applied to model the retention indices of other halogenated biphenyls. In addition, the LC retention indices of these compounds can also be modeled, although the model found was not as good as the one for GC retention indices. A set of more sophisticated descriptors might be able to support better models. A descriptor to encode the tendency to form hydrogen bonds is one example. With the availability of more retention data, these models can be improved by using more observations to develop the regression equations. Retention indices measured on other columns, i.e. those that have different polarities, should also be tested. Finally, the possibility of predicting elution order should also be considered since it would be of practical value of chromatographers.

LITERATURE CITED

(1) Hutzinger, O.; Safe, S.; Zitko, V. *The Chemistry of PCBs*; CRC Press: Boca Raton, FL, 1980.
(2) Carter, L. J. *Science* **1976**, *192*, 240-243.
(3) Robertson, L. W.; Chynoweth, D. P. *Environment* **1975**, *17* (6), 25-27.
(4) de Kok, J. J.; de Kok, A.; Brinkman, J. A. Th.; Kok, R. M. *J. Chromatogr.* **1977**, *142*, 367-383.
(5) Robertson, L. W.; Safe, S. H.; Parkinson, A.; Pellizari, E.; Pochini, C.; Mullin, M. D. *J. Agric. Food Chem.* **1984**, *32*, 1107-1111.
(6) Erickson, M. D. *Analytical Chemistry of PCBs*; Butterworth: Stoneham, MA, 1986.
(7) Mullin, M. D.; Pochini, C. M.; McCrindle, S.; Romkes, M.; Safe, S. H.; Safe, L. M. *Environ. Sci. Technol.* **1984**, *18*, 468-476.
(8) Bush, B.; Murphy, M. J.; Connor, S.; Snow, J.; Barnard, E. *J. Chromatogr. Sci.* **1985**, *23*, 509-515.
(9) Onuska, F. I.; Terry, K. A. *HRC & CC, J. High Resolut. Chromatogr. Chromatogr. Commun.* **1986**, *9*, 671-675.
(10) Fischer, R.; Ballschmiter, K. *Fresenius' Z. Anal. Chem.* **1988**, *332*, 441-446.
(11) McFarland, V. A.; Clarke, J. U. *Environ. Health Perspect.* **1989**, *81*, 225-239.
(12) Parkinson, A.; Safe, S. In *Polychlorinated Biphenyls (PCBs): Mammalian and Environmental Toxicology*; Safe, S., Ed.; Springer-Verlag: Berlin, 1987; pp 49-75.
(13) Sissons, D.; Welti, D. *J. Chromatogr.* **1971**, *60*, 15-32.
(14) Robbat, A.; Xyrafas, G.; Marshall, D. *Anal. Chem.* **1988**, *60*, 982-985.
(15) Devillers, J. *Fresinius' Z. Anal. Chem.* **1988**, *332*, 61-62.
(16) Hasan, M. N.; Jurs, P. C. *Anal. Chem.* **1988**, *60*, 978-982.
(17) Höfler, F.; Melzer, H.; Möckel, J.; Robertson, L. W.; Anklam, E. *J. Agric. Food Chem.* **1988**, *36*, 961-965.
(18) Almenningen, A.; Bastiansen, O.; Fernholt, L.; Cyvin, B. N.; Cyvin, S. J.; Samdal, S. *J. Mol. Struct.* **1985**, *128*, 59-76.
(19) Rømming, C.; Seip, H. M.; Aanesen Øymo, I.-M. *Acta Chem. Scand., Ser. A* **1974**, *28*, 507-514.
(20) Field, L. D.; Skelton, B. W.; Sternhell, S.; White, A. H. *Aust. J. Chem.* **1985**, *38*, 391-399.
(21) Dynes, J. J.; Baudais, F. L.; Boyd, R. K. *Can. J. Chem.* **1985**, *63*, 1292-1299.
(22) McKinney, J. D.; Gottschalk, K. E.; Pedersen, L. *J. Mol. Struct.* **1983**, *104*, 445-450.
(23) Tsuzuki, S.; Tanabe, K.; Nagawa, Y.; Nakanishi, H.; Osawa, E. *J. Mol. Struct.* **1988**, *178*, 277-285.
(24) Jaime, C.; Font, J. *J. Mol. Struct.* **1989**, *195*, 103-110.
(25) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(26) Kier, L. B. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109-116.
(27) Rohrbaugh, R. H.; Jurs, P. C. *Anal. Chim. Acta* **1987**, *199*, 99-109.
(28) Abraham, R. J.; Smith, P. E. *J. Comput. Chem.* **1987**, *9*, 288-297.
(29) Stanton, D. R.; Jurs, P. C. Unpublished results.
(30) Draper, N.; Smith, H. *Applied Regression Analysis*, 2nd ed.; Wiley-Interscience: New York, 1981; pp 307-312.
(31) Furnival, G. M.; Wilson, R. W., Jr. *Technometrics* **1974**, *16*, 499-511.
(32) Snee, R. D. *Technometrics* **1977**, *19*, 415-428.

RECEIVED for review April 16, 1990. Accepted July 30, 1990. This work was partially supported by the National Science Foundation Grant CHE-8815785. The Sun workstation was purchased with partial financial support of the National Science Foundation. The financial support of the University Teknologi Malaysia and the Malaysian Government is acknowledged.

Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure–Property Relationship Studies

David T. Stanton and Peter C. Jurs*

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

Intermolecular interactions that are polar in nature contribute to observed physicochemical properties such as chromatographic retention and normal boiling point. However, these types of interactions are difficult to encode with structural parameters currently available for use in SPR studies. A new series of molecular structural parameters have been developed that combine molecular surface area and partial atomic charge information to form charged partial surface area (CPSA) descriptors. These descriptors have been shown to be useful in a variety of structure–property studies. The characteristics and properties of these parameters are discussed, and their use in several structure–property studies is described.

INTRODUCTION

There are many applications of computer-assisted quantitative structure–property relationships (QSPR) that are of value in analytical chemistry. Such relationships act as tools to augment experimental analytical techniques, allowing the chemist to extract additional information. These computer-assisted techniques can be used as an aid to understanding chemical processes such as chromatographic retention. They are also of value in the identification of materials when authentic standards are not available or quantities of available materials are limited. In addition, computer-assisted QSPRs can save time by allowing the analyst to quickly estimate a property for a given molecule which might be too time-consuming to measure experimentally. However, because of limitations inherent to methods used to study the relationship between molecular structure and physical property, QSPR studies often focus on sets of compounds that are relatively similar. Also, limitations associated with the parameters used in QSPR in the past have restricted the accuracy of predictions and the variety of properties that can be studied. Thus, it is of interest to expand the utility of QSPR to a larger variety of compounds, to increase accuracy, and to extend the use of

0003-2700/90/0362-2323$02.50/0

QSPR to new properties of analytical interest.

In the computer-assisted study of the relationship between the structure of molecules and their physicochemical properties, several different approaches are used to characterize the structure of the molecule in terms that can be manipulated by a computer. The structural information used by the computer is in the form of numerical values termed structural *descriptors*. The information contained in these descriptors reflects the nature of the particular type of molecular representation used. For example, a graph theoretical representation describes the molecule as a set of vertices (atoms) and edges (bonds). Several descriptors found to be useful in QSPR studies which are based on graph theoretical representations of molecules have been reported in the literature (*1–3*). A more sophisticated representation views a molecule as a collection of nuclei bound together by overlapping electron orbitals. Such a representation can be used to derive electronic information such as partial atomic charges and dipole moments (*4, 5*). Another structural representation characterizes a molecule as a set of hard spheres connected by bonds possessing specific stretching, bending, and torsional energies. The shape of the molecule is determined by the strain placed on the bonds as the spheres are allowed to interact. Information that is related to the geometry (size and shape) of the molecule can be obtained from such a representation (*6–8*). The conclusions that can be drawn from these examples are that several useful, but significantly different, representations can be used to describe a molecule and that no single one is best. Rather, the true representation must lie in some combination of them, and some combinations of additional representations that have not yet been conceived. One of the limitations one encounters in QSPR studies is the lack of information concerning the particular structural features responsible for the observed differences in physical properties within a set of molecules. This information deficiency is related directly to the type of molecular representation employed. Thus, one goal of QSPR studies is to expand the set of representations in order to derive new descriptors which allow the chemist to better understand the link between structure and property.

In a previous QSPR study, regression equations were developed for the prediction of gas chromatographic retention indices for a series of 107 pyrazines on both a nonpolar OV-101 stationary phase and a polar Carbowax-20M stationary phase (9). Among other conclusions of that study, it was determined to be more difficult to predict the chromatographic retention for such compounds on the polar stationary phase. This had also been previously noted by Kaliszan (*10*). The difficulty is presumed to stem from the lack of information contained in the available structure-based descriptors concerning those features responsible for polar intermolecular interactions. Thus, new descriptors were sought that would improve the correlation of the structures of analytes to their retention characteristics on polar stationary phases.

The search for new parameters was based on a particular concept of how molecules might interact, particularly when the interactions are polar. Specific and nonspecific polar interactions greatly affect many physical properties, and the strength of these interactions was considered to be a function of the amount of contact between molecules. Also, the charge distribution within the molecules would act as the driving force for these interactions and should be considered in any new parameter studied. In the past, some success has been achieved by using information concerning the solvent-accessible surface area of molecules to describe intermolecular interactions that are important for different properties. For example, Camilleri et al. (*11*) reported the use of solvent-accessible surface area to correlate molecular structure with partition coefficients for a wide variety of benzene derivatives. Modifications of surface area calculations have also been reported. Amidon and co-workers (*12*) studied the correlation of surface area, in terms of hydrocarbon portions (HYSA) and functional group portions (FGSA), with the aqueous solubility of a series of monofunctional aliphatics. Dunn et al. (*13, 14*) reported similar experiments involving the division of solvent-accessible surfaces into isotropic (nonpolar) portions and hydrated (polar or solvated) portions for the prediction of partition coefficients. Iwase et al. (*15*) employed corrections for hydrophilic interactions and hydration of polar moieties in the correlation of solvent-accessible surface area and hydrophobicity. More recently, Grigoras (*16*) has reported work involving molecular surfaces expressed in the form of electrostatic molecular surface interaction terms. These structural surface descriptors include information concerning molecular surface area, partial atomic charges, and charge scaling factors. He showed these descriptors to be useful in the prediction of such properties as critical temperature and critical volume. We decided to investigate this type of structural descriptor in more detail in an attempt to find new ways of encoding the structural features responsible for polar intermolecular interactions, which are needed for a variety of QSPR studies.

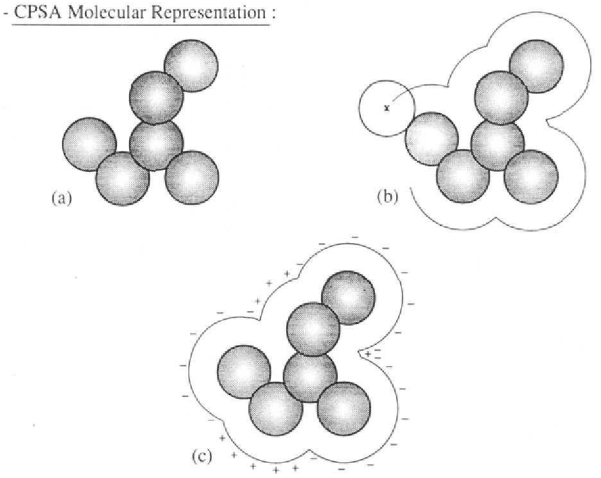


Figure 1. Diagram of the molecular representation used to derive the charged partial surface area (CPSA) descriptors.

We report here a new set of molecular structure descriptors termed charged partial surface area (CPSA) descriptors that encode features responsible for polar interactions between molecules. The molecular representation used here views a molecule as having a surface defined by the overlap of hard spheres, defined by the van der Waals radii of the atoms (Figure 1a), which is traced by a sphere representing a solvent molecule (water by default). The surface traced out by the center of the solvent sphere has been termed the solvent-accessible surface (6) and is shown diagrammatically in Figure 1b. The molecule is further defined as having a specific electron distribution, thus yielding a representation of a charged contact surface (Figure 1c) where polar intermolecular interactions can take place.

The CPSA descriptors derived from such a representation have been shown to be useful in correlations of structural features with physical properties such as chromatographic retention, boiling point, and surface tension. The utility and characteristics of the CPSA descriptors are discussed below.

EXPERIMENTAL SECTION

All computations in this study were performed on a Sun 4/110 workstation running the ADAPT software system under the UNIX operating system. The general procedure for the development of regression equations from structural descriptors using ADAPT has been outlined previously (*17, 18*). The solvent accessible surface area calculations were performed using the SAVOL

Table I. Experimentally Derived Data Used in the Structure-Property Study of Surface Tension

compound name	surface tension, dyn cm^{-1}	compound name	surface tension, dyn cm^{-1}
decane	23.74	nitrobenzene	42.70
cyclohexane	25.16	methoxybenzene	35.00
benzene	28.81	*N,N*-dimethylanaline	36.04
toluene	28.52	dimethyl phthalate	41.84
1-chlorotetradecane	30.36	1-chloronaphthylene	41.04
1-bromo-tetradecane	31.17	nitromethane	37.36
		dichloromethane	27.89
acetonitrile	29.29	dibromomethane	39.74
cyanoethane	27.24	1,2-dihydroxyethane	48.49
1-cyanobutane	27.39	1,2-dimethoxyethane	24.61
1-cyanohendecane	29.82	1,4-dioxane	33.26
methanol	22.49	pyridine	37.24
ethanol	22.32	tetrahydrofuran	27.31
2-butanol	23.37	cyclohexanone	35.05
1-hexanol	26.20	dimethyl adipate	35.86
chlorobenzene	32.96		
bromobenzene	35.80		

algorithm of Pearlman (*6*), which was modified to store the surface area data for the individual atoms of a molecule in a computer file. The partial atomic charges were obtained by using a modified version of the algorithm of Abraham and Smith (*4*). This is an empirical method of calculating partial charges, which includes both σ and π contributions and which is parameterized to reproduce experimental dipole moments. The algorithm has been implemented in FORTRAN in our laboratory as part of the ADAPT system. In addition, the method has been improved and the parameter set has been expanded to include a greater variety of atom and bond types. Calculations of the surface areas and atomic partial charges included hydrogens.

The Data Sets. The data used in this study were obtained from several sources. The gas chromatographic retention data for the series of 107 pyrazines on Carbowax-20M were taken from work previously reported by the authors (*9*). Normal boiling points (boiling point recorded at 760 mmHg) and structures for a set of 209 furans and tetrahydrofurans were obtained as ASCII files from The Beilstein Institute (supplementary material). The structures, in the form of connection tables were converted to ADAPT format with a FORTRAN program and were stored on disk. Reasonable low-energy conformations for the molecules were obtained by using molecular mechanics calculations (*17, 19*). Data for the study of surface tension were taken from Körösi and Kováts (*20*). The structures were entered into the ADAPT files by sketching them on a graphics terminal. These structures were also placed in reasonable low energy conformations using molecular mechanics calculations. Surface tension data recorded at 20 °C (Table I) were used in the modeling process.

Descriptor Analysis. Once the structures had been stored, and desired sets of structural descriptors had been calculated, the process of descriptor analysis or objective feature selection was begun. This is an objective analysis of the calculated descriptors which is necessary to reduce the number of available descriptors in order to avoid obtaining good correlations purely by chance. The analysis involves the detection and removal of those structural descriptors which exhibit high pairwise correlations with other descriptors, or which contain little discriminatory information. Pairs of descriptors that are highly correlated ($r \geq 0.950$) contain nearly identical information, and one of the pair should be discarded. Descriptors that contain a high percentage (≥90%) of identical values or a small number (≤10%) of nonzero values are of little use and are discarded. The results of this type of analysis are a reduced pool of information-rich descriptors that can then be screened in regression analysis.

Regression Analysis. Multiple linear regression techniques are used for descriptor selection rather than other methods such as factor analysis or principle components regression, because of their ease of implementation and the interpretability of the resulting equations. In most cases, least-squares multiple linear regression analysis was carried out by using a process of stepwise addition with progressive deletion (*21*). Regression analysis of this form proceeds by a stepwise addition of terms, based on *F* statistic values, which is followed by a deletion process. In the deletion process, each variable (descriptor) is held out in turn, and the stepwise addition process is repeated on the remaining pool of variables. Then each combination of two, and then three, variables are held out, and the regression is repeated. This combination of steps has the effect of uncovering potentially superior equations that may have been otherwise obscured by the existence of a descriptor that was highly correlated to the dependent variable. The final models were chosen on the basis of the values for the coefficient of multiple determination (*R*) and standard deviation of the regression (*s*). The exception to this process was the study involving the correlation of molecular structure descriptors and surface tension. For this study, the process of best subsets regression (*22*) was used as implemented in the Minitab statistical package (*23*). In this process, all possible combinations of 1-15 variables are regressed together, and the best equations are reported. Selection of the final equation in this case was based on the C_p criterion (*22*).

Each model chosen from regression analysis is examined for statistical significance (overall *F* test) and for the significance of the individual descriptors (partial *F* test). Also, the models are examined for collinearities between two or more of the descriptors. Since it is common for descriptors within models of this type to be correlated to some extent, it is important that any collinearities that do exist be noted and that they are not causing any serious problems (e.g., increased uncertainty in the regression coefficients). Statistical orthogonality of the descriptors is not a requirement.

RESULTS AND DISCUSSION

A set of 25 individual descriptors was assembled that combines information on the atomic contributions to the solvent-accessible surface area of the molecule with partial atomic charge information. This particular set was originally chosen to allow the examination of a variety of combination schemes. However, it was found in subsequent studies that the descriptors contain different information from each other such that several CPSA descriptors are often combined within a single regression equation. It was also noted that different combinations of these descriptors become important depending on the physical property and data set being considered. The equations used to calculate these descriptors are given in Table II. The values are grouped by charge type and by the method used to combine the surface area and charge information. Thus, there are three partial positive surface area descriptors (PPSAs) and an equal number of the partial negative surface area descriptors (PNSAs). Also, there is a set of three difference in partial surface area descriptors (DPSAs), six fractional charged surface area (FPSA and FNSA) descriptors (three positive and three negative), and a similar set of six total surface weighted partial surface area (WPSA and WNSA) descriptors. In addition to the charged surface area parameters, it was of interest to examine the relative influence of the most highly charged (positive and negative) atom on the overall charge of the molecule. These descriptors are the relative positive and relative negative charges (RPCG and RNCG, respectively). This information was then combined with the accessible surface area of the most highly charged atoms to obtain the relative positive and relative negative charged surface area descriptors (RPCS and RNCS, respectively).

To illustrate the calculation of the CPSA descriptors, the necessary data for methyl acetate, as an example compound, is given in Table III. The surface area and charge data for each atom in the molecule are given along with the total solvent accessible surface area and greatest positive and negative partial atomic charges. Table IV gives the values of the individual CPSA descriptors derived from these data.

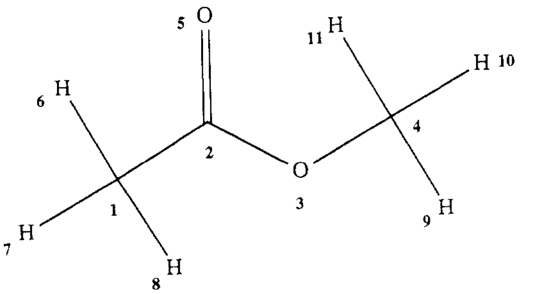
Chromatographic Retention. The first application of the CPSA descriptors in an QSPR study involved the correlation of molecular structure and chromatographic retention on Carbowax-20M for the 107 pyrazines. The set of 25 CPSA

Table II. Formulas Used in the Calculation of the CPSA Molecular Structure Descriptors

descriptor	descriptor label		formula[a,b]
partial positive surface area	PPSA-1		$\sum(+SA_i)$
partial negative surface area	PNSA-1		$\sum(-SA_i)$
total charge weighted PPSA	PPSA-2		$(\sum(+SA_i))Q^+_T$
total charge weighted PNSA	PNSA-2		$(\sum(-SA_i))Q^-_T$
atomic charge weighted PPSA	PPSA-3		$\sum(+SA_i)(Q^+_i)$
atomic charge weighted PNSA	PNSA-3		$\sum(-SA_i)(Q^-_i)$
difference in charged partial surface areas	DPSA-1		[(PPSA-1) - (PNSA-1)]
	DPSA-2		[(PPSA-2) - (PNSA-2)]
	DPSA-3		[(PPSA-3) - (PNSA-3)]
fractional charged partial surface areas	FPSA-1	FNSA-1	(charged partial surface area)/(total molecular surface area)
	FPSA-2	FNSA-2	
	FPSA-3	FNSA-3	
surface weighted charged partial surface areas	WPSA-1	WNSA-1	(CPSA)(total mol surf. area)/1000
	WPSA-2	WNSA-2	
	WPSA-3	WNSA-3	
relative positive charge	RPCG		(charge of most positive atom)/(sum total positive charge)
relative negative charge	RNCG		(charge of most negative atom)/(sum total negative charge)
relative positive charged surface area	RPCS		$(SA_{MPOS})(RPCG)$
relative negative charged surface area	RNCS		$(SA_{MNEG})(RNCG)$

[a] $(+SA_i)$ and $(-SA_i)$ are the surface area contributions of the ith positive or negative atom in the molecule. [b] Q^+_i and Q^-_i are the partial atomic charges for the ith positive and negative atoms, while Q^+_T and Q^-_T are the sum total positive and negative charges for the molecule.

Table III. Data Used for the Example Calculations of the CPSA Descriptors for Methyl Acetate



total solvent-accessible surface area = 246.7 Å^2
partial atomic charge on most positive atom = 0.3194
partial atomic charge on most negative atom = −0.3818

atom no.	solvent accessible surface area, Å^2	partial charge
1	10.1	-0.042
2	7.3	0.319
3	16.4	-0.212
4	11.4	0.002
5	39.6	-0.382
6	27.0	0.049
7	27.5	0.049
8	27.6	0.049
9	30.7	0.056
10	24.6	0.056
11	24.5	0.056

Table IV. Values of the CPSA Descriptors Calculated for Methyl Acetate

descriptor label	descriptor value	descriptor label	descriptor value
PPSA-1	180.7	FNSA-2	-0.171
PPSA-2	115.0	FNSA-3	-0.077
PPSA-3	10.8	WPSA-1	44.6
PNSA-1	66.1	WPSA-2	28.4
PNSA-2	-42.1	WPSA-3	2.67
PNSA-3	−19.0	WNSA-1	16.3
DPSA-1	114.6	WNSA-2	-10.4
DPSA-2	157.1	WNSA-3	-4.70
DPSA-3	29.9	RPCG	0.502
FPSA-1	0.732	RNCG	0.600
FPSA-2	0.466	RPCS	3.64
FPSA-3	0.044	RNCS	23.8
FNSA-1	0.268		

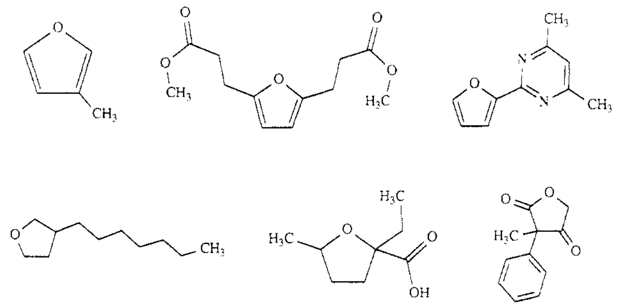


Figure 2. Examples of the compounds included in the furan and tetrahydrofuran boiling point QSPR study training set.

descriptors were combined with the variables contained in the previously reported Carbowax-20M model. After elimination of high pairwise correlations, the remaining descriptors were regressed against the retention indices for the pyrazines on Carbowax-20M. For the purpose of comparison to the previous work, the best six-variable predictive equation obtained as given in Table Va. Examination of the new model shows that three of the six descriptors are from the CPSA set. The new model yields a slightly better fit to the experimental data (R^2 = 0.988) and a 9.5% reduction in the standard deviation of the regression (s = 32.9). However, the best overall model obtained contained nine variables, six of which are of the CPSA set. The nine-variable model is shown in Table Vb. A significant improvement in the fit of the equation is obtained (R^2 = 0.994) along with a decrease of 26.5% in the standard deviation of regression (s = 26.7) as compared with the previous model. This brings the quality of the Carbowax-20M model more in line with that of the OV-101 model previously reported. Although some improvement in the fit of equations involving CPSA descriptors was expected, the number and combination of the new CPSA descriptors in the resulting models were unforeseen.

Boiling Point. As a result of the success obtained with the CPSA descriptors in modeling retention, it was of interest to study the application of the new descriptors in the modeling of other properties. In a QSPR study involving correlations of molecular structure with normal boiling point for a diverse

Table V

(a) A Six-Variable Regression Equation Obtained for the Correlation of the Molecular Structure and Gas Chromatographic Retention Indices of 107 Pyrazines on Carbowax-20M Employing CPSA Descriptors

$R = 0.994$ ($R^2 = 0.988$), $s = 32.9$ (2.0% of mean), $n = 107$
F value for analysis of variance = 1515.9, $F(0.90;6,100) = 1.77$

independent variable	regression coeff	std dev of regr coeff	partial F value
PNSA-1	4.68	0.44	113.74
PNSA-2	4.61	0.46	99.55
DPSA-3	19.55	1.34	213.10
no. of single bonds	−103.57	4.10	639.68
molecular polarizability	94.36	3.51	722.47
1st moment of inertia (included hydrogens)	0.031	0.0078	16.04
intercept	−272.65		

(b) A Nine-Variable Regression Equation Obtained for the Correlation of Molecular Structure and Gas Chromatographic Retention Indices of 107 Pyrazines on Carbowax-20M Employing CPSA Descriptors

$R = 0.997$ ($R^2 = 0.994$), $s = 26.7$ (1.6% of mean), $n = 107$
F value for analysis of variance = 1538.0, $F(0.90;6,100) = 1.77$

independent variable	regression coeff	std dev of regr coeff	partial F value
PNSA-2	3.11	0.54	32.76
DPSA-2	−0.72	0.11	40.05
DPSA-3	27.91	2.22	158.17
FNSA-2	−826.66	232.65	12.63
FPSA-3	−8776.53	1488.63	34.76
RNCS	−11.33	2.47	20.99
no. of single bonds	−116.51	5.32	479.63
molecular polarizability	94.26	3.49	729.56
1st moment of inertia (included hydrogens)	0.095	0.0079	144.78
intercept	265.98		

set of 209 furan and tetrahydrofuran related compounds, some difficulty was encountered in obtaining a good fit to experimental values. This was due in part to the great diversity of the types of substituents in the data set. To illustrate this diversity, some of the compounds from the training set are shown in Figure 2. Since the goal of the study was to model the data set as a whole, it was not possible to subset the compounds to reduce the problems related to diversity of functional groups involved. Given that properties such as the normal boiling point are functions of the strength of various intermolecular interactions in the liquid phase, it was thought that the CPSA descriptors would yield better results, based on the results obtained in the retention study of the pyrazines.

A final model based on a set of 209 furans and tetrahydrofurans was developed and was found to give good results for this diverse data set. This model, shown in Table VI, contains 11 variables. A plot of the correlation of the fitted and observed boiling points for the 209 furans and tetrahydrofurans is given in Figure 3. Results of this quality had not been obtained in the study prior to the inclusion of the CPSA descriptors to the pool available to the regression program. Five of the 11 descriptors in the model are of the CPSA variety. Again, we found it interesting that so many of the variables in the model were of the CPSA variety. We also noted that none of the five CPSA descriptors in the boiling point model were the same as those used in the pyrazine retention index model described above. This is presumably a function of the data set and property being studied. It is apparent from the examination of the gas chromatographic retention and boiling point QSPR equations that there is a significant difference in the kind of structural information contained in the individual CPSA descriptors.

Table VI. Regression Model for the Correlation of Molecular Structure and Normal Boiling Point of a Series of 209 Furans and Tetrahydrofurans Using CPSA Descriptors

$R = 0.984$ ($R^2 = 0.968$), $s = 11.2$ (5.7% of mean), $n = 209$
F value for analysis of variance = 554.5, $F(0.90;11,198) = 1.55$

independent variable	regression coeff	std dev of regr coeff	partial F value
PPSA-1	0.61	0.05	133.92
PPSA-3	10.78	1.04	108.40
FNSA-3	−1296.78	55.76	540.82
WPSA-3	−26.87	2.89	86.47
RPCG	145.05	13.00	124.41
path-1 simple molecular connectivity	61.61	4.46	191.05
no. of single bonds	−25.36	1.12	514.18
LUMO energy (from Hückel method)	11.52	2.97	15.07
valence corrected path-3 molecular connectivity	22.94	3.63	39.85
square root of KAPPA-3 (*24*)	47.15	7.31	41.57
av distance sum connectivity (*25*)	21.14	3.99	28.01
intercept	−280.94		

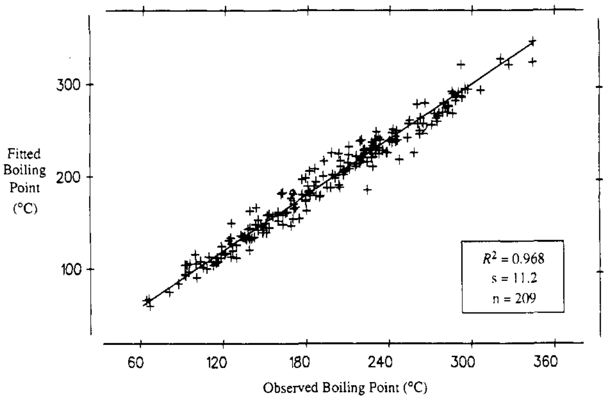


Figure 3. Correlation of the fitted and observed normal boiling points for the 209 furans and tetrahydrofurans in the boiling point QSPR study.

Surface Tension. It is a common practice in QSPR studies to use a large set of compounds to derive regression equations. This is done in order to include as much information as possible concerning the molecules under study in the resulting model. However, such large data sets are not always available, and the chemist must rely on the strength of the descriptors to encode as much information as is contained in a less than ideal data set. Since the CPSA descriptors have shown utility in studies involving quite diverse sets of compounds, it was of interest to apply them to a smaller than usual data set in order to determine the limits of our QSPR technique using the new descriptors. It was also of interest to extend the use of computer-assisted QSPR to a physical property that had not previously been studied in this fashion and is also a function of the strength of various intermolecular interactions. Examination of the literature indicated that surface tension was such a property.

By use of the methods described above, a study involving a small and very diverse series of 31 organic compounds with known surface tension values (Table I) was initiated. A regression equation that employed six descriptors was found to yield good results. This model is detailed in Table VII, and the correlation between fitted and observed surface tension

Table VII. Regression Model Obtained for the Study of the Correlation of Molecular Structure and Surface Tension Using CPSA Descriptors

$R = 0.953$ ($R^2 = 0.908$), $s = 2.32$ (6.9% of mean), $n = 31$
F value for analysis of variance = 39.14, $F(0.90;6,25) = 2.02$

independent variable	regression coeff	std dev of regr coeff	partial F value
FNSA-2	-29.16	3.64	64.18
FPSA-3	387.62	39.81	94.79
RPCS	-1.29	0.19	45.66
RNCS	0.28	0.06	23.39
7th order, valence corrected chain molecular connectivity[a]	106.71	22.82	21.86
topological symmetry measure[b]	-9.51	2.95	10.41
intercept	12.87	2.02	

[a] See ref 26. [b] See text.

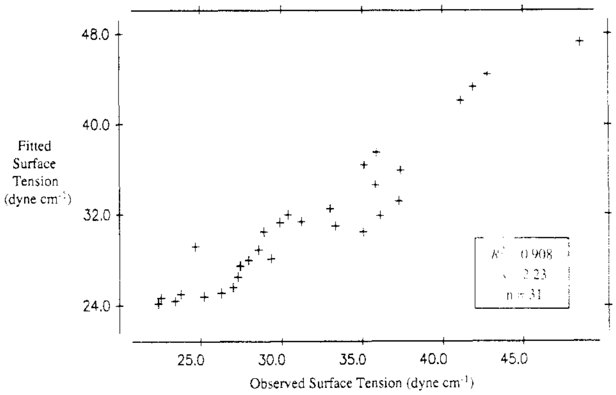


Figure 4. Correlation of the fitted and observed values for the surface tension QSPR study.

is shown graphically in Figure 4. The standard deviation of the regression (s) for this case represents 6.9% based on the mean surface tension for this data set. As was noted in other QSPRs, the majority of the information concerning the structures of the compounds in question is obtained from the CPSA type descriptors. The other descriptors within the model are topological in nature. The 7th order valence-corrected chain molecular connectivity describes the structure around the rings (*26*), while the topological symmetry describes the structural symmetry of a molecule and is calculated as the number of unique atoms in a molecule divided by the total number of atoms. A unique atom is defined as an atom having a unique chemical environment over a given bond distance. A bond distance of 1 was used in this case.

Within the model, information concerning features important for polar interactions is supplied solely by the CPSA parameters. Also, examination of the results of the regression analysis indicates that the CPSA descriptors account for the greatest amount of the variance in the training set (largest partial F values). It was also noted that only one of the four CPSA descriptors in the surface tension model had appeared in either the boiling point or retention index prediction models discussed above. While the quality of these results is not as good as that obtained for the previous QSPR examples discussed, they indicate the potential of the CPSA descriptors in the study of surface tension, and a more detailed study involving a larger data set is now in progress.

Other Characteristics of the CPSA Descriptors. Since the new CPSA descriptors were of obvious importance in the QSPR studies described, it was necessary to determine the properties of the descriptors themselves. Points of interest included degree of information overlap (intercorrelation), sensitivity to changes in conformation of molecules under study, and importance of inclusion of the solvent in the calculation of the surface area.

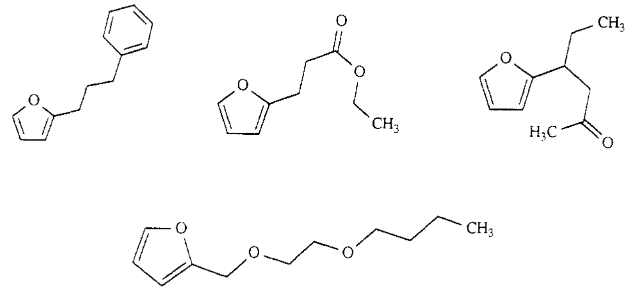


Figure 5. Structures of the furan analogues used to study the conformation dependence of the CPSA descriptors.

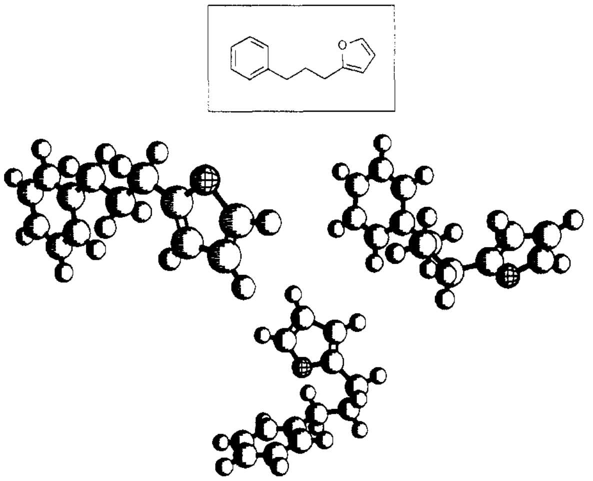


Figure 6. Conformations of 2-(propylphenyl)furan used in the study of the conformation dependence of the CPSA descriptors.

Table VIII. Average Percent Change Observed in the CPSA Descriptors as a Function of Changes in the Geometry of Four Furan Molecules

descriptor label	av % change	descriptor label	av % change
PPSA-1	2.2	FNSA-2	4.9
PPSA-2	2.2	FNSA-3	8.0
PPSA-3	2.3	WPSA-1	4.3
PNSA-1	7.3	WPSA-2	4.3
PNSA-2	7.3	WPSA-3	4.3
PNSA-3	9.0	WNSA-1	9.6
DPSA-1	5.9	WNSA-2	9.6
DPSA-2	5.9	WNSA-3	11.0
DPSA-3	4.1	RPCG	0.0
FPSA-1	2.0	RNCG	0.0
FPSA-2	2.0	RPCS	20.3
FPSA-3	2.2	RNCS	21.0
FNSA-1	4.9		

The test of sensitivity to changes in conformation was accomplished by using a set of four furan analogues (Figure 5). Each compound was placed in three different energy minimized conformations. Examples of the conformations used for the first of the test compounds are given in Figure 6. The series of 25 CPSA descriptors were then calculated for each conformation and the descriptor values were then compared. The results of the comparison are given in Table VIII. The descriptors exhibiting the largest overall variation with changing conformation are the RPCS and RPNS descriptors. Since these parameters encode information related to a single atom, changes in conformation of a molecule would have a relatively large effect on these descriptors. Also, since the remaining CPSA descriptors are based on the surface area of several atoms, changes in surface area are averaged out, thus reducing the conformational dependence of these descriptors. Another consideration is the topological nature of the method used for calculating the partial atomic charges. Geometry is not a factor in the computation of the atomic partial charges used and thus they have a stabilizing effect on the resulting CPSA descriptors. The overall average change in the set of CPSA descriptors is 6.1%. Given that the differences in conformations were large, this value seems to be reasonable.

Another aspect of the CPSA descriptors to be considered is their connection to a solvent-accessible surface. It was of interest to determine what effect the removal of the solvent in the surface area calculation would have on the usefulness or information content of the CPSA descriptors. Normally, the calculation of the surface area is done by using a sphere with a radius of 1.5 Å to approximate the contact surface formed when a water molecule interacts with the query molecule. For the test of the effects of solvent removal, the radius of the solvent was set to zero. This results in the use of the van der Waals surface of the molecule. The study of the effect of removal of the solvent was done with the set of 107 pyrazines previously described. Calculation of the descriptors and subsequent objective feature selection and regression analysis was done as previously outlined. The first test involved the direct replacement the solvent-included CPSA descriptors in the model shown in Table Va with the corresponding solventless descriptors, followed by the recalculation of the regression coefficients. Examination of the model obtained in this fashion shows that the substitute descriptors yield a poorer fit to the experimental data (R^2 = 0.984) and a larger standard deviation of regression (s = 39.3 or 2.4% of the mean retention index) as compared with the model containing the solvent-included descriptors. Also, the standard deviation of the regression coefficients increased as well. Other tests involved building models in the usual fashion with the substitute descriptors. These equations yielded generally poorer fits to the experimental data and also exhibited a decrease in the number of CPSA type descriptors in the models. It was also noted that the descriptors based on the van der Waals surface were more highly correlated between themselves than are the normal CPSA descriptors. These results suggest that the quantity of structural information in the substitute descriptors has decreased, and that the solvent-accessible surface is an important factor in describing the molecular features which are involved in intermolecular interactions. Thus, it can be concluded that the solvent-accessible surface represents a useful approximation of the contact surface involved in such interactions between molecules.

A final comparison of the solvent-included and solventless CPSA descriptors for the 107 pyrazines involved the examination of the correlation between the two sets of descriptors. The average correlation for the two sets is 0.887, with a high of 1.00 (RPCG and RNCG) and a low of 0.406 (RPCS). The high correlations for RPCG and RNCG are expected since these descriptors contain no surface area information, and such values will be identical regardless of the surface area data used. The results of this comparison suggest that different structural information is encoded in the solvent-included CPSA descriptors. This information taken together with the significant role the CPSA descriptors have played in the three SPR studies described above indicates the importance of using the solvent-accessible surface area.

CONCLUSIONS

The CPSA descriptors represent a new set of parameters for use in QSPR studies. They appear to provide important information concerning features of molecules that are responsible for strong intermolecular interactions, which are in turn important for helping to correlate structural features to physical properties such as chromatographic retention, boiling point, and surface tension. These descriptors allow for the use of smaller sets of compounds, thus extending the use of QSPR techniques to problems were the availability of experimentally derived values is limited. Through the application of the CPSA descriptors to problems of analytical importance, it may be possible to gain additional insight into the nature of solvent–solute interactions in chromatography and allow for the extension of QSPR techniques to broader sets of compounds and a larger number of analytically important properties.

ACKNOWLEDGMENT

We are grateful to Steve Dixon for his work involving the program used to calculate the partial atomic charges used in this study.

Supplementary Material Available: Identification numbers, data, and structures for the 209 furans and tetrahydrofurans (27 pages). Photocopies of the supplementary material from this paper or microfiche (105 × 148 mm, 24× reduction, negatives) may be obtained from Microforms & Back Issues Office, American Chemical Society, 1155 16th Street, NW, Washington, DC 20036. Orders must state whether for photocopy or microfiche and give complete title of article, names of authors, journal issue date, and page numbers. Prepayment, check or money order for $46.00 for photocopy ($48.00 foreign) or $10.00 for microfiche ($11.00 foreign), is required and prices are subject to change.

LITERATURE CITED

(1) Wiener, H. *J. Am. Chem. Soc*. **1947**, *69*, 2636.
(2) Randić, M. *J. Am. Chem. Soc*. **1975**, *97*, 6609.
(3) Randić, M. *J. Chem. Inf. Comp. Sci*. **1984**, *24*, 164–175.
(4) Abraham, R. J.; Smith, P. E. *J. Comput. Chem*. **1988**, *9*, 288–297.
(5) Ośmiałowski, K.; Halkiewisz, J.; Radecki, A.; Kaliszan, R. *J. Chromatogr*. **1985**, *346*, 53–60.
(6) Pearlman, R. S. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980.
(7) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950; pp 144–156.
(8) Jinno, K.; Kawasaki, K. *Chromatographia* **1983**, *17*, 445.
(9) Stanton, D. T.; Jurs, P. C. *Anal. Chem*. **1989**, *61*, 1328–1332.
(10) Kaliszan, R. *Quantitative Structure-Chromatographic Retention Relationships*; Wiley-Interscience: New York, 1987.
(11) Camilleri, P.; Watts, A.; Boraston, J. A. *J. Chem. Soc., Perkin Trans. 2*, **1988**, 1699–1707.
(12) Amidon, G. L.; Yalkowsky, S. H,; Anik, S. T.; Valvani, S. C. *J. Phys. Chem*. **1975**, *21*, 2239–2246.
(13) Dunn, W. J., III; Koehler, M. G.; Grigoras, S. *J. Med. Chem*. **1987**, *30*, 1121–1126.
(14) Koehler, M. G.; Grigoras, S.; Dunn, W. J., III *Quant. Struct.-Act. Relat*. **1988**, *7*, 150–159.
(15) Iwase, K.; Komatsu, K.; Hirono, S.; Nakagawa, S.; Moriguchi, I. *Chem. Pharm. Bull*. **1985**, *33*, 2114–2121.
(16) Grigoras, S. *J. Comput. Chem*. **1990**, *11*; 493–510.
(17) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(18) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979, pp 103–129.
(19) Allinger, N. L.; Yuh, Y. H. *Molecular Mechanics, Operating Instructions for MM2 and MMP2 Programs, 1977 Force Field*; Quantum Chemistry Program Exchange, 1980; Program 395.
(20) Körösi, G.; Kováts, E. *J. Chem. Eng. Data* **1981**, *26*, 323–332.
(21) Small, G. W.; Jurs, P. C. *Anal. Chem*. **1983**, *55*, 1121–1127.
(22) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*, 2nd. ed.; Irwin: Homewood, IL, 1985.
(23) *Minitab Reference Manual*, release 7.2; Minitab: State College, PA, 1989.
(24) Kier, L. B. *Quant. Struct.-Act. Relat*. **1985**, *4*, 109.
(25) Balaban, A. T. *Chem. Phys. Lett*. **1982**, *89*, 399–404.
(26) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Letchworth, Herts, England, 1986.

RECEIVED for review May 11, 1990. Accepted August 3, 1990. The funding for this work was provided by the Beilstein Institute. Partial funding was also provided by the National Science Foundation for the purchase of the Sun 4/110 computer system.