Piriou, F.; Lintner, K.; Lam-Thanh, H.; Toma, F.; Fermandjian, S. *Tetrahedron* **1978**, *34*, 553–556.
(58) Anteunis, M.; Gelan, J. *J. Am. Chem. Soc.* **1973**, *95*, 6502–6504.
(59) Grathwohl, C.; Shwyzer, R.; Tyn-Kyl, A.; Wüthrich, K. *FEBS Lett.* **1973**, *29*, 271–274.
(60) Despite the observation in ref 37 of unchanged conformations upon La^{3+} addition to L-4Hyp NMR spectra, a possible effect is suspected in the more flexible L-Pro part of the molecule. See also: Cockerill, A. F.; Davies, G. L. O.; Harden, R. C.; Rackham, D. M. *Chem. Rev.* **1973**, *73*, 553–588.
(61) Cis/trans ratios are pH and solvent dependent, depending on electronic densities induced by titration. See, for example, Voelter, W. V.; Oster, O. *Org. Magn. Reson.* **1973**, *5*, 547–548. Deslauriers, R.; Walter, R.; Smith, I. C. P. *Biochem. Biophys. Res. Commun.* **1972**, *48*, 854–859. More recently: Higashijima, T.; Tasumi, M.; Miyazawa, T. *Biopolymers* **1977**, *16*, 1259–1270. London, R. E.; Matwiyoff, N. A.; Stewart, J. M.; Cann, J. R. *Biochemistry* **1978**, *17*, 2277–2283.
(62) Bushweller, C. M.; O'Neil, J. W.; Halford, M. H.; Bisset, F. H. *J. Am. Chem. Soc.* **1971**, *93*, 1471–1473.
(63) Davies, D. B.; Kahled, A. *J. Chem. Soc., Perkin Trans. 2* **1976**, 1238–1244.
(64) Altona, C.; Geise, H. J.; Romers, C. *Tetrahedron* **1968**, *24*, 13–32.
(65) Ueki, T.; Bando, S.; Ashida, T.; Kakuki, M. *Acta Crystallogr., Sect. B* **1971**, *27*, 2219–2231.
(66) Pogliani, L. *Spectrosc. Lett.* **1975**, *8*(1), 37–41.
(67) Deslauriers, R.; Palva, A. C. M.; Schaumburg, K.; Smith, I. C. P. *Biopolymers* **1975**, *14*, 878–886. Deslauriers, R.; Smith, I. C. P. *Ibid.* **1977**, *16*, 1245–1257.
(68) Only a few examples of correlative X-ray/NMR studies have hitherto been reported. See, for example, Deslauriers, R.; Somorjai, E. L.; Ralston, E. *Nature (London)* **1977**, *266*, 746–748.
(69) Burns, D. M.; Ferrier, W. G.; Mullan, J. T. *Acta Crystallogr., Sect. B* **1968**, *24*, 734–737.
(70) Shomaker, V.; Trueblood, K. N. *Acta Crystallogr., Sect. B* **1968**, *24*, 63–76.
(71) Ibers, J. A.; Hamilton, W. C. In "International Tables for X-ray Crystallography"; Kynoch Press: Birmingham, England, 1974; Vol. IV, p 316.
(72) Deslauriers, R.; McGregor, W. H.; Sarantakis, D.; Smith, I. C. P. *Biochemistry* **1974**, *13*, 3443–3448.
(73) Becker, D. E. In "High Resolution NMR"; Academic Press: New York, 1969; p 204.
(74) Torchia, D. A.; Lyerla, J. R. *Biopolymers* **1974**, *13*, 97–114.
(75) See for previous examples: Deslauriers, R.; Walter, R.; Smith, I. C. P. *FEBS Lett.* **1973**, *37*, 27–31.
(76) Because of a more strained cis structure, which includes factors of different origins, the already published data on *cyclo*-(L-Pro)$_3$ and *cyclo*-(L-Pro$_2$-L-4Hyp) have been purposely disregarded for comparison, although a particular N-exo conformation was deduced from NMR: Deber, C. M.; Torchia, D. A.; Blout, E. R. *J. Am. Chem. Soc.* **1971**, *93*, 4893–4897. The crystal structure of these products (ref 39) unfortunately shows disordered distribution between a L-Pro and L-4Hyp residue preventing us from using the thermal parameters of the L-4Hyp molety for comparison.
(77) Dorman, D. E.; Bovey, F. A. *J. Org. Chem.* **1973**, *38*, 2379–2383.

BC(DEF) Parameters. 1. The Intrinsic Dimensionality of Intermolecular Interactions in the Liquid State†

Richard D. Cramer, III

Contribution from the Department of Chemistry, Research and Development, Smith Kline and French Laboratories, Philadelphia, Pennsylvania 19101.
Received June 11, 1979

Abstract: An average 95.7% of the variances in six physical properties (aqueous solvation energy, partition coefficient, boiling point, and molar refractivity, volume, and vaporization enthalpy) of 114 diverse pure liquid compounds is a linear function of two "**BC**" parameters characteristic of the compound, and derived by factor (principal components) analysis. The "**BC**" parameters are independent of the data set used in their derivation and can be identified with the "bulk" and "cohesiveness" of an individual molecule. Other physical properties which are well correlated ($r^2 > 0.9$) by the **BC** parameters include magnetic susceptibility, van der Waals' A and B, and critical temperature. Minor "**DEF**" parameters, also derived from the factor analysis, correlate further, to effect slight but significant reductions in the "unexplained" variance of critical pressure, surface tension, log (viscosity), solubility parameter, compressibility, and solvatochromic effects, as well as of the above properties. The "**BC(DEF)**" parameters are well correlated ($r^2 > 0.8$) with all the above 16 properties, except critical pressure, but only moderately correlated ($0.5 < r^2 < 0.8$) with thermal conductivity, dielectric constant, critical pressure, and the unrelated properties dipole moment, melting point, and molecular weight.

A fundamental objective in scientific research is the discovery of unifying relationships among a body of data. Ideally such relationships in chemistry are developed from a theoretical model of molecular behavior, such as the ideal gas law or the Schrodinger equation. However, some useful concepts, for example, the Hammett equation, have a primarily empirical rationale.

Noncovalent inter- and intramolecular forces, while of general chemical interest, are particularly important in the highly organized quasi-liquid structures and phenomena which are so characteristic of living organisms. Thus biochemical researchers have been prominent in demonstrating the unmistakable regularities in the liquid-state properties of molecules, particularly organic ones. Empirical additive-constitutive schemes give a completely adequate accounting of properties as diverse as partition coefficient,[1] boiling point,[2] molar volume,[3] and magnetic susceptibility.[4] Such regularities argue that, despite the potentially complex nature of intermolecular interactions within liquids, it may be that only a few relatively simple types of interaction are actually responsible for the major differences among the properties of compounds. An improved understanding of these regularities should help in the solution of a variety of biochemical problems, including the very practical challenges of drug design.[5]

Of the many mathematical methods which have been proposed for seeking regularities in chemical data,[6] the approach which best reveals the *intrinsic* linear structure of a data set is factor analysis,[7] particularly its subset methods such as principal components analysis or the Karhunen–Loeve transform. These techniques, first developed and long used in psychometrics, have recently been applied in chemical contexts to the derivation of substituent constants,[8] NMR shifts,[9] odor perception,[10] and structure/biological potency correlation.[11] Weiner has considered liquid-state properties as possible fundamental descriptors for sets of factorizable chromatographic data,[12a] and the partition coefficients of various solutes in various solvents have been factored by several groups.[12b] However, the factorization of an extremely varied set of liquid-state properties does not seem to have been attempted previously.

† Presented in preliminary form at the Great Lakes Central Section American Chemical Society Meeting, May 25, 1978.

0002-7863/80/1502-1837$01.00/0

Table I. The Data Table on Which Factor Analyses Were Performed[a]

ID#	COMPOUND NAME	B6	C6	D6	E6	F6	1 AC	2 PC	3 MR	4 BP	5 MV	6 HVP	7 X	8 CT	9 VDW A	10 VDW B	11 E	12 SLP	13 CP	14 STN	15 TCD	16 VIS	17 CMP	18 ET	19 U	20 MP	21 MW	
1	METHANE *	-0.380	-0.062	-0.018	0.027	0.014	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	1
2	ETHANE *	-0.256	-0.101	-0.040	0.005	0.015	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	2
3	PROPANE *	-0.163	-0.143	-0.024	-0.009	0.006	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	3
4	N-BUTANE *	-0.068	-0.179	-0.012	-0.025	0.005	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	4
5	2-METHYLPROPANE *	-0.073	-0.193	0.023	-0.024	-0.003	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	1	1	1	5
6	N-PENTANE	0.011	-0.209	0.005	-0.032	-0.004	1	1	1	1	1	1	1	0	0	1	1	1	0	1	0	0	0	0	1	1	1	6
7	2,2-DIMETHYLPROPANE	-0.014	-0.227	0.029	-0.018	-0.008	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	1	1	1	7
8	2,2-DIMETHYLBUTANE	0.082	-0.257	0.019	-0.033	0.006	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	8
9	CYCLOPENTANE	-0.006	-0.141	-0.043	-0.024	0.007	1	1	1	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	1	1	1	9
10	CYCLOHEXANE *	0.062	-0.173	-0.035	-0.020	0.001	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	10
11	ETHYLENE *	-0.242	-0.085	0.043	0.008	0.005	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	1	1	1	11
12	PROPYLENE *	-0.154	-0.108	0.015	0.006	0.000	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	12
13	1-BUTENE	-0.066	-0.140	0.005	0.000	0.004	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	13
14	2-METHYLPROPENE	-0.071	-0.138	0.010	0.003	0.001	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	14
15	CYCLOHEXENE						1	1	1	1	1	0	1	1	0	0	1	0	0	0	0	0	0	0	1	1	1	15
16	ACETYLENE	-0.280	0.026	-0.016	0.014	0.005	1	1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	1	1	16
17	PROPYNE	-0.183	0.009	-0.021	0.012	-0.007	1	1	1	1	1	1	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	17
18	1-PENTYNE						1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	18
19	BENZENE *	0.023	-0.045	-0.028	0.012	-0.007	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	19
20	TOLUENE *	0.098	-0.087	-0.016	0.013	-0.010	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
21	ETHYLBENZENE *	0.171	-0.123	-0.005	0.012	-0.009	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	21
22	O-XYLENE *	0.176	-0.090	-0.003	0.009	-0.013	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	22
23	M-XYLENE *	0.185	-0.116	-0.010	0.008	0.001	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	23
24	P-XYLENE *	0.183	-0.115	-0.006	0.009	-0.002	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	0	1	1	1	24
25	PROPYLBENZENE *	0.251	-0.158	0.004	0.007	0.001	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	25
26	2-PROPYLBENZENE *	0.245	-0.164	0.006	0.007	0.001	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	26
27	N-BUTYLBENZENE	0.325	-0.200	0.013	0.007	0.006	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	27
28	T-BUTYLBENZENE	0.309	-0.197	0.022	0.015	0.006	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	28
29	NAPHTHALENE *	0.322	-0.085	-0.029	0.022	0.006	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	29
30	ANTHRACENE						1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	30
31	PHENANTHRENE						1	1	0	1	0	1	1	0	0	0	0	0	1	0	0	0	0	0	1	1	1	31
32	FLUOROMETHANE	-0.298	0.026	-0.021	0.004	0.002	1	1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	1	1	32
33	CHLOROMETHANE *	-0.201	0.015	-0.022	-0.002	-0.003	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	33
34	BROMOMETHANE	-0.159	0.012	-0.038	0.007	-0.005	1	1	1	1	1	1	1	1	0	0	1	1	1	0	0	0	0	0	1	1	1	34
35	IODOMETHANE	-0.092	-0.009	-0.056	0.019	-0.009	1	1	1	1	1	1	0	1	0	0	0	0	1	1	0	1	1	0	1	1	1	35
36	CHLOROETHANE	-0.117	-0.014	-0.015	-0.006	-0.002	1	1	1	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	1	1	1	36
37	BROMOETHANE	-0.077	-0.019	-0.025	0.000	-0.008	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	0	1	1	1	1	37
38	IODOETHANE	-0.009	-0.035	-0.040	0.010	-0.009	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	1	0	1	1	1	1	38
39	1-CHLOROPROPANE *	-0.032	-0.054	-0.016	-0.016	0.002	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	1	1	1	39
40	2-CHLOROPROPANE	-0.044	-0.061	0.004	-0.008	-0.005	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	40
41	1-BROMOPROPANE	0.005	-0.050	-0.021	-0.008	-0.005	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	1	1	1	41
42	1-CHLOROBUTANE	0.045	-0.097	-0.011	-0.019	0.001	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	42
43	CHLOROBENZENE *	0.114	-0.077	-0.041	0.011	-0.010	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	43
44	1,2-DICHLOROBENZENE	0.212	-0.086	-0.058	0.003	0.000	1	1	1	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	1	1	1	44
45	1,3-DICHLOROBENZENE	0.201	-0.102	-0.051	0.006	-0.006	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	1	1	45
46	1,4-DICHLOROBENZENE	0.210	-0.104	-0.043	0.002	-0.004	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	1	1	46
47	BROMOBENZENE *	0.163	-0.067	-0.050	0.014	-0.003	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	47
48	DIMETHYLETHER *	-0.181	0.068	0.044	0.013	-0.014	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	48
49	DIETHYLETHER *	-0.032	0.011	0.082	0.006	-0.021	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	49
50	DIPROPYL ETHER	0.124	-0.086	0.097	0.001	-0.019	1	1	1	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	1	1	1	50
51	TETRAHYDROFURAN						1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	1	1	1	51
52	ANISOLE	0.144	-0.035	-0.016	-0.003	-0.029	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	1	1	52
53	METHANOL *	-0.143	0.253	-0.034	-0.018	-0.001	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	53
54	ETHANOL *	-0.076	0.218	-0.012	-0.018	0.004	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	54
55	1-PROPANOL *	-0.002	0.178	-0.002	-0.018	0.015	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	55
56	2-PROPANOL	-0.017	0.179	0.017	-0.012	0.010	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	56
57	1-BUTANOL	0.065	0.136	0.009	-0.023	0.021	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0	1	1	1	1	57
58	2-BUTANOL *	0.051	0.138	0.025	-0.012	0.019	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	58
59	T-BUTYL ALCOHOL	0.038	0.138	0.046	-0.004	0.017	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	59
60	1-PENTANOL	0.152	0.099	0.014	-0.028	0.037	1	1	1	1	1	1	1	0	0	0	1	1	0	0	0	0	1	0	1	1	1	60
61	2-METHYL-2-BUTANOL	0.128	0.114	0.043	-0.020	0.027	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	61
62	1-HEXANOL	0.220	0.050	0.028	-0.023	0.039	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	1	0	1	1	1	62
63	1-OCTANOL	0.364	-0.030	0.046	-0.025	0.058	1	1	1	1	1	1	1	0	0	0	0	1	0	1	0	1	0	1	1	1	1	63
64	CYCLOHEXANOL	0.157	0.117	0.011	-0.003	0.012	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	1	1	64
65	PHENOL	0.160	0.158	-0.032	0.021	0.025	1	1	1	1	1	1	1	1	0	0	1	0	1	0	0	0	1	0	1	1	1	65
66	2-CRESOL	0.220	0.106	-0.018	0.020	0.029	1	1	1	1	1	1	1	1	0	0	1	0	1	0	0	0	0	0	1	1	1	66
67	4-CRESOL	0.245	0.123	-0.026	0.007	0.040	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	1	1	1	67
68	4-T-BUTYLPHENOL	0.426	0.005	0.075	0.027	0.031	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	68
69	ACETONE *	-0.087	0.150	0.040	0.006	-0.025	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	69
70	2-BUTANONE	-0.017	0.110	0.051	0.007	-0.022	1	1	1	1	1	1	1	1	0	0	1	1	1	1	0	0	0	0	1	1	1	70

[a] For every compound are shown (1) the factor analyses in which its properties were used, (2) its **B, C, D, E,** and **F** values in the six-property, 114-compound factorization, and (3) the existence or absence of an experimental value for each of its properties. All listed compounds were used in the four-property factorization, those for which **B6–F6** values are shown were used in the six-property analysis, and an asterisk following

An especially important validation of any empirical relationship among data is its predictive reliability. In a following paper,[13] the relationships derived in this paper are verified by showing that many liquid-state properties of many types of compounds can be predicted with 95% accuracy from chemical structure alone.

Initial Data. The Compound/Property Table. The compounds and properties on which the factor analyses were performed and the set of eigenvectors or **BCDEF** values resulting from the most important factor analysis all appear in Table I. Note that only the presence or absence of an experimental value is indicated in Table I, the actual values being relegated to the supplementary material. In constructing this compound/property table, the first decision was selection of compounds. As our greatest interest was in biologically relevant phenomena, we began by listing all compounds whose aqueous solvation or activity coefficient in water, AC = log ($c_{H_2O_{soln}}/c_{vapor}$), had been reported or could be computed[14] from experimental data and whose partition coefficient, expressed as log ($c_{octanol}/c_{water}$), also had been measured.[15]

As can be seen in Table I, these criteria yielded a considerable diversity of structure, ranging in size from helium and hydrogen to anthracene and in polarity from alkanes to amides and water. Almost all major classes of functionality are encountered at least once. The set is not representative of chemistry as a whole; there are no ionic structures and few

Table I (*Continued*)

ID#	COMPOUND NAME	B6	C6	D6	E6	F6	1 AC	3 MR	5 MV	7 X	9 VDW A	11 E	13 CP	15 TCD	17 CMP	19 U	21 MW	
							2 PC	4 BP	6 HVP	8 CT	10 VDW B	12 SLP	14 STN	16 VIS	18 ET	20 MP		
71	2-PENTANONE						1 1	1 1	1 1	1 0	0 0	0 1	1 0	0 0	0 0	0 0	1 1	71
72	BENZALDEHYDE	0.170	0.081	-0.010	0.013	-0.010	1 1	1 1	1 1	1 0	0 0	0 1	0 0	1 0	1 0	0 0	1 1	72
73	ACETOPHENONE	0.232	0.066	0.014	0.029	-0.015	1 1	1 1	1 1	1 0	0 0	0 1	0 0	1 0	1 0	1 0	1 1	73
74	ACETIC ACID *	-0.039	0.263	-0.023	-0.009	0.008	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	74
75	PROPIONIC ACID						1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 0	1 0	1 1	75
76	BUTYRIC ACID	0.114	0.190	0.002	-0.019	0.020	1 1	1 1	1 1	1 0	1 0	0 0	1 0	1 0	1 0	0 1	1 1	76
77	METHYL ACETATE *	-0.066	0.113	0.037	0.003	-0.019	1 1	1 1	1 1	1 1	1 1	1 1	1 1	0 0	0 0	0 1	1 1	77
78	ETHYL ACETATE *	0.007	0.068	0.053	0.000	-0.015	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 0	0 0	0 1	1 1	78
79	PROPYL FORMATE *	0.004	0.050	0.044	-0.004	-0.023	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 0	0 0	0 1	1 1	79
80	ETHYL PROPIONATE *	0.075	0.026	0.066	-0.001	-0.015	1 1	1 1	1 1	1 1	1 1	1 0	1 1	0 0	0 0	0 1	1 1	80
81	METHYL BENZOATE	0.258	0.035	0.018	0.020	-0.001	1 1	1 1	1 1	1 0	0 0	0 1	0 0	0 0	0 0	0 1	1 1	81
82	ETHYLAMINE	-0.122	0.157	0.040	0.030	0.004	1 1	1 1	1 1	1 0	1 1	1 0	1 1	1 0	0 0	0 1	1 1	82
83	PROPYLAMINE	-0.045	0.115	0.047	0.031	0.004	1 1	1 1	1 1	1 0	1 1	1 0	0 1	0 0	0 0	0 1	1 1	83
84	BUTYLAMINE						1 1	1 1	1 1	1 0	1 1	1 0	0 1	0 0	0 0	0 1	1 1	84
85	PENTYLAMINE						1 1	1 1	1 1	1 0	1 1	1 0	0 1	0 0	0 0	0 1	1 1	85
86	HEXYLAMINE						1 1	1 1	1 1	1 0	1 1	1 0	0 1	0 0	0 0	0 1	1 1	86
87	DIETHYLAMINE *	0.005	0.075	0.092	0.035	-0.010	1 1	1 1	1 1	1 0	0 1	1 1	1 0	1 0	1 0	0 1	1 1	87
88	DIPROPYLAMINE						1 1	1 1	1 1	1 0	0 1	1 1	1 0	1 0	0 0	0 1	1 1	88
89	DIBUTYLAMINE						1 1	1 1	1 1	1 0	1 1	1 0	0 0	0 0	0 0	0 1	1 1	89
90	PYRROLIDINE						1 1	1 1	1 1	1 0	1 1	1 0	0 0	0 0	0 0	0 1	1 1	90
91	PIPERIDINE	0.068	0.108	0.051	0.039	-0.002	1 1	1 1	1 1	1 0	1 1	1 0	0 0	0 0	0 0	0 1	1 1	91
92	TRIMETHYLAMINE	-0.087	0.074	0.082	0.033	-0.003	1 1	1 1	1 1	1 0	0 1	1 1	1 0	0 0	0 0	0 1	1 1	92
93	TRIETHYLAMINE						1 1	1 1	1 1	1 0	1 1	1 1	1 0	1 0	0 0	0 1	1 1	93
94	ACETAMIDE	0.079	0.423	-0.029	-0.030	0.001	1 1	1 1	1 1	1 1	1 0	0 0	1 0	0 0	0 0	0 1	1 1	94
95	METHYL ACETAMIDE						1 1	1 1	1 1	1 0	0 0	0 0	0 0	0 0	0 0	0 1	1 1	95
96	DIMETHYL ACETAMIDE						1 1	1 1	1 1	1 0	0 0	0 0	0 1	0 0	0 0	0 1	1 1	96
97	ACETONITRILE *	-0.121	0.194	-0.021	-0.017	-0.030	1 1	1 1	1 1	1 1	1 1	1 0	1 1	1 0	1 0	1 1	1 1	97
98	PROPIONITRILE *	-0.050	0.154	-0.002	-0.015	-0.023	1 1	1 1	1 1	1 1	1 1	1 0	1 1	0 0	0 0	0 1	1 1	98
99	NITROETHANE	-0.042	0.154	-0.013	-0.049	-0.027	1 1	1 1	1 1	1 1	1 0	0 0	0 1	0 1	0 0	0 1	1 1	99
100	1-NITROPROPANE	0.018	0.109	0.005	-0.035	-0.010	1 1	1 1	1 1	1 1	1 0	0 0	0 0	0 0	0 0	0 1	1 1	100
101	NITROBENZENE	0.203	0.076	-0.035	0.004	-0.012	1 1	1 1	1 1	1 1	1 0	0 0	1 1	0 1	1 1	1 1	1 1	101
102	2-NITROTOLUENE	0.256	0.024	-0.018	0.010	-0.016	1 1	1 1	1 1	1 1	1 0	0 0	1 0	0 0	0 0	0 1	1 1	102
103	3-NITROTOLUENE	0.259	0.012	-0.022	0.012	-0.026	1 1	1 1	1 1	1 1	1 0	0 0	0 0	0 0	0 0	0 1	1 1	103
104	PYRIDINE	0.037	0.139	0.003	0.025	-0.007	1 1	1 1	1 1	1 1	1 1	0 0	0 1	0 0	0 0	0 1	1 1	104
105	2-METHYLPYRIDINE						1 1	1 1	1 1	1 0	1 1	0 0	1 0	0 0	0 0	0 1	1 1	105
106	3-METHYLPYRIDINE						1 1	1 1	1 1	1 0	1 1	0 0	0 0	0 0	0 0	0 1	1 1	106
107	4-METHYLPYRIDINE						1 1	1 1	1 1	1 0	1 1	0 0	0 0	0 0	0 0	0 1	1 1	107
108	2-ETHYLPYRIDINE						1 1	1 1	1 1	1 0	0 0	0 0	0 0	0 0	0 0	0 1	1 1	108
109	2,6-DIMETHYL PYRIDINE						1 1	1 1	1 1	1 0	1 1	0 0	0 0	0 0	0 0	1 0	1 1	109
110	THIOPHENOL	0.182	-0.008	-0.046	0.016	0.004	1 1	1 1	1 1	1 1	0 0	0 0	0 0	0 0	0 0	0 1	1 1	110
111	DIETHYLSULFIDE *	0.066	-0.042	0.024	0.011	-0.015	1 1	1 1	1 1	1 1	1 1	1 1	1 0	0 1	0 0	0 0	1 1	111
112	THIOANISOLE						1 1	1 1	1 1	1 0	1 0	0 0	0 0	0 0	0 0	0 1	1 1	112
113	1,3-BUTADIENE	-0.083	-0.099	0.008	0.018	-0.003	1 1	1 1	1 1	1 1	1 0	0 0	0 0	0 0	0 0	0 1	1 0	113
114	1,4-PENTADIENE						1 1	1 1	1 1	1 0	1 0	0 0	0 0	0 0	0 0	0 1	1 1	114
115	TRIFLUOROMETHANE						1 1	0 1	0 1	0 1	0 0	0 0	0 1	0 0	0 0	0 1	1 1	115
116	TETRAFLUOROMETHANE	-0.313	-0.121	0.025	-0.038	-0.001	1 1	1 1	1 1	1 0	1 0	0 0	0 1	0 0	0 0	0 1	1 1	116
117	DICHLOROMETHANE	-0.095	0.034	-0.042	-0.009	0.001	1 1	1 1	1 1	1 1	1 1	0 0	1 1	1 1	1 1	1 1	1 1	117
118	TRICHLOROMETHANE *	-0.029	-0.027	-0.033	0.000	-0.002	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	118
119	TETRACHLOROMETHANE *	0.032	-0.111	-0.033	0.000	-0.002	1 1	1 1	1 1	1 1	1 1	1 1	1 0	1 1	1 1	1 1	1 1	119
120	CHLORODIFLUOROMETHANE	-0.209	-0.027	-0.033	0.000	-0.002	1 1	1 1	1 1	1 1	1 1	0 0	0 0	1 0	0 0	0 1	1 1	120
121	CHLOROTRIFLUOROMETHANE	-0.229	-0.131	0.026	-0.040	-0.003	1 1	1 1	1 1	1 1	1 1	0 0	0 0	1 0	0 0	0 1	1 1	121
122	DICHLORODIFLUOROMETHANE	-0.131	-0.139	0.021	-0.024	-0.002	1 1	1 1	1 1	1 1	1 1	1 0	0 0	1 0	0 0	0 1	1 1	122
123	BROMOTRIFLUOROMETHANE	-0.160	-0.138	0.054	-0.031	0.001	1 1	1 1	1 1	1 0	1 0	0 0	0 1	0 0	0 0	0 1	0 1	123
124	1,1-DIFLUOROETHANE	-0.180	0.007	0.011	-0.030	-0.003	1 1	1 1	1 1	1 0	1 0	0 0	0 0	0 0	0 0	0 1	1 1	124
125	1,1-DICHLOROETHANE	-0.036	-0.029	-0.016	-0.003	-0.009	1 1	1 1	1 1	1 0	1 1	1 0	1 1	1 0	0 0	0 1	1 1	125
126	1,1,1-TRICHLOROETHANE	0.034	-0.086	-0.017	-0.004	-0.001	1 1	1 1	1 1	1 0	0 0	0 0	0 0	0 0	0 0	0 1	1 1	126
127	1,1-DICHLOROTETRAFLUOROETHANE	-0.041	-0.200	0.041	-0.043	-0.006	1 1	1 1	1 1	1 0	1 0	0 0	0 0	0 0	0 0	0 0	1 1	127
128	TRICHLOROETHYLENE	0.025	-0.058	-0.037	-0.004	-0.009	1 1	1 1	1 1	1 1	1 0	0 0	1 1	0 1	0 0	0 0	1 1	128
129	ALLYL ALCOHOL	-0.013	0.196	-0.015	-0.014	0.016	1 1	1 1	1 1	1 1	1 1	1 0	0 1	1 1	1 1	1 0	0 1	129
130	1,2-DICHLOROETHANE	-0.020	0.019	-0.023	-0.002	-0.016	1 1	1 1	1 1	1 0	0 1	1 0	0 0	1 1	0 0	1 1	1 1	130
131	1,4-DIOXANE	-0.008	0.182	0.064	0.026	-0.037	1 1	1 1	1 1	1 1	1 1	1 0	0 1	1 1	0 0	0 1	1 1	131
132	2-METHYLPYRAZINE						1 1	1 1	1 1	1 0	0 0	0 0	0 0	0 0	0 0	0 1	0 1	132
133	1,3-DICHLOROPROPANE						1 1	1 1	1 1	1 0	0 0	0 0	0 0	0 0	0 0	0 1	1 1	133
134	ETHYLENE GLYCOL						1 1	1 1	1 1	1 1	1 0	0 0	1 1	0 1	1 1	1 1	1 1	134
135	HYDROGEN *	-0.505	-0.055	-0.013	0.052	0.033	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 0	1 1	0 1	1 1	135
136	HELIUM *	-0.511	-0.074	0.026	0.031	0.032	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 0	1 1	0 0	1 1	136
137	ARGON *	-0.420	-0.049	-0.014	0.025	0.017	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 0	1 0	0 0	1 1	137
138	WATER *	-0.161	0.362	-0.080	-0.032	-0.005	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	138
139	IODINE	0.085	0.042	-0.173	0.032	-0.012	1 1	1 1	1 1	1 1	1 1	1 0	0 1	0 1	0 0	0 0	1 1	139
140	NITROGEN *	-0.420	-0.074	-0.001	0.022	0.018	1 1	1 1	1 1	1 1	1 1	1 1	1 0	1 0	0 0	0 0	1 1	140
141	OXYGEN	-0.418	-0.053	-0.021	0.025	0.018	1 1	1 1	1 1	1 0	1 1	1 1	1 0	1 0	0 0	0 0	1 1	141
142	NITROUS OXIDE						1 1	0 1	0 1	1 1	1 1	1 1	1 0	1 0	0 0	0 0	1 1	142

a name indicates a compound involved in the ten-property factorization. The 21 columns containing 1's or 0's indicate the existence or absence of an experimental value for the compound, for each of the 21 properties. Table IV shows the names and units of these properties and the supplementary material includes a table of the actual property values.

multiply functionalized or higher molecular weight substances. Nevertheless, the list would seem to comprise a remarkably broad basis for our main interest, molecules which exhibit a liquid phase under some experimentally convenient set of conditions.

The criterion for adding another property column to this table was usually the existence of a reasonably complete tabulation in the current Handbook of Chemistry and Physics.[16] The only other liquid-state properties which were available for all 138 compounds were boiling point (bp) and molar refractivity (MR), the latter by calculation from n_D and molar volume.[17] Liquid-state molar volumes (MV) and heats of vaporization (ΔH_{vap}) were tabulated or calculable from experimental data for the majority of these compounds. However, preliminary factor analyses suggested that for a few of the compounds these six property values were inconsistent with one another. Although these deviations did not affect the factor analyses themselves, their presence distorted the additive-constitutive relationships which were subsequently derived,[13] and so these compounds were omitted. Those 114 compounds for which the six properties 1–6 were both known and self-consistent are indicated in Table I by the presence of entries in their **B, C, D, E,** and **F** columns.

The remaining property values are sparser. Measurements of four more properties existed for 44 of the 138 compounds: molar magnetic susceptibility (X), critical temperature (CT),

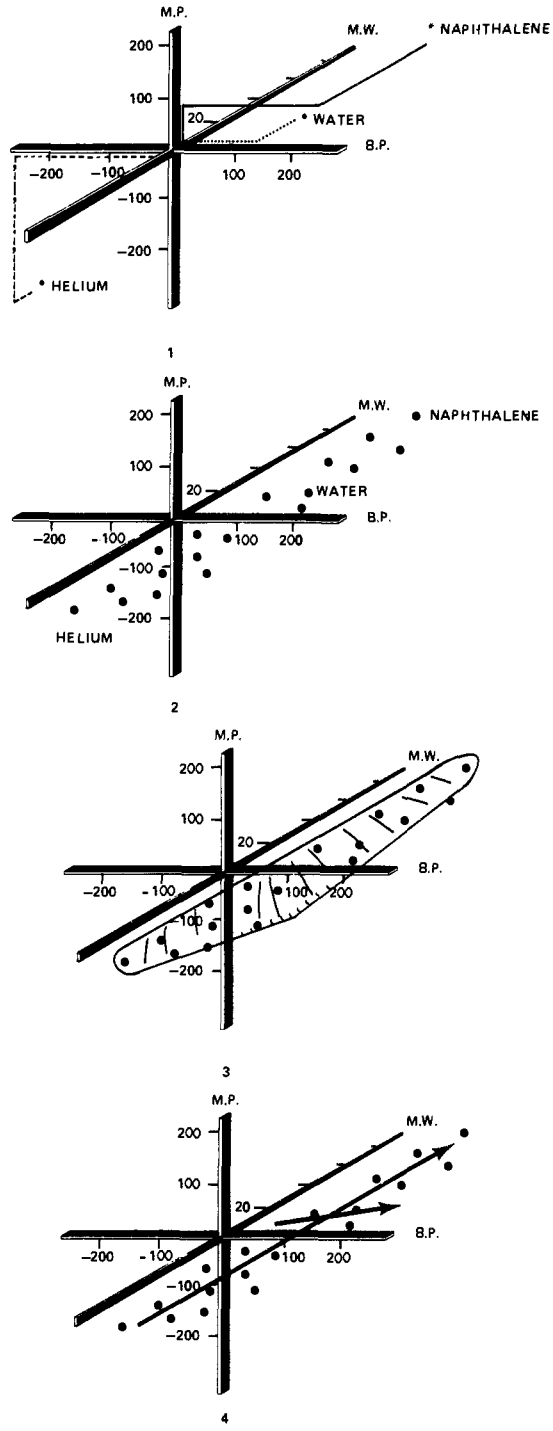

Figure 1. Geometrical representation of a simple factor analysis. Plate 1 depicts the three named compounds in a three-dimensional property space defined by their molecular weight and melting and boiling points, Plate 2 represents additional unnamed compounds. Plate 3 shows a planar figure in which the majority of the compound points lies. Plate 4 shows a pair of axes that might be used to locate compounds within the planar, or "reduced dimensionality", figure.

and van der Waals' A and B (VDW-A and -B) constants. (Because of the form of van der Waals' equation, the A value is actually studied herein as its square root.) These 44 compounds, indicated by an asterisk following their name in Table I, tend structurally to be more homologous than those in Table I as a whole.

Table I also indicates the existence of experimental values for 11 other properties not involved in the factor analyses themselves, specifically log (dielectric constant), solubility parameter, critical pressure, surface tension, thermal conductivity, log (viscosity), compressibility, E_T, the effect of the liquid as solvent upon an electronic transition,[18] gas-phase dipole moment (μ), melting point (mp), and molecular weight (mol wt). (The logarithm transform was applied as noted, when values of a property spanned more than two decades, simply to give a better fit.) These properties were excluded from factor analyses either because of their low relevance to the liquid state or an inadequate number of existing values.

Table II. Correlation Coefficients (r Values) between Pairs of the Properties Used in the Six-Property, 114-Compound Analysis

	six-property factorization (114 compounds)					
	AC	AC	MR	bp	MV	H_{vap}
1. AC	1.000	−0.450	0.219	0.668	0.080	0.764
2. PC		1.000	0.733	0.316	0.731	0.189
3. MR			1.000	0.810	0.914	0.730
4. bp				1.000	0.669	0.965
5. MV					1.000	0.593
6. ΔH_{vap}						1.000

Missing values in Table I obviously preclude a factor analysis using all 138 compounds and all 21 of the tabulated properties. The compromise which was made was to perform three analyses instead of one; a compound-intensive factorization involving the four properties known for all 138 compounds, a property-intensive factorization involving the 44 compounds for which ten properties were known or calculable, and an intermediate analysis involving 114 compounds and six properties. While attention will ultimately be limited to the intermediate analysis, where necessary these three analyses will be distinguished by the subscripts 4, 10, and 6, respectively.

Factor analysis of a table or matrix is valuable to the extent that there are linear interrelationships among several columns of data. Many linear relationships between *single* columns have of course already been reported for the properties of Table I.[19] A classical example is Trouton's rule, and a more topical example is a correlation between partition coefficient and molar volume that has been used to adduce the "hydrophobic effect".[20] Actually, the relatively low correlation coefficients in Table II suggest that many of these relationships are highly linear only for compound subsets, usually involving homologous series. Even the highest correlation coefficient, Trouton's rule ($r^2 = 0.931$) relating bp and ΔH_{vap}, represents a collinearity lower than the 96% planarity of these factor analyses. Although molar refractivity is highly correlated with molar volume and boiling point, none of the other 12 possible simple property correlations has a value of r^2 greater than 0.6. Thus *the pairwise correlations are much lower than the overall "complex correlation"* which factor analysis will show to exist within these data.

Geometric Description of Factorization. Factor analysis is an operation that can be described and visualized geometrically, provided that only three properties are considered. Figure 1 shows the progress of a hypothetical factor analysis involving three properties, mol wt, bp, and mp, plotted along the x, y, and z axes, respectively. (These properties were chosen for their familiarity and overall are not as collinear as those in the actual factor analyses.) The first panel of Figure 1 depicts the positioning of the points for three molecules, helium, naphthalene, and water, in the mol wt/bp/mp space. For example, water's position results from moving +18 units along the mol wt axis and +100 along the bp axis, while remaining stationary at 0 on the mp axis. Points for a number of other unnamed compounds have been added to the second panel of Figure 1. It can be seen that the additional points are certainly not scattered at random through the space and in fact tend to fall in a single tilted plane, as indicated in the third panel of Figure 1.

The simple collinearities, or r^2, associated with the three

possible pairings of these properties could be visualized by sighting along each of the three axes. Notice how even a very flat plane will not manifest itself as a collinearity between a pair of properties unless the plane is perpendicular to an axis.

To the extent that all the points fall into a plane, the position of an individual compound needs only two coordinates rather than the original three for its definition (along with knowledge of the existence of the plane). Choice of a particular new coordinate system is really arbitrary since any two nonparallel lines in the plane could serve as the new axes. However, in factor analysis it is usual to define the new axes initially so that the first axis is aligned along the longest data dimension and the second axis is perpendicular (orthogonal) to the first. The last panel of Figure 1 shows such a pair of axes for the mol wt/bp/mp data. These "natural" axes of a data set are sometimes called its "principal components".

Factor analysis is the statistical counterpart of the sequence in Figure 1. Analysis of the mol wt/bp/mp data shown would yield two significant "eigenvalues", corresponding to the two dimensions of the plane. Thus the "intrinsic dimensionality" of these three sets of property data is two. The set of coordinates for each compound point within the plane is called its eigenvectors. Equations for regenerating the original properties of any compound, given its eigenvectors, are also produced. Therefore, factor analysis may be described as an empirical method for seeking the simplest linear structure that exists within a set of data.

Interpretation of a factor analysis often is complicated by two secondary issues, both essentially judgmental. First is a decision about the "optimal" orientation of the axes, already alluded to. This type of issue might be exemplified by the debate in organic chemistry pitting F and R against δ_I and δ_R as descriptors of electronic effects.[8] Even more important, uncertainty can also arise as to the number of eigenvalues actually necessary to span the original data. For example, if the plane in Figure 1 happened to be squashed a little, perhaps by failing to include deviant compounds, there would be an argument for a linear data structure, having only one significant eigenvalue, rather than the planar structure. A conservative statistical rule of thumb attributes significance only to those components having eigenvalues greater than 1.0, i.e., explaining at least as much of the overall data variance as did one of the original columns. However, recognizing that chemical data are more precise than the psychometric data for which factor analysis was first developed, Malinowski has argued for relatively complex criteria,[21] which usually attribute significance to a much larger number of components. We follow Wold[22] in adopting an empirical and intermediate criterion for eigenvalue significance; those components are significant whose inclusion improves our ability to predict the properties of compounds not in the original data set.

Factorization of the Compound/Property Table. As discussed above, the data in Table I have been factored as three different subsets. The major results of the three factorizations appear in Table III, part A being the eigenvalues themselves and part B a comparison of the average ability of an increasing number of factors to reproduce increasing proportions of the variance among the original property data.

Factorization was performed using a slightly modified version of Weiner's program.[23] Before factorization each property was given an equivalent weight, or standardized, by subtracting the property's mean value and dividing the result by the property's standard deviation, so that the form of the final factorization will not be affected by the scale or units in which the properties are measured.[24] Many workers would describe this variety of factor analysis as a "principal components" analysis.

The first column of Table III, giving the results of factoring the matrix constructed from the four properties AC, PC, MR, and bp for all 138 compounds, shows that only two components are necessary to describe 97.5% of the variance in this data set. In other words, all of the 138 compound points in this four-dimensional space lie virtually within a single plane. The first "**B**" eigenvector, or factor having an eigenvalue of 2.23, will reproduce 55.8% of the original variance. Its combination with the second "**C**" eigenvector, eigenvalue = 1.67, will reproduce 97.5% of the original variance. The last two eigenvectors, which together span only 2.5% of the overall original variance, would be ignored, according to the above-mentioned rule of thumb, as probable experimental error and/or idiosyncratic properties of individual compounds.

A check on a factor analysis consists of using the eigenvectors and linear equations, reproduced for this four-property matrix within the supplementary material, to recalculate the original compound–property matrix. The results of this check are shown in part B of Table III. For example, the 138 boiling points are reproduced using the two **BC** factors with a root mean square (rms) deviation of ±18.3°C, whereas one factor gives an rms error of ±34.2°C and three factors give an rms error of ±5.6 °C. (A sample of a recalculation is shown for the most deviant of all experimental values in this matrix, the boiling point of water, in footnote *b* of Table III.)

The second factorization involves the largest number of properties, ten, but only 44 compounds. The reduction in dimensionality encountered in this factorization is particularly remarkable, just two dimensions again serving to encompass 96.7% of the variation in ten original dimensions, or properties. Even more interesting is a strong similarity between the second and first factorizations, despite the small overlap between the two data matrices involved, $2 \times 4 \times 44$ or 352 elements of a total of $(138 \times 4) + (10 \times 44)$ or 992 elements, about 35%. The similarity is immediately exhibited by the rms errors of recalculation for the four properties common to both factorizations (part B of Table II). The magnitudes of these average errors of recalculation, using either the **B** or **BC** eigenvectors, are within an average of less than 10% of each other, despite the small overlap among the data being reproduced.

The intermediate and most useful data matrix, based on the six properties of 114 compounds, can be seen in the last column of Table III to factor similarly to the two extreme matrices. The two largest eigenvectors span 95.7% of the original variance. The next eigenvector, although proportionally the largest of the three tabulated "**D**" eigenvectors, encompasses only 2.8% of the original variance. The errors in recalculations of the common properties are again of the same magnitude as those from the two preceding factorizations, whether the **B, BC, BCD,** or **BCDEF** factor sets are used. For this factorization only, the six equations used in recalculation are given at the top of Table IV.

The similarities among these three different sets of **BCDEF** vectors, suggested by the similarity of errors remaining after recalculation, are shown by their correlation coefficients (Table V) to be completely general. For example, the $\mathbf{B}_{10}$ and $\mathbf{C}_{10}$ vectors have collinearities of $r = 0.996$ and 0.967, respectively, with the $\mathbf{B}_4$ and $\mathbf{C}_4$ vectors, despite the overlap of only 35% between the two matrices. *Such collinearities among the results of predominantly independent analyses suggest an underlying similarity in the important mechanisms of intermolecular interaction.* All of the corresponding **B, C,** and **D** vector pairs have collinearities of $r = 0.938$ or higher in Table V (underlined r's), except the $\mathbf{D}_4$ vector, which as the next-to-the-last eigenvector of its matrix may include some idiosyncratic contributions from individual compounds. Visualization of the collinearities is facilitated by some sample plots of **BCD** vector pairs in Figure 2. In contrast, because the eigenvectors from a given factor analysis are defined to be perpendicular (cf. Figure 1 and its textual description), one ex-

Table III. Results of the Three Factorizations

no. of factors	matrix = four properties, 138 compds	matrix = ten properties, 44 compds	matrix = six properties, 114 compds
	A. Eigenvalues and, in Parentheses, Cumulative Percentage of Variance Reproduced[a]		
1. "B"	2.23 (55.8)	7.53 (75.3)	3.87 (64.4)
2. "C"	1.67 (97.5)	2.14 (96.7)	1.87 (95.7)
3. "D"	0.070 (99.3)	0.186 (98.6)	0.168 (98.5)
4. "E"	0.029 (100.0)	0.045 (99.0)	0.045 (99.2)
5. "F"		0.043 (99.4)	0.029 (99.7)
6.		0.025 (99.7)	0.017 (100.0)
7.		0.017 (99.9)	
8.		0.010 (99.97)	
9.		0.003 (99.99)	
10.		0.001 (100.0)	
	B. Rms of Errors in Recalculating Six of the Properties[b] for All Compounds, Using an Increasing Number of Factors		
	One Factor (**B**)		
AC (H_2O)	1.84	1.75	1.75
PC	1.06	1.04	1.01
MR	3.27	3.42	3.28
bp	34.22	41.44	34.76
MV		13.10	15.13
ΔH_{vap}		1.48	1.33
	Two Factors (**B, C**)		
AC (H_2O)	0.25	0.26	0.26
PC (octanol/H_2O)	0.15	0.31	0.25
MR	1.84	1.55	1.76
bp	18.30	13.18	19.75
MV		8.27	9.79
ΔH_{vap}		0.49	0.51
	Three Factors (**B, C, D**)		
AC (H_2O)	0.22	0.16	0.23
PC (octanol/H_2O)	0.13	0.10	0.11
MR	0.51	1.50	1.75
bp	5.57	6.27	13.92
MV		5.07	2.20
ΔH_{vap}		0.36	0.40
	Five Factors (**B, C, D, E, F**)		
AC		0.13	0.14
PC		0.10	0.07
MR		1.28	0.49
bp		3.05	3.21
MV		1.42	0.39
ΔH_{vap}		0.24	0.21

[a] Cumulative % variance is $(100/F)\sum_{n=f}^{n=F}\lambda_n$, where F is the number of properties in the matrix, n is a column number, f is the current column number, and λ_n is the nth eigenvalue. Its interpretation is identical with that of r^2 (×100) for a regression equation. [b] An example of property calculation is the boiling point of water, based on the results of the six-property, 114-compound factorization, for which the eigenvectors are given in Table I and the property equation in Table IV:

no. of factors used	calculation	result	error
1	66.39 + 532.5**B**[= 66.39 + 532.5(−0.161)]	= −19.3	−119.3
2	66.39 + 532.5**B** + 223.6**C**	= 61.6	−38.4
3	66.39 + 532.5**B** + 223.6**C** − 365.4**D**	= 90.8	−9.2
5	66.39 + 532.5**B** + 223.6**C** −365.4**D** −250.8**E** −794.6**F**	= 102.8	+2.8

(The boiling point of water is the single most deviant observation in Table I with respect to the one- and two-factor equations.)

pects noncorresponding eigenvectors to have zero intercorrelations. In Table V this would be strictly true only among the $\mathbf{B}_{10}$, $\mathbf{C}_{10}$, and $\mathbf{D}_{10}$ vectors, since from half to two-thirds of the elements in the other six vectors must be deleted before the numbers in Table V can be computed.

Since all three of the factor analyses give such similar results, we need henceforth consider only the intermediate six-property 114-compound factorization. This analysis seems the most useful because its compounds include the same amount of functional variety and almost the same size and polarity variation as does the four-property 138-compound matrix, while the additional columns allow its third and perhaps its fourth and fifth eigenvectors to contain a greater proportion of general trends than will the minor eigenvector(s) of the four-property matrix.

A decision about the number of factors needed to span the original data is not yet possible, although clearly the **B** and **C** factors are "real" and dominant. As indicated above, the statistical rule of thumb ($\lambda > 1$) suggests that the remaining factors **D, E,** and **F** are much too small to be meaningful. On the other hand, the large decreases in rms error of recalculation as the minor factors are added suggest that their inclusion might usefully improve predictions. Only the BCDE and BCDEFG models can be eliminated from further consideration, the latter because a **G** factor, as the sixth and last factor in the intermediate factorization, would perforce contain ex-

Table IV. Regression Equations Expressing 21 Properties[d] as Functions of a Compound's $\mathbf{B}_6$, $\mathbf{C}_6$, $\mathbf{D}_6$, $\mathbf{E}_6$, and $\mathbf{F}_6$ Values, Derived from the Data of Table I

1.[f]	activity coefficient log (c_{H_2O}/c_{gas})	= 1.241(±0.027) + 5.09(±0.14)**B** + 13.54(±0.21)**C** + 3.36(±0.70)**D** + 6.83(±1.35)**E** + 7.39(±1.69)**F**	114	0.998	0.14
2.[f]	log (partition coefficient) ($c_{octanol}/c_{H_2O}$)	= 1.604(±0.014) + 3.65(±0.08)**B** − 7.66(±0.11)**C** − 5.74(±0.37)**D** − 0.31(±0.71)**E** + 5.09(±0.90)**F**	114	0.998	0.08
3.[f]	molar refractivity	= 22.94(±0.09) + 52.89(±0.52)**B** − 21.58(±0.74)**C** + 4.21(±2.5)**D** + 83.6(±4.8)**E** − 12.02(±6.0)**F**	114	0.999	0.51
4.[f]	boiling point	= 66.39(±0.62 + 532.50(±3.3)**B** + 223.6(±4.8)**C** − 365.4(±15.9)**D** − 250.8(±31)**E** − 794.6(±39)**F**	114	0.9996	3.31
5.[f]	molar volume	= 90.02(±0.24) + 139.9(±1.3)**B** − 90.12(±1.8)**C** + 245.6(±6.1)**D** − 108.5(±12)**E** − 3.1(±14)**F**	114	0.999	1.27
6.[f]	heat of vaporization	= 8.197(±0.041) + 14.92(±0.22)**B** + 9.615(±0.32)**C** − 8.07(±1.08)**D** − 12.8(±2.1)**E** + 14.5(±2.6)**F**	114	0.997	0.22
7.	magnetic susceptibility	= 54.58(±1.3) + 111.4(±7.2)**B** − 51.8(±9.9)**C** − 16.1(±34.6)**D** + 42.9(±70)**E** + 13.0(±79)**F**	89	0.963	6.31
8.	critical temp	= 248.6(±4.7) + 770.4(±25)**B** + 343.1(±36)**C** − 927(±117)**D** − 68(±230)**E** − 1810(±280)**F**	83[a]	0.992	20.71
9.	(van der Waals A)$^{1/2}$	= 4.144(±0.062) + 6.93(±0.34)**B** − 0.23(±0.47)**C** − 0.03(±1.7)**D** + 1.2(±3.8)**E** − 10.6(±5.3)**F**	53	0.991	0.19
10.	van der Waals B	= 0.121(±0.003) + 0.225(±0.02)**B** − 0.07(±0.03)**C** + 0.15(±0.10)**D** + 0.11(±0.2)**E** − 0.12(±0.30)**F**	53	0.975	0.01
11.	log (dielectric constant)	= 0.728(±0.06) + 0.57(±0.30)**B** + 2.60(±0.43)**C** − 1.97(±1.38)**D** − 5.1(±3.2)**E** − 2.3(±3.8)**F**	66	0.883[e]	0.23
12.	solubility parameter	= 8.97(±0.23) + 5.22(±1.6)**B** + 12.1(±1.8)**C** − 15.0(+6.7)**D** − 12.9(±13)**E** − 7.7(±16)**F**	54[b]	0.912	0.80
13.	critical pressure	46.3(±2.3) − 7.1(±12.3)**B** + 56.2(±19)**C** − 243(±58)**D** − 51(±118)**E** − 230(±141)**F**	76[b]	0.769	9.75
14.	surface tension	= 23.5(±0.9) + 36.8(±5.3)**B** + 22.2(±7.4)**C** − 101(±25)**D** + 58(±51)**E** − 174(±57)**F**	46[b]	0.953	2.55
15.	thermal conductivity	= 3.17(±0.23) + 3.19(±2.0)**B** + 4.5(±1.5)**C** + 2.1(±5.6)**D** − 13.2(±21)**E** − 14.5(±19)**F**	29[c]	0.808	0.43
16.	log (viscosity)	= −0.23(±0.08) + 2.44(±0.46)**B** + 1.26(±0.63)**C** − 0.94(±1.9)**D** − 8.1(±4.5)**E** + 5.06(±4.9)**F**	37	0.920	0.19
17.	isothermal compressibility	= 11.6(±0.7) − 22.6(±6.0)**B** − 10.8(±4.9)**C** + 57.6(±14)**D** + 11(±43)**E** + 22(±38)**F**	25	0.935	1.19
18.	E_T	= 38.4(±1.1) + 9.4(±10)**B** + 47.9(±7.2)**C** − 41.7(±21)**D** − 147(±71)**E** + 63(±59)**F**	25	0.980	1.73
19.	dipole moment	= 1.33(±0.15) + 1.5(±0.8)**B** + 5.0(±1.1)**C** − 1.5(±3.8)**D** − 10(±7.2)**E** − 20.4(±9)**F**	110	0.733	0.77
20.	melting point	= 77.2(±7.4) + 295(±40)**B** + 194(±58)**C** − 443(±194)**D** + 339(±376)**E** − 406(±471)**F**	112	0.851	39.6
21.	molecular weight	= 87.8(±4.8) + 149.6(±26)**B** − 78.3(±37)**C** − 239(±125)**D** + 20(±241)**E** − 386(±303)**F**	114	0.779	25.8

A sample calculation using eq 4 appears in Table III. The number of compounds, correlation coefficient, and standard error of regression appear after each equation as n, r, and s. Parenthesized numbers following a value are its 95% confidence intervals. [a] Excluding 1,1-difluoroethane. [b] Excluding water. [c] Excluding water and chloromethane. [d] Units of the equations: 1, 2, dimensionless; 3, cm^3 mol^{-1}; 4, °C; 5, cm^3 mol^{-1}; 6, kcal mol^{-1}; 7, cgs molar; 8, °C; 9, L atm$^{1/2}$ mol^{-1}; 10, L mol^{-1}; 11, dimensionless; 12, cal cm^{-3}; 13, atm; 14, dyn cm^{-1}; 15, cal s^{-1} cm^{-2} (cal/cm)$^{-1}$ × 10^4; 16, dimensionless; 17, m^2 mol^{-1} × 10^{10}; 18, see ref 18; 19, D; 20, °C; 21, g mol^{-1}. [e] If dipole moment is included as a sixth explanatory variable, the following equation is obtained (improvement significant at 99% level): log (dielectric constant = 0.459(±0.06) + 0.232(±0.04)μ + 0.19(±0.20)**B** + 1.52(±0.33)**C** − 2.17(±0.84)**D** − 4.2(±1.9)**E** + 0.2(±2.4)**F** (n = 64, r = 0.961, s = 0.14). However, dipole moment did not improve the equations for properties 12, 13, or 15. [f] Equations used to "recalculate" properties, in the right-hand column of Table IIIB.

Table V. Correlation Coefficients for the **BCDEF** Parameter Sets, Defined by the Three Factorization Experiments of Table III, Based on the 44 Compounds for Which All Ten Properties Are Known[a]

	$\mathbf{B}_4$	$\mathbf{C}_4$	$\mathbf{D}_4$	$\mathbf{B}_6$	$\mathbf{C}_6$	$\mathbf{D}_6$	$\mathbf{E}_6$	$\mathbf{F}_6$	$\mathbf{B}_{10}$	$\mathbf{C}_{10}$	$\mathbf{D}_{10}$	$\mathbf{E}_{10}$	$\mathbf{F}_{10}$
$\mathbf{B}_4$	1.000	−0.189	0.345	0.998	−0.161	−0.055	−0.205	−0.461	−0.996	0.027	−0.056	−0.024	−0.002
$\mathbf{C}_4$		1.000	0.256	−0.173	0.995	0.117	−0.333	−0.271	0.207	0.967	0.144	−0.003	−0.008
$\mathbf{D}_4$			1.000	0.330	0.318	−0.629	−0.833	−0.315	−0.321	0.427	−0.676	0.328	−0.108
$\mathbf{B}_6$				1.000	−0.149	−0.007	−0.226	−0.467	−0.998	0.037	−0.014	−0.037	−0.026
$\mathbf{C}_6$					1.000	0.021	−0.358	−0.242	0.185	0.980	0.057	−0.011	0.034
$\mathbf{D}_6$						1.000	0.235	−0.299	0.003	−0.023	0.938	−0.093	−0.329
$\mathbf{E}_6$							1.000	0.292	0.210	−0.433	0.345	−0.226	0.319
$\mathbf{F}_6$								1.000	0.476	−0.336	−0.217	−0.423	0.396
$\mathbf{B}_{10}$									1.000	0.000	−0.001	0.001	−0.002
$\mathbf{C}_{10}$										1.000	0.001	−0.002	0.002
$\mathbf{D}_{10}$											1.000	−0.001	0.001
$\mathbf{E}_{10}$												1.000	0.001
$\mathbf{F}_{10}$													1.000

[a] The individual $\mathbf{B}_6$, $\mathbf{C}_6$, $\mathbf{D}_6$, $\mathbf{E}_6$, and $\mathbf{F}_6$ values appear in Table I and the remaining **BCDEF** values in the Supplementary Material.

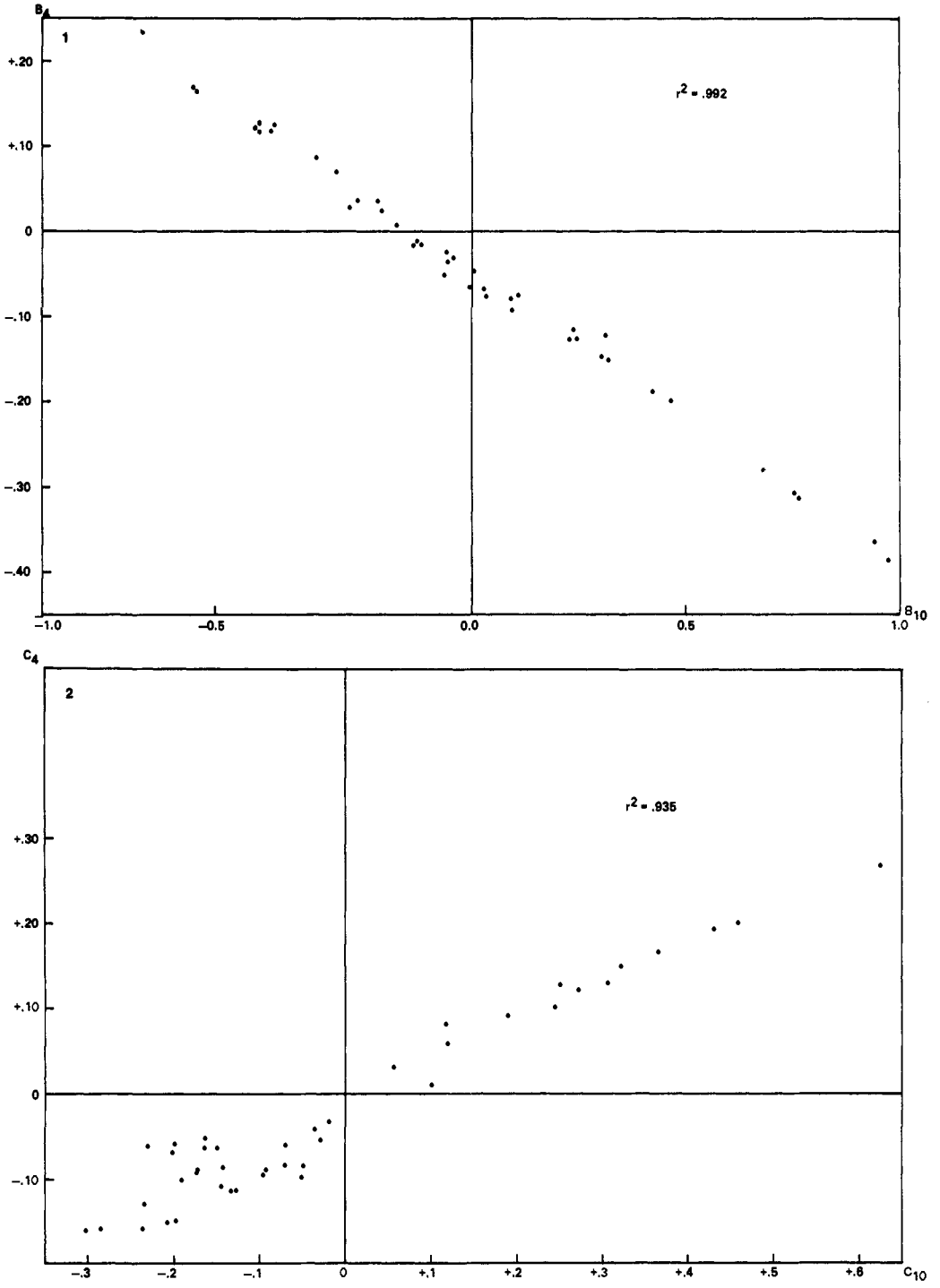

Figure 2. Collinearity between the eigenvectors derived from largely independent factor analyses of the property data. Plate 1 shows the first, or $\mathbf{B}_4$ eigenvector from the four-property 142-compound analysis ($\mathbf{B}_4$) plotted against the first eigenvector from the ten-property 44-compound analysis ($\mathbf{B}_{10}$). Plate 2 shows the plot of $\mathbf{C}_4$ against $\mathbf{C}_{10}$, where $\mathbf{C}_4$ and $\mathbf{C}_{10}$ are defined analogously.

perimental error, and the former because, given the similar magnitudes of the **E** and **F** eigenvalues, the argument that the **E** and not the **F** factor is meaningful would be difficult to sustain. Throughout this paper the **BC, BCD,** and **BCDEF** models are all considered to be viable candidates. The property prediction experiments[13] also are inconclusive, showing that the **BC** model is the least likely to give misleading predictions under any circumstance, but that the **BCDEF** model gives more accurate predictions for many properties of most compounds. To emphasize the continuing uncertainty as to the most desirable number of factors, when the **BCDEF** label is used generically, the **DEF** portion will be enclosed by parentheses.

The BCDEF Property Equations. In Table IV are listed regression equations which express 21 physical properties as a function of a compound's $\mathbf{B}_6$, $\mathbf{C}_6$, $\mathbf{D}_6$, $\mathbf{E}_6$, and $\mathbf{F}_6$ values, derived by least-squares analysis of the data indicated by Table I. Note that these equations are the ones used later[13] to predict the properties of compounds not included in Table I, the necessary **B, C, D, E,** and **F** values being obtained either from a subset of the compound's properties or from its structure alone. An example of the use of these equations appears at the foot of Table III, as mentioned above. In general, one should not do as that example suggests and attempt to predict a dependent variable using only some of the terms of a lengthy regression equation, because the variable sets from which the equation

Table VI. Relative Importance of the Individual Eigenvectors (**BCDEF** Values in Table I) in the Equations (Table IV) Which Fit the Sets of Values for 21 Properties[a]

property	*bB*	*cC*	weight of *dD*	*eE*	*fF*	property variance not fitted[b]
1. activity coeff	0.424	0.548	0.012	0.007	0.005	0.005
2. log (partition coeff)	0.474	−0.484	−0.033	−0.0005	0.005	0.004
3. molar refractivity	0.817	−0.162	0.003	0.015	−0.001	0.002
4. boiling point	0.799	0.163	−0.024	−0.004	−0.009	0.001
5. molar volume	0.713	−0.223	0.056	−0.006	−0.0001	0.002
6. heat of vaporization	0.735	0.230	−0.018	−0.007	0.005	0.005
7. magnetic susceptibility	0.749	−0.169	−0.005	0.003	0.001	0.073
8. critical temp	0.764	0.165	−0.041	−0.001	−0.013	0.016
9. (van der Waals A)$^{1/2}$	0.954	−0.016	−0.0002	0.002	−0.011	0.017
10. van der Waals B	0.803	−0.117	0.023	0.005	−0.003	0.049
11. log (dielectric const)[c]	0.223	0.493	−0.034	−0.023	−0.007	0.220
12. solubility parameter	0.362	0.409	−0.046	−0.010	−0.004	0.168
13. critical pressure	−0.088	0.340	−0.134	−0.007	0.021	0.409
14. surface tension	0.619	0.181	−0.076	0.011	−0.022	0.091
15. thermal conductivity	0.363	0.249	0.011	−0.018	−0.012	0.348
16. log (viscosity)	0.639	0.161	−0.011	−0.025	0.013	0.151
17. compressibility	−0.643	−0.150	0.073	0.0003	0.005	0.129
18. E_T	0.246	0.608	−0.048	−0.045	0.012	0.041
19. dipole moment	0.192	0.303	−0.008	−0.015	−0.019	0.463
20. melting point	0.513	0.164	−0.034	0.007	−0.005	0.277
21. molecular weight	0.451	−0.115	−0.032	0.001	−0.009	0.393

[a] Relative importances are computed as $x(s_X)^2/[(1/r^2)\sum_{X=B}^{X=F} x(s_X)^1]$ where x is a coefficient in the regression equation for that property in Table IV; r^2 is the correlation coefficient for the same equation; s_X is the standard deviation of the eigenvector in Table I. (These standard deviations are $s_B = 0.185$, $s_C = 0.129$, $s_D = 0.039$, $s_E = 0.020$, $s_F = 0.016$). [b] Equal to $1 - r^2$ for the equation in Table IV. [c] Note footnote *e* in Table IV. With dipole moment included, this line would read: *bB*, 0.074; *cC*, 0.288; *dD*, −0.037; *eE*, 0.004; *fF*, 0.001; unexplained, 0.076. Dipole moment "explains" 0.520 of the variance.

was derived are probably both intercorrelated and have nonzero means, implying that the best fitting value of any particular coefficient depends on the presence or absence of other terms in the equation. However, the **BCDEF** eigenvectors are defined to be orthogonal to one another and to have zero means, so the "best", or least-squares, two-term **BC** equation usually closely resembles the first two terms of the corresponding **BCDEF** equation in Table IV. The actual "best" two-term **BC** and three-term **BCD** equations appear in the supplementary material, as do complete lists of residuals, i.e., calculated–experimental values, for the **BC, BCD,** and **BCDEF** least-squares equations.

Casual inspection of the coefficients in Table IV might give the impression that the **D, E,** and **F** terms are almost as influential as the **B** and **C** terms in calculating property values. However, this impression would be based on the tacit assumption that **B, C, D, E,** and **F** parameters will have about the same range of values, whereas inspection of Table I shows the much greater numerical spread of **B** and **C** values. To give a clearer picture of the relative importance of the various terms, the coefficients in Table IV have been reweighted, to correct for differences in parameter set spreads, in the scale on which the property is measured, and in the equation's correlation coefficient, to produce Table VI. For example, the first line in Table VI indicates that 42.4% of the original variance in the aqueous activity coefficients of the substances in Table I is accounted for by the *bB* term of eq 1 in Table IV, 54.8% by the *cC* term, and only 1.2, 0.7, and 0.5% by the *dD*, *eE*, and *fF* terms. The final entry of the line indicates that 0.5% of the activity coefficient variance remains unexplained.

Table VI indicates again that the **B** and **C** terms are always very much more influential than the **D, E,** and **F** terms in explaining property values. However, comparison of the *dD*, *eE*, and *fF* columns with the "unexplained" column shows that the minor terms may often substantially reduce the amount of variance which the equation would otherwise leave unexplained. The usual statistical method for deciding on whether a term should be included in a regression equation is to compute the "unexplained" variance with and without the term. Comparison of the ratio of the two variances with tabulated F-test values then shows whether the improvement in fit is large enough to be an unlikely chance occurrence. These variance ratios, comparing the **BCDEF** equations with the **BCD** and **BC** equations and the **BCD** with the **BC** equations, appear in Table VII. From the asterisked entries, indicating statistically significant reductions in unexplained variance, it is apparent that the addition of a **D** term, followed by the **EF** terms, does effect significant reductions in the unexplained variance of seven and nine properties, respectively. Altogether, the use of the **DEF** parameters in combination significantly improves equation fit for 13 of 18 liquid-state properties, an improvement which is particularly notable for properties 11–18, since these are the properties which are less completely explained by the **BC** parameters alone. However, it should also be noted that the **D, E,** and **F** parameters are not random vectors, being first defined so as to force a reduction in the variance of the first six properties, and second defined orthogonally, so that the likelihood of chance correlations is probably a bit greater than F-test theory suggests. No final conclusion about the "reality" of the **D, E,** and **F** vectors is yet possible.

Assessment of the BC(DEF) Correlations. It is desirable to have some comparison of the data-fitting quality of the **BC(DEF)** parameters with other parameters that might be used to predict physical properties. For example, despite the relatively low correlation coefficients in Table II, the reader may still suspect that much of the high quality of the equations in Table IV is a trivial artifact of the dependency of many of these properties on molecular weight or molecular volume. In the first six columns of Table VIII, the r^2 (fraction of total variance fit) and s (variance not fit) for some of the equations in Table IV can be compared with r^2 and s for the corresponding molecular weight and molar volume equations. (Properties 17–21 were not tested because of too few values or low relevance to the liquid state.) It is evident that the **BC** parameters are invariably superior to either molecular weight or molecular volume in fitting the property data, usually by an enormous margin. In fact, comparison of the r^2 in Table VIII for molecular weight or molecular volume equations with the

Table VII. Comparison of the Data-Fitting Qualities of the Two-Factor "**BC**" Equation, the Three-Factor "**BCD**" Equation, and the Five-Factor "**BCDEF**" Equation[a]

	BC	**BCD**		**BCDEF**		
	r^2	r^2	variance ratio (vs. **BC**)	r^2	variance ratio (vs. **BC**)	variance ratio (vs. **BCD**)
1. activity coeff	0.983	0.987	1.31	0.995	2.6**	3.40**
2. log (PC)	0.959	0.992	5.12**	0.996	2.0**	10.25**
3. molar refractivity	0.971	0.971	1.00	0.998	14.5**	14.50**
4. boiling point	0.964	0.982	2.00**	0.999	8.0**	16.0**
5. molar volume	0.894	0.993	15.1**	0.998	3.5**	53.0**
6. $\Delta H_{\text{vaporization}}$	0.972	0.982	1.6*	0.995	3.5**	5.6**
7. magnetic susceptibility	0.925	0.926	1.01	0.927	1.01	1.02
8. critical temp	0.902	0.948	1.88**	0.984	3.2**	6.1**
9. (vdw A)$^{1/2}$	0.976	0.977	1.04	0.983	1.36	1.41
10. vdw B	0.932	0.950	1.36	0.951	1.02	1.39
11. log (dielectric const)	0.709	0.739	1.11	0.780	1.19	1.32
12. sol parameter	0.729	0.817	1.48	0.832	1.09	1.61*
13. critical pressure	0.163	0.523	1.75**	0.591	1.17	2.05**
14. surface tension	0.711	0.818	1.59	0.909	2.00**	3.18**
15. thermal conductivity	0.589	0.613	1.06	0.652	1.11	1.18
16. log (viscosity)	0.717	0.750	1.13	0.849	1.65	1.87*
17. compressibility	0.373	0.864	4.61**	0.871	1.05	4.86**
18. E_T	0.749	0.864	1.85	0.959	3.31**	6.12**
19. dipole moment	0.408	0.412	1.01	0.537	1.27	1.28
20. melting point	0.656	0.708	1.18	0.723	1.05	1.24
21. molecular weight	0.531	0.584	1.13	0.607	1.05	1.19

[a] r^2 is the squared correlation coefficient for regression of the property against the listed eigenvectors, either from Table IV or from the supplementary material; variance ratio is $(1 - r^2$, this equation$)/(1 - r^2$, **BC** or **BCD** equation) as indicated by column heading. A * following a variance ratio implies an inequality of variances significant at the 95% level (F test), i.e., there is considered to be a less than 1 in 20 chance that the improvement in fit produced by the additional regression term(s) is caused by chance. A ** following a variance ratio implies an inequality of variance significant at the 99% level. Degrees of freedom for the F test are $n - 3$ for the **BC** equation, $n - 4$ for the **BCD**, and $n - 6$ for the **BCDEF**, where n is given in Table IV.

first column of Table VI indicates that the one-term equation, in **B** alone, would be a much better correlate of these 16 physical properties than is either molecular weight or molecular volume.

A second type of comparison for the **BC** equations might be with *any* other possible two-parameter equation. One approach to identifying the *best* possible two-parameter property equations would be to generate all the two-parameter equations possibly calculable from the other properties indicated by Table I. While this procedure seems unduly tedious, a reasonable approximation involves application of stepwise regression to Table I, which yields for each property the best of the two-parameter equations having as one of its parameters the property most highly correlating with the property being estimated. This study yielded the final three columns of Table VIII, the two "best" parameters themselves being identified numerically by the middle of the three columns. For example, molar refractivity correlates better with property 4, boiling point, and property 5, molar volume, than with any other pair of properties from Table I that includes molar volume, the property which alone most highly correlates with molar refractivity. The r^2 of 0.907 for this correlation is significantly lower than the r^2 of 0.971 for the molar refractivity **BC** equation.

Overall, the competition between the **BC** equation and these "best possible" regressions against two other properties appears from the variance ratios in the last column of Table VIII to be even. In four instances the **BC** equation is significantly superior and in seven instances the "best stepwise" equation is significantly superior. This result suggests to us that the **BC** parameters cannot be very far from the optimum in all-around data-fitting power. Inspection of the IDs of those property pairs which yield correlations superior to the **BC** correlations does suggest that adding critical pressure to the original factor matrix and repeating the subsequent work might yield an even more powerful set of parameters, although there would probably have to be a third major factor. However, the range of structural variety among compounds whose critical pressure is known is so much narrower than that of Table I that the resulting parameters would have limited applicability.

Discussion

Two variables suffice to explain 95% of the variance in several physical properties of the liquids studied. This previously unsuspected and completely empirical result seems hard to assimilate, perhaps because of a tendency for scientists to visualize only two-dimensional relationships among data. Instead of the typical finding, displayed in an x–y plot, that two variables can be reduced to one, our central finding is that among these data four to ten variables can be reduced to two. It perhaps bears repeating that, from Tables III and II, this reduction of many properties to a plane is much greater, both proportionally and absolutely, than the reduction of any pair of these properties to a line.

A second way of describing these results would be to say that, to the extent that the **B** and **C** parameters are the only significant eigenvalues, a two-term equation has been found for describing the physical properties of liquid molecules that is analogous to the Taft or Swain–Lupton–Unger equations for describing chemical reactivities. Consequently there appear to be no more than two types of intermolecular interactions which produce significant, and truly independent, differences in these observable macromolecular properties. Put differently, any theory of the behavior of liquids which explains these properties in terms of more than two adjustable parameters seems likely to contain redundant information. A widely accepted theory of liquids whose dimensionality conforms to this restriction is the scaled particle theory.[25] It is also interesting that the **BC(DEF)** parameters describe certain mixture equilibria, the interaction of pure substances with water and with lipid, as well as the properties of pure liquids.

The primary value of these findings may be empirical. The

Table VIII. Comparison of Data-Fitting Quality of Two-Factor (**BC**) Equations with Those of Molecular Weight, Molecular Volume, and the "Best Stepwise" Two-Property Equation Found[a]

	BC eq		mol wt eq		molar vol eq		"best stepwise"		
	n	r^2	r^2	variance ratio (vs. **BC**)	r^2	variance ratio (vs. **BC**)	r^2	property IDs	variance ratio (vs. **BC**)
1. AC	114	0.982	0.007	58.4**	0.008	58.3**	0.949	2, 6	3.0**
2. log (PC)	114	0.959	0.379	15.1**	0.534	11.4**	0.882	1, 6	2.88**
3. MR	114	0.971	0.509	16.9**	0.836	5.66**	0.907	4, 5	3.21**
4. bp	114	0.964	0.393	16.9**	0.447	15.4**	0.981	8, 13	0.53**
5. MV	114	0.894	0.380	5.8**			0.941	3, 10	0.56**
6. ΔH_{vap}	114	0.972	0.269	26.1**	0.352	23.1**	0.957	1, 4	1.54*
7. mag susc	89	0.925	0.709	3.9**	0.805	2.6**	0.942	3, 21	0.77
8. CT	83	0.902	0.325	6.9**	0.344	6.7**	0.982	4, 13	0.18**
9. (vdw A)$^{1/2}$	53	0.976	0.749	10.5**	0.828	7.17**	0.995	4, 10	0.21**
10. vdw B	53	0.932	0.742	3.8**	0.911	1.31	0.989	8, 9	0.16**
11. log (e)	66	0.709	0.011	3.4**	0.005	3.42**	0.728	10, 19	0.76
12. sol parameter	54	0.729	0.000	10**	0.155	3.12**	0.820	3, 6	0.66
13. CP	76	0.163	0.082	1.10	0.127	1.04	0.721	9, 10	0.33**
14. surface tension	46	0.711	0.435	1.96*	0.309	3.99**	0.826	4, 13	0.60*
15. thermal conductivity	29	0.589	0.439	1.36	0.027	2.37*	0.778	10, 21	0.54
16. log (visc)	37	0.717	0.089	3.2**	0.135	3.06**	0.797	6, 20	0.72

[a] Variance ratio is defined at the foot of Table VII. A ratio >1 implies inferiority compared with the **BC** equation; a ratio <1 superiority. * The inequality in variance (fit to the data) is significant at the 95% level, i.e., the probability of obtaining this large a difference in fit if the equations are actually of equal quality is less than 1 in 20. ** The inequality in variance is significant at the 99% level.

BC(DEF) parameters have been shown above to be good to excellent correlates for almost every liquid physical property whose value for a variety of substances is known. It has also been found[13] that the **BC(DEF)** parameters for a new substance can often be calculated from structure alone and used to make accurate predictions of physical properties. Therefore, it seems worthwhile to try to derive a **BC(DEF)** equation to "explain" any set of observations which one wishes to explain in terms of previous measurements on the same substances, for example, mixture behaviors, chromatographic data, and biological potencies. However, the **BC(DEF)** parameters do not correlate well with properties which are dependent on specific, non-orientation-averaged, intermolecular interactions, such as melting point.

If the **BC(DEF)** parameters are to be used thus, as the "explanatory" variables for other experimental observations, it seems desirable to have a mechanistic rationalization of why particular molecules have the **BC(DEF)** values that they do. Factor analysts traditionally address this question by rotating the "principal component" axes of their dimensionally simplified data, as mentioned above, in the hopes of aligning them with some set of observations whose relevance is suggested by theory.[26] Although such manipulations might clarify the mechanistic significance of the **BC(DEF)** parameters, it is not easy to imagine a parameter which might be more relevant theoretically to the liquid state than those already involved in the factor analyses,[27] while being well defined for such a large variety of compounds as are found in Table I. Perhaps the **BC(DEF)** vectors are themselves the most appropriate "fundamental" parameters available to describe liquid-state interactions at the molecular level. However, further studies would be desirable.

In any case, a reasonably straightforward mechanistic interpretation of the **B** and **C** parameters already exists. Inspection of the values in Table I shows that the "largest" molecules, naphthalene, *tert*-butylphenol, and octanol, have the highest **B** values, 0.322, 0.426, and 0.364, and the smallest molecules, helium and hydrogen, have the smallest **B** values, −0.505 and −0.511. The **B** values also follow the additive-constitutive relationship that would be expected if **B** is a measure of some aspect of molecular *bulk*.[13]

That the **C** parameter is a measure of molecular *cohesiveness* is suggested by its extrema, the very polar water and acetamide having the highest **C** values, 0.362 and 0.423, and the nonpolar 2,2-dimethylbutane the lowest, −0.257. There is always a tendency for molecules having higher bulk to be more cohesive, and thus a measure of absolute "bulk" would be somewhat correlated with a measure of absolute "cohesiveness". Because the factor analysis instead constrains the **B** and **C** axes to be orthogonal,[28] the **C** axis should more exactly be thought of as "cohesiveness, given that the molecule has this **B** value". Thus, although molecules such as hydrogen and helium of course have the lowest cohesiveness in absolute terms, for a molecule of *given* "bulk" as measured by **B** the cohesiveness of branched hydrocarbons is least.

The normalized equation coefficients in Table VI are consistent with these mechanistic interpretations of **B** and **C**. The largest bB value is for (van der Waals A)$^{1/2}$, a parameter which in the van der Waals equation is thought to express the volume occupied by the molecules themselves under conditions of minimal intermolecular attraction. All of the properties 3-10 are known to be strongly dependent on molecular bulk and do in fact have the highest bB values. The lowest bB values, for critical pressure, dielectric constant, E_T, and dipole moment, are for properties whose relationship to molecular bulk seems least compelling. Similarly for cC, the largest coefficients are associated with electrostatic interactions (the most important non-bulk-related cause of cohesive forces). The relative signs of bB and cC are also interesting. Increased bulk and cohesiveness are seen to increase boiling point, along with the related ΔH_{vap}, CT, and solubility parameter, dielectric constant, surface tension, conductivity, and viscosity, while decreasing compressibility. Bulk and cohesiveness also increase a substance's affinity for water as opposed to the vapor phase (activity coefficient). However, increased bulk but decreased cohesiveness favor a substance in partitioning from water into lipid (partition coefficient), and produce an increase in molar volume, the related refractivity and magnetic susceptibility, and the van der Waals parameters. These trends are all consistent with the postulates that **B** describes the bulk and **C** the bulk-corrected cohesiveness of a molecule.

Mechanistic interpretation of the minor **D, E,** and **F** parameters is less certain. In fact, the lack of collinearity shown in Table V between E_6 and E_{10} and between F_6 and F_{10} suggests that the **E** and **F** vectors are strongly dependent on the particular properties and/or compounds in the factor matrix,

and thus have no consistent mechanistic origin. The **D** parameter has by far its lowest value for iodine, −0.17, and its highest values for ethers, esters, and amines, in the 0.04–0.10 range. The extreme value for iodine suggests that negative **D** values imply large *dispersion* interactions, for a given bulk and cohesive (Coulombic) interaction, because iodine has by far the largest number of electrons per unit volume of any compound in Table I. The low dispersion interactions thereby implied for ethers, amines, and esters might be caused by the small diameter of these functionalities relative to their adjoining methylene groups, leading to large average intergroup separations, lower packing densities, and thus lower average dispersion forces between molecules containing such groups. A slightly different mechanistic interpretation is suggested by the *dD* values of Table VI. The properties most affected are molar volume, critical pressure, surface tension, and compressibility, iodine, for example, contrasting with ethyl propionate in having a much smaller molar volume, less compressibility, and a higher surface tension and critical pressure. One might therefore prefer to identify **D** with *deformability* or compressibility, molecules with higher **D** values having higher deformability.

Deviant property values in the derivation of the property equations of Table IV indicate either experimental error or atypical molecular behavior. Complete lists of residuals, or "calculated-experimental" derivations, can be found in the supplementary material. Here we discuss only the very few experimental observations which were excluded from property-equation derivation, as noted in Table IV. These experimental values differed from their calculated values by more than five times the average error, so that their inclusion seemed likely to degrade seriously the ability of the resulting regression equation to predict the property values for the remaining compounds. For water, the experimental solubility parameter, critical pressure, surface tension, and thermal conductivity all are far away from the values calculated by the **BCDEF** equations, and thus water does seem to have a "unique" aspect. On the other hand, with regard to the majority of its properties, notably those involving mixtures, water seems to be a fairly typical molecule. The other "outlying" observations are the critical temperature of 1,1-difluoroethane, having a tabulated value of 387 °C but a calculated value of 124 °C, and the thermal conductivity of chloromethane, tabulated value of 4.6×10^4 units, calculated value of 3.05×10^4. No explanation of these deviations is apparent.

Considerable publicity has been given recently to numerous high correlations, mainly within homologous and isomeric analogue series, between various of these physical properties and "molecular connectivity", an index whose rationale depends upon an analogy between the properties of molecules and the properties of graphs.[29] In view of the above factorization results, it seems reasonable to suppose that molecular connectivity correlations are artifacts, perhaps representing alternative axes for compound subsets within "**BC(DEF)** space".

Finally, many research results, particularly involving physical aspects of biological systems, have been interpreted in terms of a "hydrophobic interaction", which as the name suggests is meant to imply that the driving force in partitioning substances from water into lipid is a bulk-related entropically disfavorable reorganization of water molecules about a nonpolar substance.[30] Although such a reorganization probably occurs to some extent, the notion that this phenomenon plays any important role in partitioning energetics has recently come under repeated criticism.[31] In this connection, it should be noted that partitioning data have been shown above to be governed by exactly the same types of interactions which describe the behavior of pure substances, overall and aspecific bulk and cohesiveness. Furthermore, the positive coefficient of the *bB* term for the aqueous activity coefficient in Table VI suggests that, in opposition to the prediction of the "hydrophobic interaction", increases in molecular bulk *favor* the interaction of a substance with water. Finally, the molar volume/partition coefficient correlation which is used to justify the "hydrophobic interaction" in its most popular presentations is seen here to be a relatively minor manifestation of an underlying, much more general, interrelationship among liquid-state properties.

Acknowledgments. I thank S. Wold, A. J. Hopfinger, S. Unger, and T. Spurling for their encouragements and helpful criticism.

Supplementary Material Available: Values of all experimental data used, as indicated in Table I; the eigenvectors for the four-property, 138-compound and the ten-property, 44-compound factorizations; the equations analogous to those in Table IV but based on **BC** or **BCD** only; residuals for these equations as well as those in Table IV (31 pages). Ordering information is given on any current masthead page.

References and Notes

(1) A. Leo, C. Hansch, and D. Elkins, *Chem. Rev.*, **71**, 525 (1971); G. G. Nys and R. F. Rekker, *Chim. Ther.*, **5**, 521 (1973); A. Leo, P. Y. C. Jow, and C. Hansch, *J. Med. Chem.*, **18**, 865 (1975).
(2) C. R. Kinney, *J. Am. Chem. Soc.*, **60**, 3032 (1938). See "Lange's Handbook of Chemistry and Physics", 11th ed., McGraw-Hill, New York, 1973, Table 10–13.
(3) O. Exner, *Collect. Czech. Chem. Commun.*, **32**, 1 (1967). Kopp (1885) first observed additivity for this property (S. Glasstone, "Textbook of Physical Chemistry", 2nd ed., Van Nostrand-Reinhold, Princeton, N.J., 1946, Chapter 8.
(4) References cited to Pascal rules by C. A. Hutchinson in "Determination of Organic Structures by Physical Methods", Vol. 1, E. A. Braude and F. C. Nachod, Eds., Academic Press, New York, 1955, Chapter 7, p 259.
(5) R. D. Cramer, III, *Annu. Rep. Med. Chem.*, **11**, 301 (1976); G. Redl, R. D. Cramer, III, and C. E. Berkoff, *Chem. Soc. Rev.*, **3**, 273 (1974); Y. C. Martin, "Quantitative Drug Design", Academic Press, New York, 1978.
(6) (a) B. R. Kowalski, Ed., "Chemometrics: Theory and Applications", American Chemical Society Symposium Series, No. 52, Washington, D.C., 1977; (b) P. C. Jurs and T. L. Isenhour, "Chemical Applications of Pattern Recognition", Wiley, New York, 1975; (c) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **96**, 918 (1974).
(7) (a) H. H. Harman, "Modern Factor Analysis", University of Chicago Press, Chicago, (b) R. J. Rummel, "Applied Factor Analysis", Northwestern University Press, Evanston, Ill., 1970. (c) A recent and important contribution describes hazards with this technique and lists more applications: C. G. Swain, H. E. Bryndza, and M. S. Swain, *J. Chem. Inf. Comput. Sci.*, **19**, 19 (1979).
(8) S. H. Unger, Ph.D. Thesis, Massachusetts Institute of Technology, 1969; S. Wold and M. Sjostrom, *Chem. Scr.*, **2**, 49 (1972); **6**, 114 (1974); **9**, 200 (1976).
(9) P. H. Weiner, E. R. Malinowski, and A. Levinstone, *J. Phys. Chem.*, **74**, 437 (1970).
(10) J. R. McGill and B. R. Kowalski, *Anal. Chem.*, **49**, 596 (1977); S. S. Schiffman, *Science*, **185**, 112 (1974).
(11) M. W. Weiner and P. H. Weiner, *J. Med. Chem.*, **16**, 665 (1973); G. K. Menon and A. Cammarata, *J. Pharm. Sci.*, **66**, 304 (1977); W. J. Dunn, S. Wold, and Y. C. Martin, *J. Med. Chem.*, **21**, 922 (1978).
(12) (a) P. H. Weiner and J. F. Parcher, *Anal. Chem.*, **45**, 302 (1973); (b) R. Franke in "Biological Activity and Chemical Structure", Vol. 2, J. A. K. Buisman, Ed., Elsevier, Amsterdam, 1977, p 251; G. L. Dunn and S. Wold, *Acta Chem. Scand., Ser. B*, **32**, 536 (1978).
(13) R. D. Cramer, III, *J. Am. Chem. Soc.*, following paper in this issue.
(14) J. Hine and P. K. Mookerjee, *J. Org. Chem.*, **40**, 292 (1975); R. Wolfenden, *Biochemistry*, **17**, 201 (1978). For example, the activity coefficient of water itself is calculated:

$$c_{H_2O}/c_{vapor} \text{ at equilibrium and } 298\text{ K} = \log \frac{(1000\text{ g L}^{-1}/18\text{ g mol}^{-1})}{(23.7/760\text{ atm})/(0.082\text{ L atm K}^{-1}\text{ mol}^{-1} \times 298\text{ K})} = \log (55.66/1.28 \times 10^{-3}) = 4.64$$

(15) A. J. Leo, Pomona College Medicinal Chemistry Project, Claremont, Calif., semiannually distributed update. See also ref 1a.
(16) "CRC Handbook of Chemistry and Physics", 58th ed., CRC Press, Cleveland, Ohio, 1977–1978. Properties cited are listed as follows: no. 4, 5, 20, and 21; C81-C548; no. 6, C733-C745; no. 7, E127-E135; no. 8 and 13, F89-F90; no. 9 and 10, D178; no. 11, E56-E58; 12, C726-C732; no. 14, F46-F48; no. 15, E4; no. 16, F52-F57; no. 17, F16-F20; no. 19, E63-E65. Where different property values were reported for different experimental conditions, the liquid-state value was preferred. Otherwise the value for room temperature and atmospheric pressure was used.
(17) Computed using the Lorentz expression:

$$\frac{(N_D^2 - 1)}{(N_D^2 + 2)} \frac{MW}{D}$$

where N_D is the refractive index, MW is molecular weight, and D is density.
(18) C. Reichardt, *Angew. Chem., Int. Ed. Engl.*, **4**, 29 (1965).

(19) C. Hansch, J. E. Quinlan, and G. L. Lawrence, *J. Org. Chem.*, **33**, 347 (1968); A. Leo, C. Hansch, and C. Church, *J. Med. Chem.*, **12**, 766 (1969); A. Cammarata, S. J. Yan, and K. S. Rogers, *ibid.*, **14**, 1211 (1971).
(20) A. Leo, C. Hansch, and P. Y. C. Yow, *J. Med. Chem.*, **19**, 611 (1976). See also ref 28.
(21) E. R. Malinowski in ref 6a, Chapter 3, p 53.
(22) S. Wold, *Technometrics*, in press; S. Wold and M. Sjostrom in ref 6a, Chapter 12, p 243.
(23) P. H. Weiner, *Chemtech*, **7**, 321 (1977). Studies were carried out at the University of Pennsylvania Medical School Computer Facility.
(24) Inspection of Figure 1 will show that, unless data are standardized, the use of widely different measurement scales for different properties will tend to introduce spurious structure. For example, if boiling points were recorded in hundreds of degrees rather than degrees, variance along the boiling-point axis would virtually disappear. See discussion of "autoscaling" in ref 6c.
(25) R. A. Pierotti, *Chem. Rev.*, **76**, 717 (1976).
(26) D. G. Howery in ref 6a, Chapter 4, p 73. See also ref 6–12. By the criteria in ref 7c, our procedure of factor analysis would be characterized by the following quote: "the best way to obtain correct parameters is to . . . narrow the scope of the study to a full subset having no missing data and then use principal components analysis followed by a valid transformation." According to these authors' findings, popular methods of transformation can give physically absurd results. This seems to justify our decision to perform no transformation whatsoever, beyond the principal components analysis described.
(27) F. M. Richards, *Annu. Rev. Biophys. Bioeng.*, **6**, 151 (1977), gives an instructive discussion of the difficulties in defining "molecular volume", one plausible major component of **BCDEF** space.
(28) An example of this phenomenon is cited by Harman (ref 7a). Given a set of data on the falling times of various balls through various media, the factor analyst presumably would discover that two variables correlate the observations. These two variables would not be identical with weight and volume, however, because the weights and volumes of balls are partially correlated. Instead one variable would probably be weight, but the second would be "volume corrected for weight".
(29) The earliest reference is to M. Randic, *J. Am. Chem. Soc.*, **97**, 6609 (1975), and a recent one to T. DiPaolo, L. B. Kier, and L. H. Hall, *J. Pharm. Sci.*, **68**, 39 (1979). A review is L. B. Kier and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research", Academic Press, New York, 1976.
(30) C. Tanford, "The Hydrophobic Effect", Wiley, New York, 1973, and references cited therein.
(31) P. Mukerjee, *Adv. Colloid Interface Sci.*, **1**, 241 (1967); O. W. Howarth, *J. Chem. Soc., Faraday Trans. 1*, **71**, 2303 (1975); R. A. Wolfenden and C. A. Lewis, *J. Theor. Biol.*, **59**, 231 (1976); R. D. Cramer, III, *J. Am. Chem. Soc.*, **99**, 5408 (1977); K. Shinoda, *J. Phys. Chem.*, **81**, 1300 (1977); J. H. Hildebrand, *Proc. Natl. Acad. Sci. U.S.A.*, **76**, 194 (1979). M. H. Abraham, *J. Am. Chem. Soc.*, **101**, 5477 (1979), has very recently responded to the Cramer and Wolfenden criticisms. In brief reply, the central issue should perhaps be "Are there any experimental data which *require* a 'hydrophobic effect'?" instead of "Can the experimental data be manipulated so as to *allow* postulation of a 'hydrophobic effect'?" More specifically, Abraham asserts "hydrophobicity" to be an attribute of hydrocarbon but *not* of the completely apolar rare gases. Of what value can such a construct be?

BC(DEF) Parameters. 2. An Empirical Structure-Based Scheme for the Prediction of Some Physical Properties†

Richard D. Cramer, III

Contribution from the Department of Chemistry, Research and Development, Smith Kline and French Laboratories, Philadelphia, Pennsylvania 19101.
Received June 11, 1979

Abstract: Based on either a hierarchically organized additive-constitutive model or a subset of four physical properties, for calculation of intermediate **BC(DEF)** values where **BCDEF** are the principal components of a matrix of six physical properties of 114 compounds, all experimental values of 18 common physical properties for 139 additional compounds of diverse structure have been "predicted". The rms difference between the 1142 predicted and experimental values is 22% of the variance in the experimental values, corresponding to a "correlation coefficient" or "r" of 0.88. For the 118 compounds and 10 properties to which application of the **BC(DEF)** model is clearly warranted, the rms difference between the 749 predicted and actual values is 6% of the overall variance; that is, the "r" is 0.97. Predictions using the **BC(DEF)** model are at least as accurate as those of existing additive-constitutive models for individual properties. There is no significant difference in predictive accuracy between **BCDEF** values derived from the additive-constitutive model and **BCDEF** values derived from the property subset. The five-factor **BCDEF** model is more accurate than the two-factor **BC** model for compounds having reasonable structural similarity to any of the 114 used to derive the **BCDEF** scale, but the two-factor model is the less likely to give completely misleading results for very different structures.

In the preceding paper,[1] analysis of a collection of physical-property data for a variety of pure liquid compounds showed that more than 95% of the variance in most of the properties can be explained in terms of a two-, three-, or five-component "**BC(DEF)**" model, where the components are derived by factorization of a matrix constructed from the values of activity coefficient, partition coefficient, boiling point, molar volume, refractivity, and heat of vaporization for 114 compounds. In this paper, the generality and utility of this model will be investigated by "predicting" the experimentally known properties of 139 compounds not among the 114 used for derivation of the model.

Prediction of a property using the **BC(DEF)** scheme has two steps: (1) calculation of the **BC(DEF)** values for the compound, either from previously known properties or from its structure alone; (2) calculation of the property, from the **BC(DEF)** values and the appropriate previously derived "property equation" (Table IV[1]).

Although structurally based schemes have been proposed for calculating some of the physical properties encompassed by the **BC(DEF)** models,[2] little attention has been given to scope and limitations. One notable exception is Exner's discussions of the significance of the long-known additive-constitutive behaviors of molar volume and parachor.[3] Types of information which add to the utility of any predictive scheme include answers to the following questions: (1) What kinds of molecules (and properties) can the scheme confidently be applied to? (2) What must be known about a molecule in order to calculate an unknown property? (3) How accurate are the results? These questions provide an outline for the following description of our data and methods.

Scope of the BC(DEF) Model. In choosing the 139 compounds whose properties were to be predicted, the major objectives were a large number of examples of values for the rarer properties and a structurally diverse data set. The completed

† Presented in part at the 177th National Meeting of the American Chemical Society, Honolulu, Hawaii, 1979.

0002-7863/80/1502-1849$01.00/0