(11) Leroi, J.; Masson, J.; Renon, H.; Fabries, J.; Sannier, H. Accurate Measurement of Activity Coefficients at Infinite Dilution by Inert Stripping and Gas Chromatography. *Ind. Eng. Chem., Process Des. Dev.* **1977**, *16*, 139−144.
(12) Richon, D.; Antoine, P.; Renon, H. Infinite Dilution Activity Coefficients by Linear and Branched Alkanes from C_1 to C_9 in n-Hexadecane by Inert Gas Stripping. *Ind. Eng. Chem. Process Des. Dev.* **1980**, *19*, 144−147.
(13) Wright, D. A.; Sandler, S. I.; DeVoll, D. Infinite Dilution Activity Coefficients and Solubilities of Halogenated Hydrocarbons in Water at Ambient Temperature. *Environ. Sci. Technol.* **1992**, *26*, 1828−1831.
(14) Hussam, A.; Carr, P. W. Rapid and Precise Method for the Measurement of Vapor/Liquid Equilibria by Headspace Gas Chromatography. *Anal. Chem.* **1985**, *57*, 793−801.
(15) Li, J. Solvatochromic and Thermodynamic Studies in Gas Chromatography and Gas−Liquid Equilibria, Ph.D. Dissertation, University of Minnesota, Minneapolis, MN, 1992.
(16) Abraham, M. H.; Whiting, G. S.; Fuchs, R.; Chambers, E. J. Thermodynamics of Solute Transfer from Water to Hexadecane. *J. Chem. Soc., Perkin Trans.* **1990**, *2*, 291−300.
(17) Dallas, A. J. Solvatochromic and Thermodynamic Studies of Chromatographic Media, Ph.D. Dissertation, University of Minnesota, Minneapolis, MN, 1995.
(18) Abraham, M. H.; Grellier, P. L.; McGill, R. A. Determination of Olive Oil-Gas and Hexadecane-Gas Partition Coefficients, and Calculation of the Corresponding Olive Oil−Water and Hexadecane-Water Partition Coefficients. *J. Chem. Soc., Perkin Trans.* **1987**, *2*, 797−803.
(19) Tochigi, K.; Kojima, K. The Determination of Group Wilson Parameters to Activity Coefficients by Ebulliometer. *J. Chem. Eng. Jpn.* **1976**, *9*, 267−273.
(20) Tochigi, K.; Tiegs, D.; Gmehling. J.; Kojima, K. Determination of New ASOG Parameters. *J. Chem. Eng. Jpn.* **1990**, *23*, 453−463.
(21) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE J.* **1975**, *21*, 1086−1099.
(22) Hansen, H. K.; Rasmussen, P.; Fredenslund, A.; Schiller, M.; Gmehling, J. Vapor-Liquid Equilibria by UNIFAC Group Contribution. 5. Revision and Extension. *Ind. Eng. Chem. Res.* **1991**, *30*, 2352−2355.

(23) Sherman, S. R.; Trampe, D. B.; Bush, D. M.; Schiller, M.; Eckert, C. A.; Dallas, A. J.; Li, J.; Carr, P. W. Compilation and Correlation of Limiting Activity Coefficients of Nonelectrolytes in Water. *Ind. Eng. Chem. Res.* **1996**, *35*, 1044−1058.
(24) Shing, K. S. Infinite-Dilution Activity Coefficients from Computer Simulation. *Chem. Phys. Lett.* **1985**, *119*, 149−151.
(25) Yalkowsky, S. H.; Valvani, S. C. Solubilities and Partitioning 2. Relationships between Aqueous Solubilities, Partition Coefficients, and Molecular Surface Areas of Rigid Aromatic Hydrocarbons. *J. Chem. Eng. Data* **1979**, *24*, 127−129.
(26) Medir, M.; Giralt, F. Correlation of Activity Coefficients of Hydrocarbons in Water at Infinite Dilution with Molecular Parameters. *AIChE J.* **1982**, *28*, 341−343.
(27) Dutt, N. V. K.; Prasad, D. H. L. Estimation of Infinite Dilution Activity Coefficients of Hydrocarbons in Water from Molar Refraction. *Fluid Phase Equilib.* **1989**, *45*, 1−5.
(28) Mackay, D.; Shiu, W. Y. Aqueous Solubility of Polynuclear Aromatic Hydrocarbons. *J. Chem. Eng. Data* **1977**, *22*, 399−402.
(29) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(30) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; The American Chemical Society: Washington, D.C., 1979; pp 103−129.
(31) Wessel, M. D. Computer-Assisted Development of Quantitative Structure−Property Relationships And Design of Feature Selection Routines. Ph.D. Dissertation, Pennsylvania State University, University Park, PA, 1996.
(32) Sutter, J. M.; Jurs, P. C. Selection of molecular structure descriptors for quantitative structure−activity relationships. In *Adaption of Simulated Annealing to Chemical Problems*; Kalivas, J. H., Ed.; Elsevier Science: Amsterdam, 1995.
(33) 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.
(34) 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(5)*, 841−851.
(35) Barr, R. S.; Newsham, D. M. T. Phase Equilibria in Very Dilute Mixtures of Water and Chlorinated Hydrocarbons. Part I.-Experimental Results. *Fluid Phase Equilib.* **1987**, *35*, 189−205.
(36) Abraham, M. H.; Grellier, P. L.; Mana, J. Limiting Activity Coefficients in Triethylamine and 30 Solvents by a Simple Gas−Liquid Chromatographic Method. *J. Chem. Thermodyn.* **1974**, *6*, 1175−1179.
(37) Gillespie, P. C.; Cunningham, J. R.; Wilson, G. M. Total Pressure and Infinite Dilution Vapor Liquid Equilibrium Measurements for the Ethylene Oxide/Water System. *AIChE Symp. Series* **1985**, *81(244)*, 26−40.
(38) Sorenson, J. M.; Arlt, W. *Liquid−Liquid Equilibrium Data Collection, Binary Systems*, DECHEMA Chemistry Data Series, Vol. V, Part 1; DECHEMA: Frankfurt, 1979.
(39) Macedo, E. A.; Rasmussen, P. *Liquid−Liquid Equilibrium Data Collection, Supplement 1*, DECHEMA Chemistry Data Series, Vol. V, Part 4; DECHEMA: Frankfurt, 1987.
(40) Shaw, D. G. *Hydrocarbons with Water and Seawater, Part 1: Hydrocarbons C_5 to C_7*, IUPAC Solubility Data Series Vol. 37; Pergamon: Oxford, 1989.
(41) Stewart, J. P. P. MOPAC 6.0, *Quantum Chemistry Program Exchange*; Program 455; Indiana University, Bloomington, IN.
(42) Lowrey, A. H.; Cramer, C. J.; Urban, J. J.; Famini, G. R. Quantum Chemical Descriptors for Linear Solvation Energy Relationships. *Computers Chem.* **1995**, *19*, 209−215.
(43) Cao, C., Distance-Edge Topological Index − Research on Struccture-Property Relationships of Alkanes. *Huaxue Tongbao*. **1996**, *22*, 1238−1244.
(44) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(45) Broyden, C. G. The convergence of a class of double-rank minimization algorithms. *J. Inst. Maths. Appl.* **1970**, *6*, 76.
(46) Fletcher, R. A new approach to variable metric algorithms. *Comput. J.* **1970**, *13*, 317.
(47) Goldfarb, D. A family of variable-metric methods derived by variational means. *Math. Comput.* **1970**, *24*, 23.
(48) Shanno, D. F. Conditioning of quasi-Newton methods for function minimization. *Math. Comput.* **1970**, *24*, 647.
(49) Fletcher, R. *Practical Methods of Optimization, Vol. 1, Unconstrained Optimization*; Wiley: New York, 1980.
(50) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*; Irwin: Homewood, IL, 1990.

CI970092K