A General QSPR Treatment for Dielectric Constants of Organic Compounds

Sulev Sild and Mati Karelson*

Institute of Chemical Physics, University of Tartu, Jakobi Street 2, Tartu, 51014, Estonia

Received August 28, 2001

Multilinear regression and neural network methods have been used to develop QSPR models for the prediction of the dielectric constant (ϵ) and Kirkwood function $(\epsilon - 1)/(2\epsilon + 1)$ of organic liquids. Both methods can provide acceptable models for the prediction of these properties. The QSPR models developed from the training set of 155 diverse compounds use theoretical molecular descriptors encoding electronic properties of the molecule and the intermolecular interaction between molecules. The QSPR models for the Kirkwood function appear to be more reliable than the models for the dielectric constant. The average prediction error of the best model for the dielectric constant is 27.0%. The average prediction error of the best model for the Kirkwood function is 4.1%.

INTRODUCTION

The static dielectric constant, also called the relative permittivity (ϵ), is a well defined molecular bulk property. This property is relatively easy to measure, and experimental data for many organic and inorganic compounds are readily available from the literature. The dielectric constant is a macroscopic property that is measured as a ratio of the capacitance of a condenser with the material as dielectric to its capacitance with vacuum as dielectric. Thus, the dielectric constant represents the ability of a substance to separate charge and/or to orient its molecular dipoles in external electric field. Solvent effects described through their dielectric properties may play an important role in the chemical reactivity and equilibrium. For instance, in the case of the Menschutkin reaction, the reaction rate is accelerated with increasing solvent polarity, and reaction rates are well correlated with the Kirkwood function $(\epsilon - 1)/(2\epsilon + 1)$ of the solvent.[1] However, this correlation is restricted only to solvents with nonspecific solvent effects dominating. In general, the dielectric constant can be very useful parameter to consider when reaction environment needs to be optimized to achieve better control over the chemical reaction. The dielectric constant is therefore often applied in the theoretical studies of solvent effects.[1,2] For example, the nonspecific solvent effects can be accounted for within the quantum chemical models by using the self-consistent reaction field (SCRF) method.[3,4] This method is based on the Kirkwood-Onsager model, where the function of the dielectric constant and molecular cavity parameters (e.g. cavity radius) is used in the Hamiltonian to account for the electric field response of the solvent.

The dielectric constant is frequently used as a practical parameter to characterize the polarity of organic solvents. Nonpolar solvents have generally lower dielectric constant values than polar solvents. However, in the presence of specific intermolecular interactions, the dielectric constant alone is insufficient as a universal measure for the solvent polarity. Therefore, a large number of alternative solvent polarity scales have been proposed.[5,6] Various functions of the dielectric constant are correlated with many of those empirical solvent scales. For example, Katritzky et al. proposed multiparameter equation for $E_T(30)$ scale that combined Kirkwood function $(\epsilon - 1)/(2\epsilon + 1)$ and the index of refraction.[7] Koppel and Palm proposed an alternative multiparameter approach for the solvent polarity.[8] In the last model, the parameters, accounting for nonspecific interactions, were based on the function of dielectric constant.

The ability to predict dielectric constants theoretically is valuable in the molecular design of new materials. The ability to make fast and reliable predictions over a wide range of diverse chemical structures will substantially increase the productivity and speed of the research. However, the theoretical calculation of the dielectric constant of a fluid from the first principles is a complex problem. The value of dielectric constant is strongly related both to the chemical structure of a molecule and to intermolecular interactions. In addition, external conditions (temperature, pressure, etc.) need to be accounted for to describe this macroscopic property. Currently, several models and methods have been used, however leading to diverse results. A recent summary[9] by Tomasi et al. describes several of those calculation methods, including the calculation of the square of the total dipole moment of system (fluctuation method), the polarization response method, integral equations, and molecular Ornstein−Zernike theory. An alternative and mathematically simpler option for the prediction of dielectric constant would be the respective quantitative structure−property relationship (QSPR). The QSPR method relies on various statistical techniques to find correlation between investigated property and predefined set of theoretical molecular descriptors. The QSPR method has been successfully applied to many different technologically relevant physical properties of chemical compounds, as summarized in a recent review.[10]

Two QSPR models for the dielectric constant have been reported recently in the literature. Cocchi et al. used multiple linear regression analysis and multivariate partial least-squares method to develop QSPR models for several phys-

* Corresponding author phone: (+372-7) 375-254; fax: (+372-7) 375-264; e-mail: mati@theor.chem.ut.ee.

10.1021/ci010335f CCC: $22.00
Published on Web 02/01/2002

icochemical properties of organic solvents.[11] Based on a training set of 23 and a testing set of 20 compounds, they proposed three-parameter multiple linear regression model with squared correlation coefficient (R^2) of 0.9564. The dielectric constant values span the range from 2 to 41. The root-mean-squared (RMS) errors for the training and testing sets were 2.262 and 4.650, respectively. They also proposed an alternative 15 variable partial least squares (PLS) model with R^2 of 0.9744. The RMS error for the training and testing sets were 1.576 and 3.213, respectively. Another QSPR model has been developed by Schweitzer et al.[12] In this case, the neural network technique was used to develop a 10 parameter nonlinear QSPR model from a data set of 497 organic compounds. The dielectric constant values span the range from 1 to 40. The best model had RMS errors for the testing set of 97 compounds, and the training set of 350 of compounds were 2.33 and 3.77, respectively.

In many cases, the use of the function of dielectric constant is much more practical than the dielectric constant itself. For a function of dielectric constant to be theoretically predicted, the development a specific QSPR model for this particular function would be more feasible. Accordingly, in this work we have developed several QSPR models for dielectric constant (ϵ) and to Kirkwood function $(\epsilon - 1)/(2\epsilon + 1)$. We also describe techniques that were used for QSPR treatment, perform detailed analysis of developed QSPR models, and compare our results with other existing QSPR models.

DATA SET

The QSPR treatment started with the assembly of the data set. The experimental dielectric constant data were compiled from the CRC Handbook.[13] The Kirkwood function $(\epsilon - 1)/(2\epsilon + 1)$ values were calculated from the respective dielectric constant values. A total of 155 compounds with extensive structural diversity were selected as the data set (see Table 1). The quality and robustness of the predictive power of a QSPR model depends heavily on the diversity of data set. To select significant descriptors for the QSPR model that captures all the underlying interaction mechanisms, it is advisable to have as many as possible structural features represented in the data set. The working data set included saturated and nonsaturated hydrocarbons, aromatic and nonaromatic cyclic compounds. Also, the compounds chosen contained aldehyde, amino, amide, ether, ester, hydroxyl, nitrate, nitrile, nitro, phosphate, and thiol functionalities. The structures include C, H, N, O, S, P, and halogen atoms. Since the dielectric constant depends significantly on the temperature, all property values collected were measured at 293 K. The dielectric constant values for organic liquids can range from about 1.8 to 180. However, for the majority of organic liquids the property values are less than 50. Considering this, we limited the dielectric constant values in our data set to a range between 1.87 and 46.5. An additional set of 46 compounds was also selected from the CRC Handbook to form the external prediction set (see Table 2). This prediction set was used to evaluate the predictive quality of the QSPR models obtained.

Molecular descriptors were calculated for each compound in the training and prediction sets using the CODESSA program.[14] Overall, more than 600 molecular descriptors were calculated. Traditionally, the theoretical descriptors were divided into the following categories: constitutional, topological, geometric, electrostatic, and quantum chemical.[15] The constitutional descriptors (such as molecular weight, number of oxygen atoms, etc.) can be derived only on the basis of molecular formula. Topological descriptors are constructed from the two-dimensional connectivity matrix of the molecule by using algorithms from graph theory. Geometrical descriptors require three-dimensional coordinates of atoms, which are usually extracted from the output of quantum-chemical or molecular mechanics calculations. Quantum chemical descriptors are directly extracted or calculated from the output of semiempirical or ab initio quantum-chemical calculations. Electrostatic descriptors are usually calculated from partial atomic charges and three-dimensional coordinates of atoms. Partial atomic charges (PC) can be obtained using either quantum-chemical or empirical methods. In the present study, semiempirical AM1 method was used to calculate quantum-chemical descriptors and three-dimensional coordinates, for partial atomic charges semiempirical AM1[16] and the empirical Zefirov method[17] were used. All AM1 calculations were performed with the MOPAC 6.0 program.[18] The initial geometries were generated using standard bond lengths and angles.[13]

MULTILINEAR REGRESSION MODELS

Linear QSPR models can be developed with several statistical techniques, such as multivariate linear regression, partial least-squares regression, and principal components regression.[14] In this study, we applied stepwise multivariate regression analysis to select significant descriptors for linear QSPR models.[14,15] The QSPR treatment started with the reduction of the number of molecular descriptors. If two descriptors were highly intercorrelated, then only one descriptor was selected, and also descriptors with insignificant variance were rejected. This procedure helps to speed up the descriptor selection and reduces the risk of including unrelated descriptors by chance. The size of the descriptor pool was reduced to 112 descriptors. The QSPR models were developed with CODESSA program by using the best multilinear regression analysis algorithm.

The QSPR analysis for dielectric constant resulted in a six-parameter model with the correlation coefficient R^2 of 0.9447, as summarized in Table 3 and Figure 1. The cross-validated R^2_{cv} (calculated with leave one out procedure) for this model is 0.936. The RMS errors for training and prediction sets are 2.368 and 3.743, respectively. This multilinear regression (MLR) model includes four descriptors related to the electronic properties of the molecule. The *image of Onsager-Kirkwood solvation energy*, the *total hybridization component of the molecular dipole*, and the *maximal atomic orbital electronic population* describe electronic structure of the molecule in semiempirical AM1 level. Electrostatic interactions are described by the *topological electronic index (all bonds) [Zefirov's PC]*. The *maximum atomic orbital electronic population* is a quantum-chemical descriptor that is defined by the atomic orbital from the valence shell with the highest electron density value, calculated from MO coefficients. This descriptor describes electron density distribution across the valence shell, and its value is generally determined by electron density of the *s* orbital. Essentially, this descriptor depends on the number

Table 1. Experimental and Calculated Property Values for Dielectric Constant and Kirkwood Function $(\epsilon - 1)/(2\epsilon + 1)$ for the Training Set

	dielectric constant			Kirkwood function		
structure	exp.	MLR	NN	exp.	MLR	NN
acetonitrile	36.640	36.411	38.468	0.480	0.478	0.480
allyl alcohol	20.700	22.746	23.020	0.465	0.470	0.465
benzaldehyde	17.850	21.518	21.615	0.459	0.478	0.464
benzene	2.283	4.250	4.741	0.230	0.244	0.249
benzonitrile	25.900	24.867	26.465	0.472	0.480	0.466
benzoyl chloride	23.000	19.245	18.023	0.468	0.487	0.452
benzoyl fluoride	22.700	21.162	20.715	0.468	0.492	0.459
benzylamine	5.180	5.623	6.175	0.368	0.355	0.380
2-bromo-2-methylpropane	10.980	7.108	5.624	0.435	0.396	0.402
1-bromooctane	5.096	3.784	3.557	0.366	0.367	0.375
1-bromopropane	8.090	7.294	5.806	0.413	0.407	0.403
2-bromopropane	9.460	7.912	6.218	0.425	0.413	0.407
butanenitrile	24.830	21.518	21.942	0.470	0.449	0.468
2-butanol	17.260	17.510	16.298	0.458	0.444	0.457
butyl acetate	5.070	4.396	3.682	0.365	0.363	0.367
butylamine	4.710	7.276	7.513	0.356	0.368	0.391
butyl nitrate	13.100	15.414	13.661	0.445	0.426	0.450
1-chlorobutane	7.276	9.039	6.530	0.404	0.408	0.409
2-chlorobutane	8.564	10.000	7.374	0.417	0.417	0.418
chloroethane	9.450	12.301	9.559	0.425	0.429	0.430
1-chloroheptane	5.521	6.419	4.628	0.375	0.383	0.389
1-chlorohexane	6.104	7.584	5.413	0.386	0.395	0.399
2-chloro-2-methylpropane	9.663	9.978	7.330	0.426	0.409	0.417
1-chloropentane	6.654	8.651	6.226	0.395	0.405	0.408
1-chloropropane	8.588	10.322	7.663	0.417	0.417	0.418
cyclohexane	2.024	2.376	3.579	0.203	0.206	0.194
cyclohexanone	16.100	17.877	16.566	0.455	0.444	0.466
cyclohexene	2.218	2.784	3.889	0.224	0.235	0.228
cyclohexylamine	4.547	5.167	6.017	0.351	0.351	0.348
cyclopentane	1.969	2.955	3.908	0.196	0.216	0.198
dibenzylamine	3.446	4.124	5.028	0.310	0.314	0.310
1,2-dibromoethane	4.961	7.613	5.134	0.363	0.367	0.377
dibutylamine	2.765	3.056	4.474	0.270	0.324	0.273
N,*N*-dibutylacetamide	19.100	17.166	15.296	0.462	0.429	0.444
N,*N*-dibutylformamide	18.400	19.046	17.771	0.460	0.443	0.454
1,2-dichloroethane	10.420	7.478	5.048	0.431	0.386	0.382
1,2-dichloropropane	8.370	6.696	4.612	0.415	0.386	0.379
diethanolamine	25.750	32.015	32.667	0.471	0.502	0.476
diethylamine	3.680	6.293	6.480	0.321	0.362	0.333
N,*N*-diethylacetamide	32.100	22.988	23.234	0.477	0.470	0.466
diethyl ether	4.267	5.759	4.844	0.343	0.374	0.367
N,*N*-diethylformamide	29.600	25.699	26.978	0.475	0.483	0.472
2,3-dimethyl-1,3-butadiene	2.102	2.867	3.802	0.212	0.234	0.240
2,2-dimethylbutane	1.869	1.515	3.111	0.183	0.203	0.190
2,3-dimethylbutane	1.889	1.586	3.137	0.186	0.205	0.191
N,*N*-dimethylformamide	38.250	37.654	39.575	0.481	0.516	0.480
2,2-dimethylpentane	1.915	0.892	2.813	0.189	0.195	0.188
2,3-dimethylpentane	1.929	0.889	2.824	0.191	0.196	0.188
2,4-dimethylpentane	1.902	0.951	2.836	0.188	0.198	0.188
2,2-dimethylpropanal	9.051	16.572	14.900	0.421	0.429	0.463
2,4-dimethylpyridine	9.600	11.414	10.466	0.426	0.414	0.429
2,6-dimethylpyridine	7.330	8.395	7.849	0.404	0.393	0.405
1,2-ethanediamine	13.820	13.063	12.544	0.448	0.416	0.465
ethanol	25.300	25.414	26.882	0.471	0.464	0.472
ethanolamine	31.940	32.499	32.242	0.477	0.520	0.481
ethyl acetate	6.081	6.694	5.105	0.386	0.386	0.391
ethylbenzene	2.446	2.898	3.954	0.245	0.237	0.239
ethyl benzoate	6.200	5.322	4.092	0.388	0.382	0.368
ethylene glycol	41.400	36.132	38.816	0.482	0.464	0.478
ethyl hexanoate	4.450	3.163	3.084	0.348	0.347	0.356
ethyl isopentyl ether	3.955	2.597	3.023	0.332	0.338	0.323
ethyl isothiocyanate	19.600	15.292	13.995	0.463	0.449	0.452
ethyl nitrate	19.700	19.381	18.658	0.463	0.442	0.459
ethyl propanoate	5.760	6.576	5.163	0.380	0.382	0.394
2-ethylpyridine	8.330	9.241	8.589	0.415	0.403	0.406
4-ethylpyridine	10.980	13.563	12.619	0.435	0.422	0.442
ethyl 4-pyridinecarboxylate	8.950	6.440	5.922	0.421	0.409	0.395
fluorobenzene	5.465	8.249	6.235	0.374	0.416	0.399
furfural	42.100	42.307	42.806	0.482	0.516	0.474
glycerol	46.530	38.911	40.059	0.484	0.488	0.480
1-heptanol	11.750	14.173	12.296	0.439	0.418	0.436
2-heptanone	11.950	15.482	13.525	0.440	0.430	0.458
4-heptanone	12.600	15.024	12.983	0.443	0.428	0.455
1-heptene	2.092	1.719	3.225	0.211	0.230	0.219
1-hexanol	13.030	15.774	13.695	0.445	0.447	0.447
2-hexanone	14.560	18.610	17.498	0.450	0.445	0.465
hexyl acetate	4.420	3.245	2.979	0.348	0.348	0.352
hexylamine	4.080	4.777	5.630	0.336	0.352	0.338

Table 1 (Continued)

	dielectric constant			Kirkwood function		
structure	exp.	MLR	NN	exp.	MLR	NN
iodobenzene	4.590	3.651	4.610	0.353	0.401	0.389
1-iodobutane	6.270	2.850	4.129	0.389	0.384	0.384
iodoethane	7.820	4.413	4.880	0.410	0.405	0.397
1-iodohexane	5.350	1.916	3.637	0.372	0.374	0.378
iodomethane	6.970	4.610	5.034	0.400	0.409	0.399
1-iodopentane	5.780	2.626	3.951	0.381	0.384	0.384
1-iodopropane	7.070	3.585	4.505	0.401	0.392	0.390
isobutyl acetate	5.068	5.332	4.303	0.365	0.370	0.381
isobutylbenzene	2.318	1.688	3.259	0.234	0.219	0.227
isopropylamine	5.627	7.282	7.521	0.378	0.375	0.392
isopropylbenzene	2.381	2.222	3.551	0.240	0.227	0.232
m-dichlorobenzene	5.020	7.280	5.044	0.364	0.409	0.378
methanol	33.000	34.761	37.184	0.478	0.496	0.480
N-methylaniline	5.960	7.682	7.377	0.384	0.366	0.377
3-methyl-1-butanol	15.630	15.568	13.794	0.454	0.436	0.437
methylcyclohexane	2.024	1.751	3.229	0.203	0.199	0.192
methyl cyclohexanecarboxylate	4.870	3.358	3.214	0.360	0.348	0.358
methylcyclopentane	1.985	2.312	3.536	0.198	0.211	0.195
2-methylhexane	1.922	1.015	2.858	0.190	0.198	0.189
3-methylhexane	1.920	0.916	2.841	0.190	0.197	0.188
methyl hexanoate	4.615	3.687	3.356	0.353	0.355	0.361
methyl nitrate	23.900	19.822	19.275	0.469	0.454	0.459
2-methylpentane	1.886	1.603	3.153	0.186	0.206	0.191
3-methylpentane	1.886	1.539	3.141	0.186	0.204	0.190
2-methylpropanenitrile	24.420	21.277	21.622	0.470	0.450	0.468
2-methyl-2-propanethiol	5.475	9.583	8.727	0.374	0.416	0.436
methyl propanoate	6.200	6.321	4.948	0.388	0.385	0.389
2-methyl-1-propanol	17.930	17.928	16.815	0.459	0.443	0.453
2-methylpyridine	10.180	10.829	9.936	0.430	0.414	0.430
4-methylpyridine	12.200	14.616	13.800	0.441	0.429	0.448
N-methyl-2-pyrrolidone	32.550	26.510	28.158	0.477	0.488	0.473
m-xylene	2.359	2.828	3.896	0.238	0.234	0.237
nitrobenzene	35.600	34.599	36.913	0.479	0.514	0.475
nitromethane	37.270	39.587	41.015	0.480	0.496	0.478
1-nonanol	8.830	11.443	9.184	0.420	0.417	0.418
5-nonanone	10.600	12.203	9.868	0.432	0.409	0.440
1-nonene	2.180	0.495	2.631	0.220	0.208	0.209
o-chloroaniline	13.400	18.025	16.338	0.446	0.475	0.421
octane	1.948	0.506	2.607	0.194	0.187	0.186
octanenitrile	13.900	14.038	12.459	0.448	0.416	0.443
1-octanol	10.300	12.405	10.448	0.431	0.405	0.416
2-octanone	9.510	13.690	11.453	0.425	0.417	0.451
1-octene	2.113	1.115	2.926	0.213	0.219	0.215
octylamine	3.580	3.172	4.791	0.316	0.330	0.308
o-dichlorobenzene	10.120	8.758	6.252	0.429	0.421	0.395
o-nitrotoluene	26.260	22.578	23.005	0.472	0.483	0.460
o-xylene	2.562	2.857	3.975	0.255	0.238	0.239
pentanenitrile	20.040	19.039	18.637	0.463	0.441	0.463
2-pentanone	15.450	19.734	19.073	0.453	0.450	0.468
3-pentanone	17.000	19.495	18.751	0.457	0.451	0.467
phenyl pentanoate	4.300	2.193	2.493	0.344	0.351	0.338
propanenitrile	29.700	24.571	25.986	0.475	0.456	0.473
1-propanol	20.800	20.936	20.622	0.465	0.460	0.464
2-propanol	20.180	20.769	20.559	0.464	0.456	0.465
propyl acetate	5.620	5.873	4.482	0.377	0.376	0.381
propylbenzene	2.370	2.276	3.575	0.239	0.225	0.232
propyl butanoate	4.300	3.322	3.080	0.344	0.353	0.354
p-xylene	2.274	2.867	3.879	0.230	0.228	0.236
pyridine	13.260	13.907	13.043	0.445	0.428	0.447
pyrrolidine	8.300	8.336	8.088	0.415	0.377	0.392
quinoline	9.160	10.339	9.448	0.422	0.406	0.418
styrene	2.474	3.662	4.300	0.248	0.240	0.253
tert-butylbenzene	2.359	1.574	3.208	0.238	0.220	0.226
1,1,2,2-tetrachloroethane	8.500	6.716	4.538	0.417	0.378	0.367
tetrahydropyran	5.660	6.772	5.585	0.378	0.371	0.377
tetranitromethane	2.317	2.565	2.375	0.234	0.269	0.240
tribromofluoromethane	3.000	4.395	3.976	0.286	0.298	0.332
tributylamine	2.340	1.159	3.411	0.236	0.256	0.233
tributyl phosphate	8.340	6.993	7.390	0.415	0.399	0.438
1,1,1-trichloroethane	7.243	8.317	5.964	0.403	0.408	0.391
2,2,2-trifluoroethanol	27.680	29.977	31.925	0.473	0.509	0.475
1,2,3-trimethylbenzene	2.656	2.119	3.583	0.262	0.230	0.232
1,2,4-trimethylbenzene	2.377	2.127	3.507	0.239	0.225	0.230
1,3,5-trimethylbenzene	2.279	2.197	3.496	0.230	0.221	0.229
2,2,4-trimethylpentane	1.943	0.233	2.520	0.193	0.185	0.186
trimethyl phosphate	20.600	23.612	21.489	0.464	0.477	0.478
tripropyl phosphate	10.930	10.994	10.279	0.434	0.430	0.455

Table 2. Experimental and Calculated Property Values for Dielectric Constant and Kirkwood Function $(\epsilon - 1)/(2\epsilon + 1)$ for the Prediction Set

	dielectric constant			Kirkwood function		
structure	exp.	MLR	NN	exp.	MLR	NN
acetone	21.010	24.229	25.290	0.465	0.466	0.475
acrylonitrile	33.000	26.303	28.179	0.478	0.479	0.471
bromobenzene	5.450	6.775	5.379	0.374	0.411	0.391
1-butanol	17.840	17.722	16.366	0.459	0.451	0.454
2-butanone	18.560	19.782	19.185	0.461	0.453	0.469
butyl phenyl ether	3.734	2.405	2.873	0.323	0.354	0.345
chlorobenzene	5.690	8.321	5.901	0.379	0.420	0.393
2-chloroethanol	25.800	19.914	18.927	0.472	0.471	0.461
3-chloro-1,2-propanediol	31.000	26.402	26.779	0.476	0.479	0.471
cyclohexanol	16.400	15.955	22.350	0.456	0.442	0.394
cyclohexyl acetate	5.080	3.540	3.067	0.366	0.346	0.352
cyclohexyl butanoate	4.580	1.397	2.160	0.352	0.324	0.328
cyclohexyl propanoate	4.820	2.211	2.494	0.359	0.334	0.337
dibenzyl ether	3.821	2.564	2.890	0.326	0.346	0.351
dibutyl ether	3.083	1.589	2.474	0.291	0.337	0.324
diethoxymethane	2.527	2.073	2.324	0.252	0.311	0.299
diethylene glycol	31.820	27.873	29.112	0.477	0.456	0.471
N,*N*-dimethylacetamide	38.850	27.800	29.986	0.481	0.494	0.475
dipropylamine	2.923	4.527	5.318	0.281	0.342	0.294
ethyl 2-bromopropanoate	9.400	3.647	3.543	0.424	0.386	0.370
ethyl methyl carbonate	2.985	1.136	2.143	0.285	0.340	0.326
1-fluorooctane	3.890	5.281	3.864	0.329	0.364	0.355
1-fluoropentane	3.931	8.052	5.726	0.331	0.390	0.385
1-hexanethiol	4.436	7.830	7.393	0.348	0.412	0.425
isobutyl chlorocarbonate	9.100	5.138	3.945	0.422	0.375	0.376
isobutyl vinyl ether	3.340	4.614	4.167	0.305	0.378	0.374
2-methyl-1,3-butadiene	2.098	3.404	4.192	0.211	0.249	0.248
2-methyl-2-butanethiol	5.087	8.671	8.005	0.366	0.410	0.429
3-methyl-2-butanone	10.370	20.778	20.494	0.431	0.455	0.470
2-methylcyclohexanone	14.000	16.014	14.131	0.448	0.434	0.461
methyl heptanoate	4.355	2.264	2.659	0.346	0.344	0.344
1-methylnaphthalene	2.915	2.857	3.927	0.280	0.224	0.246
methyl pentanoate	4.992	4.599	3.859	0.363	0.365	0.372
4-methyl-2-pentanone	13.110	16.239	14.542	0.445	0.436	0.459
1-nitrooctane	11.460	15.060	13.226	0.437	0.432	0.450
pentane	1.837	2.223	3.488	0.179	0.213	0.193
1,5-pentanediol	26.200	32.111	33.386	0.472	0.494	0.478
1-pentanethiol	4.847	8.786	8.098	0.360	0.421	0.432
1-pentene	2.011	2.871	3.874	0.201	0.242	0.229
propyl chlorocarbonate	11.200	6.277	4.653	0.436	0.387	0.386
1,3-propylene glycol	35.100	34.738	36.801	0.479	0.494	0.479
tribromoacetaldehyde	7.600	12.820	9.928	0.407	0.412	0.426
trichlorofluoromethane	3.000	5.372	3.849	0.286	0.324	0.337
trichloromethane	4.807	8.133	5.616	0.359	0.424	0.381
trichloronitromethane	7.319	9.122	6.638	0.404	0.364	0.373
tripropylamine	2.380	3.329	4.368	0.240	0.280	0.229

of electrons in the valence shell. Molecules that contain atoms with lone electron pairs (like oxygen, nitrogen, and halogen atoms) have higher descriptor values. The *image of the Onsager-Kirkwood solvation energy* is calculated by the eq 1, where μ is the dipole moment and M is the molecular weight. The *topological electronic index*[19] T^E is calculated over all pairs of bonded atoms according to eq 2, where q_i and q_j are the Zefirov's partial charges, and r_{ij} is the distance between bonded atoms i and j.

$$E_{Solv} = \frac{\mu^2}{M} \tag{1}$$

$$T^E = \sum_{(i<j)}^{N} \frac{|q_i - q_j|}{r_{ij}^2} \tag{2}$$

The other two descriptors, the *number of hydrogen-acceptor sites [Zefirov's PC] and fractional hydrogen bonding surface area [Semi-MO PC]*, involved in the model, describe hydrogen bonding interaction between the molecules. The *fractional hydrogen bonding surface area [Semi-MO PC]* is a sum of solvent accessible surface areas over hydrogen bonding donor and acceptor sites, normalized by the total molecular surface area. This descriptors accounts for the intermolecular hydrogen bonding interactions and is calculated according to eq 3, where S_{tot} is total molecular surface area and S_D and S_A are surface areas of hydrogen bonding donors and acceptors, respectively.

$$FHBSA = \frac{\sum_D S_D + \sum_A S_A}{S_{tot}} \tag{3}$$

The QSPR analysis for Kirkwood function resulted also in a six-parameter model with the R^2 of 0.9616, as summarized in Table 4 and Figure 2. The cross-validated R^2 for this model is 0.957. The RMS errors for training and prediction sets are 0.0187 and 0.0372, respectively. The molecular descriptors involved are *maximum atomic orbital electronic population*, the *number of hydrogen-acceptor sites [Zefirov's PC]*, *total dipole moment of the molecule*, *topological electronic index (all pairs) [Zefirov's PC]*, *maximum atomic charge for an H atom*, and *molecular weight*. Two descriptors, the *maximum atomic orbital electronic population* and the *number of hydrogen-acceptor sites*, are identical with the model for the dielectric constant. In addition, the *topological electronic index* is calculated according to eq 1; however, this index is calculated over all atom pairs in the molecule.

ARTIFICIAL NEURAL NETWORK MODELS

Artificial neural networks (NN) are novel and powerful technique to build models that can effectively solve complex real world problems. These techniques are loosely inspired by the way the densely interconnected, parallel structure of the brain processes information. Neural networks are constructed from a number of highly interconnected nonlinear processing units (also called neurons) that are joined together with weighted connections in several ways to form various types of networks. Probably the most widely used type is a feed-forward multilayer neural network. Another common type is the Kohonen or self-organizing map (SOM). These networks have the capacity to learn, memorize, and find relationships among the data. The most common tasks approached by the use of neural networks are modeling and classification problems. Numerous application areas in chemistry[20,21] include QSAR/QSPR studies, spectroscopy (IR, NMR, and UV spectras), protein folding, process control in chemical industry, etc.

The NN module from the CODESSA program[14] was used to calculate NN models. Before the neural network treatment started, both the experimental property (dielectric constant or Kirkwood function) and descriptor values were normalized to a range from 0.1 to 0.9. Feed-forward multilayer neural network models with a sigmoid activation function were chosen for the prediction of dielectric constant and Kirkwood function. All networks had one input layer, one hidden layer, and one output layer, and they were trained with a back-propagation algorithm.[22] The data set used as a training set

Table 3. Best Six Parameter Multilinear Regression Model for the Dielectric Constant ($R^2 = 0.9447$, $F = 421.45$, and $s^2 = 5.873$)

X	ΔX	t-test	descriptor
1462	1.355	1.078	intercept
95392	3.701	25.771	image of the Onsager-Kirkwood solvation energy
5860	0.424	13.818	the number of H-acceptor sites [Zefirov's PC]
42094	2.348	17.926	fractional HBSA (HBSA/TMSA) [Semi-MO PC]
−8419	0.963	−8.741	topographic electronic index (all bonds) [Zefirov's PC]
−3464	0.484	−7.170	total hybridization component of the molecular dipole
3607	0.782	4.614	max atomic orbital electronic population

Table 4. Best Six Parameter Multilinear Regression Model for the Kirkwood Function ($R^2 = 0.9616$, $F = 618.50$, and $s^2 = 0.0004$)

X	ΔX	t-test	descriptor
−0.01953	0.0118	−1.654	intercept
0.1953	0.00744	26.250	maximum atomic orbital electronic population
0.0327	0.00295	11.068	the number of H-acceptor sites [Zefirov's PC]
0.0216	0.00159	13.518	total dipole of the molecule
−0.0307	0.00343	−8.964	topographic electronic index (all pairs) [Zefirov's PC]
0.444	0.0476	−9.334	maximum net atomic charge for a H atom
−2.75e-4	4.80e-5	−5.727	molecular weight

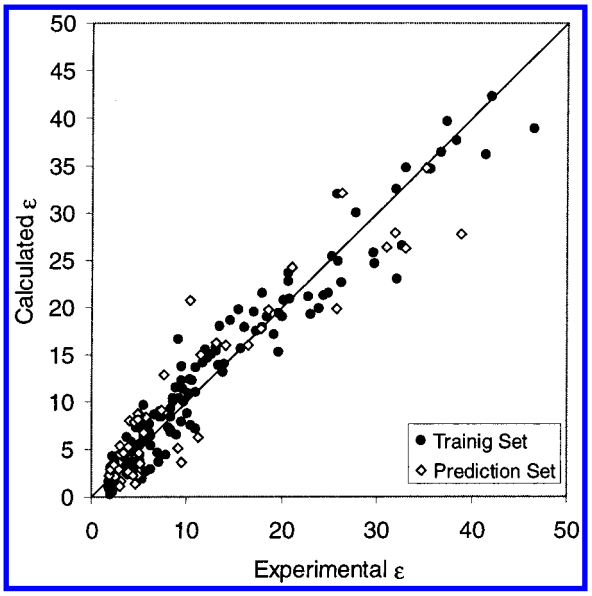

Figure 1. Plot of predicted vs experimental values of dielectric constant, multilinear regression model ($R^2 = 0.9447$).

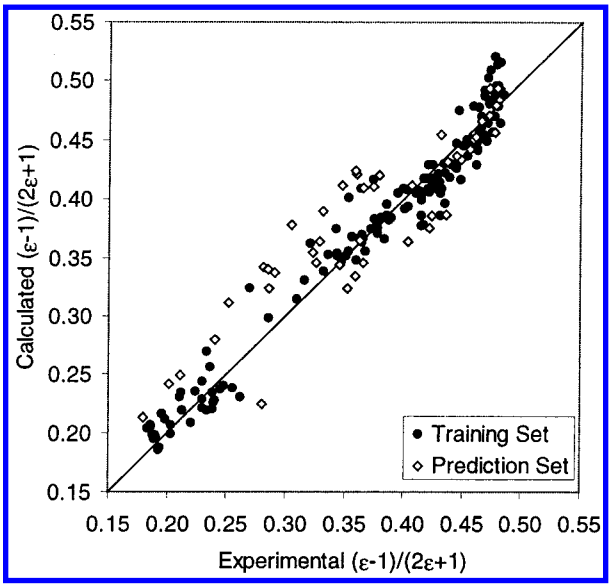

Figure 2. Plot of predicted vs experimental values of Kirkwood function, multilinear regression model ($R^2 = 0.9616$).

was the same as in the MLR treatment. However, from this data set 1/3 of structures were randomly selected and moved into separate validation set. The neural network weights were initialized with random values from the range between −0.5 and 0.5. According to the training set error, the neural network weights were then adjusted with back-propagation algorithm to minimize the prediction error. The validation set error was monitored by the training algorithm to perform automatic early stopping in order to avoid over-training of the neural network. Also, early stopping significantly reduces time that is spent to train the network.

Significant descriptors for NN models were selected using a stepwise forward selection algorithm. This algorithm starts by evaluating NN models ($1 \times 1 \times 1$) with one descriptor as an input parameter. From this step a number of models with lowest prediction errors was selected for the next step. In the next step, the number of hidden units was increased by one, and to each selected model from the previous step one additional input parameter was added from a pool of available descriptors. Again, the best models were selected, and this stepwise procedure was repeated until models with six input parameters were obtained. We found that five descriptor models ($5 \times 5 \times 1$) were sufficient both for the dielectric constant and Kirkwood function. The six descriptor models ($6 \times 6 \times 1$) did not give significant improvement over $5 \times 5 \times 1$ models. The inclusion of additional hidden units did not give significant imrovement either. The number of parameters in the model was kept as small as possible to avoid overfitting of the network.

The NN model (see Figure 3) for the dielectric constant consisted of the following descriptors: *fractional hydrogen bonding surface area FHBSA [Semi-MO PC]*, *image of the Onsager-Kirkwood solvation energy*, *topographic electronic index (all bonds) [Zefirov's PC]*, *total hybridization component of the molecular dipole*, and *the number of H-acceptor sites [Semi-MO PC]*. Importantly, all descriptors present in this NN model were also included in the best six-parameter linear model. The statistical fit of this NN model obtained is quite comparable to the MLR model. The training and prediction sets had RMS errors 2.29 and 3.548, respectively. The R^2 between predicted and experimental values for the training set is 0.948.

The NN model (see Figure 4) for the Kirkwood function consisted of the following descriptors: *difference in charged*

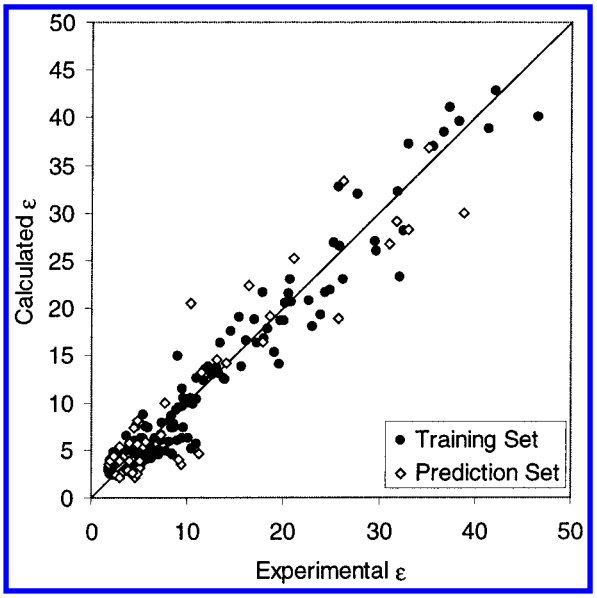

Figure 3. Plot of predicted vs experimental values of dielectric constant, neural network model ($R^2 = 0.948$).

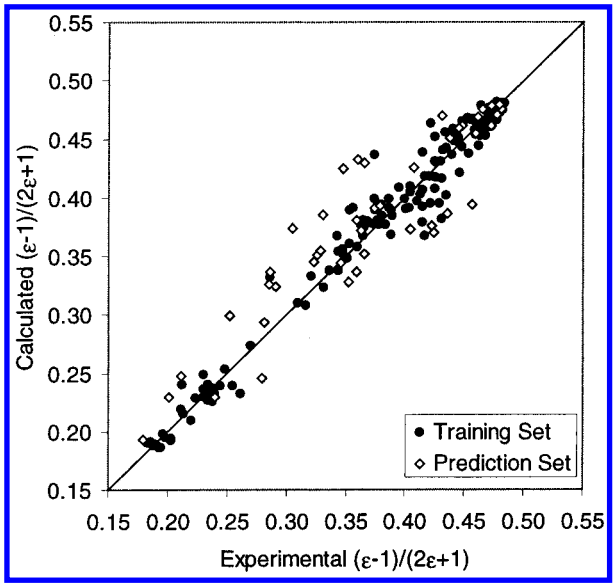

Figure 4. Plot of predicted vs experimental values of Kirkwood function, neural network model ($R^2 = 0.976$).

partial surface areas (DPSA-3) [Zefirov's PC], fractional hydrogen bonding surface area (FHBSA) [Semi-MO PC], *hydrogen bonding donors' surface area (HDSA2) [Semi-MO PC]*, *topographic electronic index (all bonds) [Zefirov's PC]*, and *total dipole moment of the molecule*. The training and prediction sets had RMS errors 0.0149 and 0.0341, respectively. The R^2 value for the training set is 0.976. Compared with MLR model of the Kirkwood function, this model is statistically better fit both for the training and prediction sets.

In the NN model for the Kirkwood function, only the *total dipole moment of the molecule* is identical descriptor to MLR model. However, the information covered by other descriptors is very similar to the descriptors used in the MLR model. For instance, topographic electronic index (T^E) is calculated over all bonds instead of all pairs of atoms. Also, the descriptors related to intermolecular interactions have slightly different definitions. The *difference in charged partial surface areas (DPSA-3*) is a charge distribution related descriptor that encodes information about polar interactions

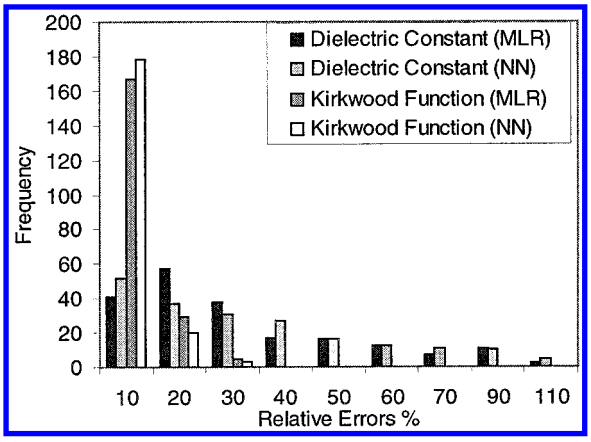

Figure 5. Distribution of relative prediction errors (all 201 compounds), calculated with multilinear regression (MLR), and neural network methods for the dielectric constant and the Kirkwood function.

between molecules. This descriptor is calculated according to eq 4, where q is the partial charge and S is the surface area of atoms. The *hydrogen-bonding donors' surface area* describes hydrogen-bonding interactions between the molecules. This descriptor is calculated according to eq 5, where S_{tot} is the total surface area of the molecule, q_D is the AM1 calculated partial charge of hydrogen-bonding donor (H) atom, and S_D is the surface area of this atom.

$$DPSA3 = \sum_i q_i^+ S_i - \sum_j q_j^- S_j \tag{4}$$

$$HDSA2 = \sum_D \frac{q_D \sqrt{S_D}}{\sqrt{S_{tot}}} \tag{5}$$

DISCUSSION

The analysis of descriptors included in all four QSPR models indicates that at least two different types of descriptors are necessary to predict dielectric constant and the Kirkwood function. The first class of descriptors is related to molecular electronic properties, e.g. *total molecular dipole moment*, *topographic electronic index*, and *maximal atomic orbital electronic population*. The second class of descriptors is necessary to describe intermolecular interactions (hydrogen bonding, electrostatic interactions between the molecules).

The NN models developed for the prediction of dielectric constant did not give noticeable improvement over the respective MLR models. Therefore, the MLR method seems to be more appropriate for the prediction of dielectric constant, since NN models demand much more computational time and labor. However, in the case of Kirkwood function, NN models performed better than the MLR model. The theoretical prediction of the dielectric constant for organic liquids is a complicated task that requires information both from the microscopic and macroscopic level. Obviously, the limiting factor here is not in the use of particular QSPR method but rather the molecular descriptors that fail to account for all the details of the underlying system (e.g. quantum-chemical descriptors from gas-phase calculations, inadequate descriptors to describe specific interactions between the molecules).

Table 5. Comparison of Different QSPR Models for the Dielectric Constant

model	training set R^2	training set RMS	prediction set RMS
present MLR	0.945	2.368	3.743
present NN	0.948	2.291	3.548
Cocchi et al. (MLR)	0.956	2.262	4.650
Cocchi et al. (PLS-GOLPE)	0.974	1.576	3.213
Schweitzer et al. (NN)	N/A	3.770	2.330

The comparison of QSPR models for the dielectric constant and Kirkwood function shows that the prediction of the Kirkwood function is more accurate. The histogram with relative errors is given in the Figure 5. The average prediction errors for the best dielectric constant model calculated for the entire data set of 201 compounds are 27.0% (MLR) and 29.1% (NN). The average prediction errors for the best Kirkwood function model are 5.4% (MLR) and 4.1% (NN). Thus, the prediction of Kirkwood function values from the predicted dielectric constant values would be much less reliable. The Kirkwood function values are calculated for each of 201 compounds from the predicted dielectric constant values by using both MLR and NN models result in the RMS errors 0.0923 and 0.0525, respectively. The respective RMS errors from predicted Kirkwood function values are 0.0242 and 0.0209.

The statistical fit of the MLR and NN models for the dielectric constant is between to other two previously published QSPR models (see Table 5). The R^2 values are smaller than Cocchi et al.[11] MLR and PLS models that were, however, derived for a much more restricted data set (training set of 23 structures and test set of 20 structures). The RMS error values are quite similar, except for the prediction set of Cocchi's MLR model and the training set of Cocchi's PLS model. Three descriptors used in their MLR model are also related to the electronic structure of the molecule and intermolecular interactions. Better results for smaller data sets suggest that more accurate predictions could be achieved when specific QSPR models are developed for compounds with common functionality, e.g. carbohydrates, alcohols, amines, etc. The RMS errors of present QSPR models are slightly smaller than Schweitzer et al.[12] NN model, which is quite expected since their data set was bigger. However, their models consisted of 10 descriptors. In addition to descriptors related to electronic structure and intermolecular interactions, their model also included several topological and constitutional descriptors. They reported that 11 compounds had absolute error values greater than 10, while 64 of the compounds had absolute error greater than 5. They also reported that 37 compounds had relative error larger than 100%, while 115 of the compounds had relative error larger than 50%.

CONCLUSION

The theoretical prediction of the dielectric constant and the Kirkwood function values from the chemical structure alone is feasible within certain error limits by using molecular descriptors encoding electronic properties of the molecule and intermolecular interactions between molecules. Both MLR and NN methods can provide acceptable models for the prediction of these properties. The QSPR models for the Kirkwood function appear to be more reliable than models for the dielectric constant. The average prediction error for the best dielectric constant model is 27.0%. The average prediction error for the best Kirkwood function model is 4.1%.

ACKNOWLEDGMENT

This work was partially supported by Estonian Science Foundation grant no. 4548.

REFERENCES AND NOTES

(1) Reichart, C. *Solvents and Solvent Effects in Organic Chemistry*; VCH: Weinheim, 1990.

(2) Karelson, M. In *Handbook of Solvents*; Wypych, G., Ed.; ChemTec Publishing: Toronto, 2001; Chapter 11, pp 639−682.

(3) Karelson, M. M.; Tamm, T.; Katritzky, A. R.; Cato, S. J.; Zerner, M. C. Application of Self-Consistent Reaction Field Method in Semi-empirical Quantum Chemical Calculations. *Tetrahedron Comput. Methodol.* **1989**, *2*, 295−304.

(4) Tapia, O.; Goscinski, O. Self-consistent Reaction Field Theory of Solvent Effects. *Mol. Phys.* **1975**, *29*, 1653.

(5) Katritzky, A. R.; Tamm, T.; Wang, Y.; Sild, S.; Karelson, M. QSPR Treatment of Solvent Scales. *J. Chem. Inf. Comput. Sci.* **1999,** *39*, 684−691.

(6) Katritzky, A. R.; Tamm, T.; Wang, Y.; Karelson, M. Unified Treatment of Solvent Properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 692−698.

(7) Fowler, F. W.; Katritzky, A. R.; Rutherford, R. J. D. The Correlation of Solvent Effects on Physical and Chemical Properties *J. Chem. Soc. (B)* **1971**, 460.

(8) Koppel, I.; Palm, V. In *Advances in Linear Free Energy Relationships*; Chapman, N. B., Shorter, J., Eds.; Plenum: London, 1972; pp 203−280.

(9) Tomasi J.; Mennucci, B.; Cappelli, C. In *Handbook of Solvents*; Wypych, G., Ed.; ChemTec Publishing: Toronto, 2001; Chapter 8, pp 419−504.

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

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

(12) Schweitzer, R. C.; Morris, J. B. The development of a quantitative structure property relationship (QSPR) for the prediction of dielectric constants using neural networks. *Anal. Chim. Acta* **1999**, *384*, 285−303.

(13) *CRC Handbook of Chemistry and Physics*, 81st ed.; Lide, D. R., Ed.; CRC Press: Boca Raton, FL, 2000.

(14) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA Reference Manual;* University of Florida; 1994.

(15) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.

(16) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(17) Zefirov, N. S.; Kirpichenok, M. A.; Izmailov, F. F.; Trofimov, M. I. Scheme for the Calculation of the Electronegativities of Atoms in a Molecule in the Framework of Sanderson's Principle. *Dokl. Akad. Nauk (Engl. Transl.)* **1987**, *296*, 883−887.

(18) Stewart, J. J. P. *MOPAC 6.0;* QCPE No 455, 1989.

(19) Osmialowski, K.; Halkiewicz, J.; Kaliszan, R. Quantitative Chemical Parameters in Correlation Analysis of Gas−Liquid Chromatographic Retention Indices of Amines. *J. Chromatogr.* **1986**, *361*,63−69.

(20) Burns, J. A.; Whitesides, G. M. Feed forward neural networks in chemistry: Mathematical systems for classification and pattern recognition. *Chem. Rev.* **1993**, *93*, 2583−2601.

(21) Svozil, D; Kvasnicka, V; Pospíchal, J Introduction to multilayer feed-forward neural networks. *Chemometrics Intelligent Laboratory Systems* **1997**, *39*, 43−62.

(22) Mitchell, T. M. *Machine Learning;* McGraw-Hill: New York, 1997.

CI010335F