Quantum-Mechanical QSAR/QSPR Descriptors from Momentum-Space Wave Functions

Errol F. McCoy and Matthew J. Sykes*

School of Chemistry, Physics and Earth Sciences, Flinders University, GPO Box 2100, Adelaide, SA, Australia 5001

Received September 2, 2002

It is shown that quantum-mechanical descriptors obtained as parameters from the one-dimensional radial distribution function of electron momentum can be used to predict molecular activities or properties to a precision that compares favorably with the more traditional QSAR/QSPR methods. The distribution function is derived from momentum space ab initio wave functions. The predictive value of the descriptors is illustrated by their application to the estimation of McGowan's volume, gas-chromatographic retention time, gas−hexadecane partition coefficient, second hyperpolarizability, and tadpole narcotic activity.

INTRODUCTION

The commercial exploitation of an organic compound—as a medicinal drug for example—is likely to require, at some stage of its development, the determination of biological or chemical activities or properties related to the intended end use of the compound. It is desirable therefore to have at hand relatively straightforward and inexpensive procedures enabling the efficient and accurate prediction of a molecular activity or property—especially when its direct measurement by experiment is, for one reason or another, to be avoided if at all possible. The procedures that are conventionally used for indirect determinations of activities make use of molecular "descriptors" which include suitable molecular properties and physical-organic constructs obtained from both experimental and computational sources. Molecular descriptors are ultimately related to molecular structure. Hence the relationships between activities and the descriptors on which they depend are generally known as Quantitative Structure−Activity Relationships (QSARs) or Quantitative Structure−Property Relationships (QSPRs) depending upon whether the property of interest is to be characterized as biological or nonbiological, respectively. Much of the early development of QSAR theory is due to the work of Hansch and co-workers inspired by the pioneering work of Hansch and Fujita.[1]

Over the past 30 years, a good deal of attention has been given to the development of molecular descriptors which have been categorized as conventional or quantum-chemical.[2] The conventional descriptors have included relatively simple properties of the isolated molecule such as molecular weight, mean polarizability, and molecular volume.[3] Descriptors derived using graph theory have also been used with remarkable success. An example is the Wiener index obtained by summing all interatomic distances.[4] More sophisticated descriptors have emerged in recent years from the laboratories of Kamlet, Abraham, Taft, and others.[5] Many of these descriptors have been designed to allow for solute−solvent interactions and have been used to estimate properties such as hexadecane−gas and olive oil−gas partition coefficients,[6] tadpole narcosis,[7] and brain-blood barrier permeation.[8]

Most QSAR/QSPR studies use a combination of conventional and quantum-mechanical descriptors. For instance, Katritzky et al.[2] used a combination of four quantum-mechanical and two conventional descriptors to estimate gas−liquid chromatography (GLC) retention times. The almost endless array of descriptors that has been proposed has led to the development of software capable of handling large data sets and many descriptors. For example, the programs CODESSA (Comprehensive Descriptors for Statistical and Structural Analysis) and ADAPT (Automated Data Analysis and Pattern recognition Toolkit) were developed for this purpose by the Katritzky and Jurs groups, respectively.

In a related approach,[9−11] momentum-space electron density distributions were used to generate molecular similarity measures. For instance, Cooper et al.[9] used similarity measures for the total densities of a group of phospholipids to predict, with reasonable success, their ability to inhibit HIV.

The structure of a molecule determines its electronic wave function which in turn determines many of its physicochemical properties. It is reasonable to suppose that the biological activity of a molecule—its potency as a drug for example—will also, often enough, be dependent upon the molecular electronic wave function. Given that it is now possible to perform accurate ab initio calculations routinely on molecules of moderate size at reasonable cost, it would clearly be advantageous to have available a procedure for estimating a molecular activity or property from descriptors which are derived from, and in turn characterize, the electronic wave function of a molecule.

It is possible in principle to determine any molecular property that is ultimately dependent upon the molecule's electronic ground state wave function. In practice this can sometimes be achieved using rigorous procedures. It can,

* Corresponding author phone: +44-114-222-9529; e-mail: m.j.sykes@sheffield.ac.uk. Present address: Department of Chemistry, University of Sheffield, Sheffield, UK, S3 7HF.

10.1021/ci025597b CCC: $25.00
Published on Web 01/10/2003

for instance, be achieved for certain properties using quantum-mechanical expectation values. But when rigorous procedures are tedious or difficult (as with say molecular hyperpolarizability) or virtually impossible (as with say anaesthetic potency or GLC retention times), then it is necessary to resort either to less rigorous model calculations or to the use of QSAR or QSPR methods. The latter approach is explored in this article, using descriptors obtained from accurately determined molecular electronic wave functions. A preliminary report of the proposed method and its application to estimating polarizability and hyperpolarizability was given by McCoy and Sykes.[12]

AN OUTLINE OF THE METHOD

The method used in this study for obtaining wave-mechanical descriptors may be summarized as follows:

(1) The electronic ground-state wave functions of the molecules of interest are computed using ab initio methods. The wave functions so obtained are functions of the electron position coordinates. It is convenient to refer to them as r-space (position-space) wave functions.

(2) Using Fourier transform methods, the r-space functions are transformed to the equivalent k-space (momentum-space) functions. This is done for two main reasons. First of all, and most importantly, the electron density distribution of a molecule in k-space (obtained as the square of the modulus of the k-space wave function) possesses an inversion center at the k-space origin regardless of the r-space point group to which the molecule belongs. Thus, the k-space density distribution has a uniquely defined origin. In r-space, the density distribution does not in general have a uniquely defined origin (unless of course the molecule happens to be centrosymmetric in r-space). The second reason for transforming to k-space is that, in k-space, the more interesting part of the electron density distribution (from the viewpoint of biological activity or molecular property estimation) is more conveniently and more compactly located near to the k-space origin. The corresponding part of the density distribution in r-space is associated with the outermost valence regions of the molecule (since the momentum of an electron is relatively large when it is near to a nucleus and approaches zero when it is distant from any nucleus).

(3) By integration of the k-space electron density distribution over the polar angle coordinates, a radial distribution function in k-space is obtained. The k-space radial distribution function of a molecule, which we denote by $D_2(k)$, is typically smooth, unimodal and skewed toward the origin—but is otherwise featureless.

(4) That part of the distribution between the origin and the peak of the distribution is curve-fitted using a low-degree polynomial. It has been our experience that four coefficients are generally sufficient to obtain a good fit to the low-k end of the k-space radial distribution function. The polynomial coefficients so obtained are used as molecular descriptors.

(5) For a given collection of molecules with known ("observed") values for the biological activity or molecular property under consideration, multiple linear regression is used to obtain a QSAR (or QSPR) which expresses the relationship between the property, or a suitable function of the property such as its logarithm, and the four k-space molecular descriptors described above. The QSAR or QSPR so obtained can then be used to predict the property for an unknown molecule whose k-space descriptors are known.

It is perhaps worth mentioning that the method outlined above is distinctly different to the approach taken by Cooper et al.[9] who sought to quantify the similarity between pairs of molecules in terms of a generalized overlap integral between their total electron density distributions in momentum space. Our approach, on the other hand, has been to extract potentially useful descriptors from the total electron density distribution in momentum space, averaged over the polar angles. What the two methods depend on and evidently share however is an implicit realization that the density distribution in momentum space is centrosymmetric with a clearly defined origin at zero momentum.

THE ELECTRON DENSITY DISTRIBUTIONS IN R-SPACE

In the independent-particle approximation, the ground state electronic wave function of a molecule is an antisymmetrized product of molecular orbitals (MOs). Molecular orbitals are commonly expressed as linear-combinations of basis functions which, in r-space, are real, atom-centered functions. The *i*th MO in r-space may thus be written as

$$\phi_i(\boldsymbol{r}) = \sum_{\mu} c_{\mu i}\chi_{\mu}(\boldsymbol{r} - \boldsymbol{r}_{\mu}) \tag{1}$$

In this equation, the $c_{\mu i}$ are MO expansion coefficients and the χ_{μ} are basis functions. $\boldsymbol{r}_{\mu}$ is the position of the atom center associated with the μth basis function.

In ab initio MO methods, which include Hartree−Fock (HF), higher level post-HF methods, and methods based on density functional theory (DFT), the MO expansion coefficients are determined by the linear variation method. These procedures are implemented in the popular GAUSSIAN 94 program which was used in this study.[13]

The basis functions used in ab initio methods are Gaussian Type Functions (GTF) which are categorized as primitive or contracted Gaussians. Contracted gaussians are linear combinations of primitive gaussians with fixed expansion coefficients

$$\chi(\boldsymbol{r} - \boldsymbol{r}_{\mu}) = \sum_{\mathrm{s}} d_{\mu \mathrm{s}}\theta_{\mathrm{s}}(\boldsymbol{r} - \boldsymbol{r}_{\mu}) \tag{2}$$

where $\theta_{\mathrm{s}}(\boldsymbol{r} - \boldsymbol{r}_{\mu})$ is a primitive Gaussian function, and the $d_{\mu \mathrm{s}}$'s are fixed constants within a given basis set. Combining eqs 1 and 2, we obtain

$$\phi_{\mathrm{i}}(\boldsymbol{r}) = \sum_{\mu}\sum_{\mathrm{s}} c_{\mu i} d_{\mu \mathrm{s}}\theta_{\mathrm{s}}(\boldsymbol{r} - \boldsymbol{r}_{\mu}) \tag{3}$$

Primitive r-space Gaussians are functions of the form

$$\mathrm{g}_{lmn}(\alpha;\boldsymbol{r}) = N_{lmn}(\alpha)x^l y^m z^n \exp(-\alpha r^2) \tag{4}$$

where $N_{lmn}(\alpha)$ is a normalizing constant. The exponents l, m, and n are integers, and α is a constant which determines the size (radial extent) of the function. The normalizing

Table 1. Fourier Transforms $\Theta_s(\boldsymbol{k})$ of Gaussian Primitives $\theta_s(\boldsymbol{r})$[a]

$\theta_s(\boldsymbol{r})$	$\Theta_s(\boldsymbol{k})$
$g_{000}(\alpha;\boldsymbol{r})$	$G_{000}(\beta;\boldsymbol{k})$
$g_{100}(\alpha;\boldsymbol{r})$	$-\mathrm{i}\, G_{100}(\beta;\boldsymbol{k})$
$g_{010}(\alpha;\boldsymbol{r})$	$-\mathrm{i}\, G_{010}(\beta;\boldsymbol{k})$
$g_{001}(\alpha;\boldsymbol{r})$	$-\mathrm{i}\, G_{001}(\beta;\boldsymbol{k})$
$g_{110}(\alpha;\boldsymbol{r})$	$-G_{110}(\beta;\boldsymbol{k})$
$g_{101}(\alpha;\boldsymbol{r})$	$-G_{101}(\beta;\boldsymbol{k})$
$g_{011}(\alpha;\boldsymbol{r})$	$-G_{011}(\beta;\boldsymbol{k})$
$g_{200}(\alpha;\boldsymbol{r})$	$(2/\sqrt{3})G_{000}(\beta;\boldsymbol{k}) - \mathrm{G}_{200}(\beta;\boldsymbol{k})$
$g_{020}(\alpha;\boldsymbol{r})$	$(2/\sqrt{3})G_{000}(\beta;\boldsymbol{k}) - \mathrm{G}_{020}(\beta;\boldsymbol{k})$
$g_{002}(\alpha;\boldsymbol{r})$	$(2/\sqrt{3})G_{000}(\beta;\boldsymbol{k}) - \mathrm{G}_{002}(\beta;\boldsymbol{k})$

[a] Note that $G_{lmn}(\beta;\boldsymbol{k})$ is generally not the transform of $g_{lmn}(\alpha;\boldsymbol{r})$. The constants α and β are related by $\alpha\beta = 1/4$.

constants for the primitive Gaussians required for this work are

$$N_{000}(\alpha) = (2^3\alpha^3/\pi^3)^{1/4} \tag{5a}$$

$$N_{001}(\alpha) = N_{010}(\alpha) = N_{100}(\alpha) = (2^7\alpha^5/\pi^3)^{1/4} \tag{5b}$$

$$N_{011}(\alpha) = N_{101}(\alpha) = N_{110}(\alpha) = (2^{11}\alpha^7/\pi^3)^{1/4} \tag{5c}$$

$$N_{002}(\alpha) = N_{020}(\alpha) = N_{200}(\alpha) = (2^{11}\alpha^7/9\pi^3)^{1/4} \tag{5d}$$

Primitive Gaussians in momentum space are similarly defined

$$G_{lmn}(\beta;\boldsymbol{k}) = N_{lmn}(\beta)k_x{}^l k_y{}^m k_z{}^n \exp(-\beta \mathrm{k}^2) \tag{6}$$

where k_x, k_y, and k_z are the components of the momentum variable $\boldsymbol{k}$ and β is a constant. The normalizing constant $N_{lmn}(\beta)$ in k-space is similar in form to the normalizing constant in r-space $N_{lmn}(\alpha)$.

THE TRANSFORMATION TO MOMENTUM SPACE

The Fourier transform[14] of the ith molecular orbital $\phi_i(\boldsymbol{r})$ is defined by

$$\Phi_i(\mathbf{k}) = (2\pi)^{-3/2}\int_{\mathrm{all}\,\mathbf{r}} \exp(-i\mathbf{k}\cdot\mathbf{r})\phi_i(\mathbf{r})\mathrm{d}\mathbf{r}$$

$$= \sum_{\mu}\sum_{\mathrm{s}} c_{\mu i} d_{\mu \mathrm{s}} \exp(-i\mathbf{k}\cdot\mathbf{r}_\mu)\Theta_{\mathrm{s}}(\mathbf{k}) \tag{7}$$

where $\Theta_s(\boldsymbol{k})$ is the Fourier transform of the r-space Gaussian primitive $\theta_s(\boldsymbol{r})$. When atomic units are used, the variables $\boldsymbol{r}$ and $\boldsymbol{k}$ represent position and momentum, respectively. The Fourier transform pairs relevant to this study are given in Table 1.

The electron density in r-space at some point $\boldsymbol{r}$ is given by

$$\rho(\boldsymbol{r}) = \sum_{\mathrm{i}} \lambda_i \phi_i(\mathbf{r})\phi_i{}^*(\boldsymbol{r}) \tag{8}$$

where λ_i is the occupancy number (0, 1, or 2) of the ith MO. The asterisk denotes complex conjugation. Similarly, the electron density in k-space at some point $\boldsymbol{k}$ is given by

$$\rho(\boldsymbol{k}) = \sum_{i} \lambda_i \Phi_i(\boldsymbol{k})\Phi_i{}^*(\boldsymbol{k}) \tag{9}$$

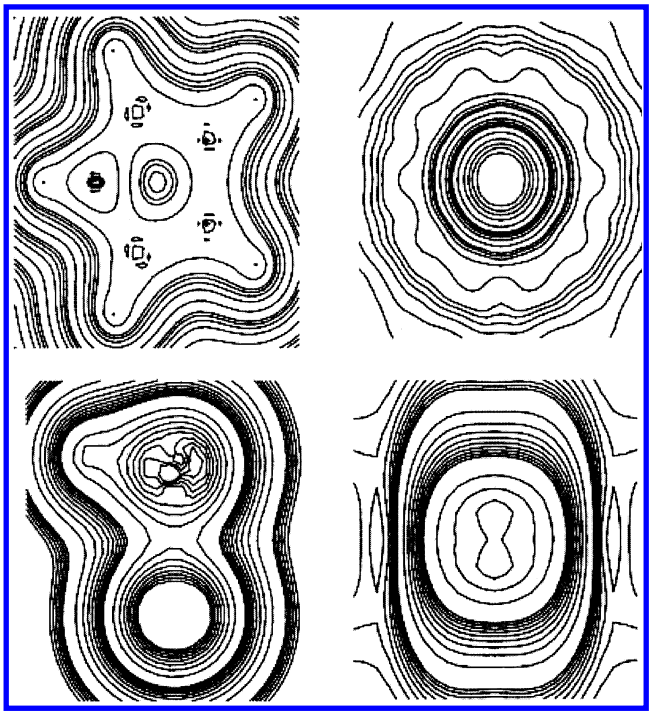

Figure 1. Comparisons of position space (left) and momentum space (right) valence densities. The densities are shown for pyrrole at the top and hydrogen oxyfluoride at the bottom.

It is worth noting that, with the Fourier transform defined as in eq 7, Parseval's theorem[14] ensures that normalization of the wave function is preserved on transformation to k-space.

It follows that

$$\int_{-\infty}^{+\infty} \rho(\boldsymbol{r})\, \mathrm{d}\boldsymbol{r} = \int_{-\infty}^{+\infty} \rho(\boldsymbol{k})\, \mathrm{d}\boldsymbol{k} \tag{10}$$

It is a property of Fourier transforms that, since $\phi(\boldsymbol{r})$ is a real function, then $\Phi(\boldsymbol{k})$ is complex with real and imaginary parts that are even and odd, respectively.[14] It follows that $\Phi(\mathbf{k})\Phi^*(\mathbf{k})$ and therefore $\rho(\boldsymbol{k})$ are even functions. That is, the electron density distribution in k-space possesses an inversion center regardless of the point symmetry group (in r-space) to which the molecule belongs. This is evident in Figure 1 which compares the valence electron density distributions in r-space and k-space for representative molecules.

THE RADIAL DISTRIBUTION FUNCTION IN MOMENTUM SPACE

We define the distribution function $D_n(k)$ by

$$D_n(k) = \int_0^{\pi} \mathrm{d}\theta \int_0^{2\pi} \mathrm{d}\phi\, k^n \sin\theta\, \rho\,(k) \tag{11}$$

$$\rho(\boldsymbol{k}) = \rho(k_x,k_y,k_z) = \rho(k\sin\theta\cos\phi,\, k\sin\theta\sin\phi,\, k\cos\theta) \tag{12}$$

where k, θ, and ϕ are spherical coordinates in momentum space. The momentum space radial distribution function is the function $D_2(k)$. The integration is facilitated if use is made of the inversion symmetry to write

$$D_n(k) = 4\int_0^{\pi/2} \mathrm{d}\theta \int_0^{\pi} \mathrm{d}\phi\, k^n \sin\theta\, \rho\,(k) \tag{13}$$

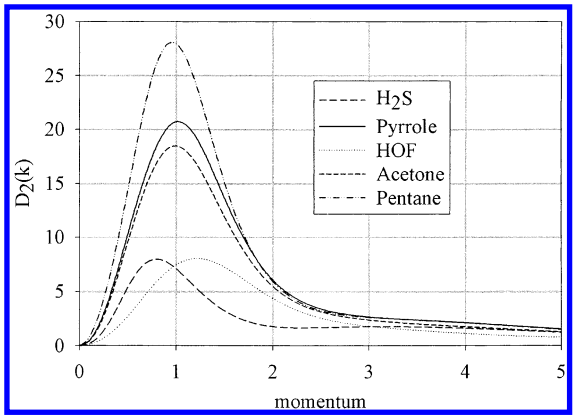

Figure 2. Five exemplar plots of spherically averaged momentum distributions, $D_2(k)$. All values are in atomic units.

Table 2. Comparisons of Calculated Moments of Momentum m_0 and $m_2/2$ with Corresponding Quantities N and $-E$[a]

molecule	N	m_0	$-E$	$m_2/2$
hydrogen sulfide	18	17.99	399.39	398.91
pyrrole	36	36.00	210.18	209.38
hydrogen oxyfluoride	18	18.00	175.54	173.53
acetone	32	32.00	193.17	192.55
pentane	42	42.00	197.78	198.84
nitrogen	14	14.00	109.53	108.56
sulfur hexafluoride	70	69.99	997.15	988.42
naphthalene	68	68.00	385.91	384.98
methanol	18	18.00	115.73	115.26
trichloroethylene	64	63.96	1457.37	1453.62

[a] The molecules were all geometry optimized and their wave functions determined using the hybrid density functional B3LYP and the 6-31+G(d) basis set.

We have evaluated this integral by numerical integration using 16-point Gauss-Legendre quadrature.[15] The value of the integral was determined in most cases for $k = 0.1$, 0.2, 0.3, ..., 5.0 atomic units. Plots of the radial distribution function in k-space for a number of representative molecules are given in Figure 2.

Two simple numerical checks are available to test the accuracy of the integration procedure. If we define the moments of momentum about the origin by the relation

$$m_n = \int k^n \rho(k)\mathrm{d}k \tag{14}$$

then the zeroth moment m_0 is the number N of electrons in the molecule and the second moment m_2 is twice the average electron kinetic energy $\bar{T}$ (in atomic units) of the molecule.

$$N = m_0 = \int k^0 \rho(k)\mathrm{d}k = \int_0^\infty D_2(k)\mathrm{d}k \tag{15}$$

$$\bar{T} = \frac{1}{2}m_2 = \frac{1}{2}\int k^2 \rho(k)\mathrm{d}k = \frac{1}{2}\int_0^\infty D_4(k)\mathrm{d}k \tag{16}$$

According to the virial theorem $\bar{T}$ is equal in magnitude but opposite in sign to the total electron energy E which, in ab initio calculations, is generally calculated using the r-space wave function. The results of applying these tests to a number of molecules is given in Table 2. There is evidently very good agreement between the corresponding quantities although it is apparent that the values of $m_2/2$ are consistently lower than the corresponding values of $-E$.

The differences between $m_2/2$ and $-E$ are almost certainly due to a lack of precision in the numerical integration procedures. It is reasonable to expect that integration over the polar angles (eq 13) using a 16-point Gauss-Legendre quadrature formula is a reasonably accurate procedure and that the observed discrepancies are more likely to be attributable to the integrations required by eq 16. The latter integral was evaluated using Simpson's rule with a step-size for k of 0.1 atomic units. The upper limit of integration was set equal to 200 atomic units of momentum for the test molecules in Table 2. Bearing in mind that an s-type core electron has a high probability of being found close to an atomic center, with a correspondingly very high momentum, it is quite possible that the upper limit of integration was set too low for the complete numerical integration of $D_4(k)$. It may also be the case that the integration step size was too large to achieve high precision. In any event, because of the reasons mentioned, we consider that the differences (ca. 1%) are quite small and, for the purposes of QSAR/QSPR studies, are not considered to be of any concern.

THEORETICAL MODEL CHEMISTRIES AND BASIS SETS

A theoretical model chemistry is a defined level of theory which is capable of being applied uniformly to a collection of molecules in order to explore structures, energies, and other physical properties with the use of a computer program. The model is tested by systematic comparison with experiment and, if the comparison is favorable, the model then acquires predictive value in situations where experimental data are not available.[16] The models that are currently popular are essentially ab initio and utilize the independent particle (orbital) approximation. Defining the level of theory requires decisions about the method (Hartree−Fock (HF), post-HF, density functional theory (DFT), etc.) and about the size and quality of the basis set from which the orbitals are constructed. The choice is inevitably a tradeoff between accuracy and computational cost.

In this study we anticipated the need to perform accurate computations on a reasonably large number of molecules of moderate size. Accordingly we were attracted to the DFT methods which are becoming increasingly more popular on account of their reputed accuracy and applicability to molecules of moderate size at a fraction of the cost required for HF and post-HF methods. We chose for our structure−activity studies to use the B3LYP density functional together with the 6-31+G(d) basis set. On the "down" side, however, we were conscious of the fact that DFT is a relative newcomer to the molecular modeling armament and that the general accuracy of DFT generated k-space wave functions is as yet an unknown quantity. The accuracy of the k-space functions is by no means guaranteed by merely acknowledging the documented successes of DFT based on r-space wave functions.

From Table 2 it is apparent that, notwithstanding a slight systematic discrepancy between the values of $-E$ and $m_2/2$, the virial theorem is satisfied and, to that extent, the accuracy of the DFT generated k-space functions is supported. It is perhaps worth noting that the orbitals obtained using DFT are Kohn−Sham (KS) orbitals which are not identical to the molecular orbitals of HF theory. Nevertheless, the shape and

symmetry of KS and HF orbitals are remarkably similar. It was reported by Stowasser and Hoffmann[17] that there is commonly a small systematic difference between KS and HF orbital energies which can be allowed for by linear scaling. This difference may be related to the small systematic difference between the values of $-E$ and $m_2/2$. However, as was remarked earlier, it is our considered opinion that this difference is almost certainly due to a lack of precision in evaluating the integral in eq 16.

Additional support for the accuracy of DFT generated k-space functions is given by a comparison between $D_2(k)$ obtained using both HF and DFT methods. For a selection of test molecules (water, benzene, carbon dioxide, hydrogenoxyfluoride, and methane) it was observed that the transforms were almost identical for the two methods so long as the same basis set was used. It has been suggested that DFT may handle lone pair electrons less adequately than other valence electrons.[18] Our observations have been that the existence of lone pairs did not appear to profoundly influence the generally good agreement between the $D_2(k)$ functions obtained using the two methods (HF and DFT). Furthermore, it is not noticeable that the "outlier" points in the plots described below are predominantly associated with molecules containing lone pairs.

Finally we note that the Dyson orbitals[19] generated experimentally by electron momentum (e,2e) spectroscopy[20] are generally in excellent agreement with the Dyson orbitals calculated from DFT generated k-space wave functions (see for example ref 21).

Ultimately, of course, the predictive value of the defined model must be the determining factor. If systematic errors are present, it is likely in any case that they will be largely accommodated by the regression analysis (discussed below) and will not therefore detract to any great extent from the reliability of the predictions. We see it as important for this type of study that the same theoretical model be used for all molecules in a given analysis. While it would no doubt be of interest to compare models based upon different functionals and different basis sets, we have (for reasons of economy) chosen to use the same "middle-of-the-road" theoretical model for all of the calculations reported in this article. This, we believe, is appropriate for the purpose of introducing and advocating a novel approach to obtaining QSAR/QSPR quantum-mechanical descriptors—which can and presumably will be optimized at some future point in time.

OBTAINING DESCRIPTORS FROM THE RADIAL DISTRIBUTION FUNCTION

It is reasonable to surmise that certain useful activities/properties of a molecule are likely to be primarily associated with the nature of the electron density distribution in the vicinity of the r-space perimeter of the molecule. It is in this region, remote from the nuclear centers, that the electron momentum will be least. Therefore, having regard to the reciprocal relationship between r-space and k-space, it is reasonable to suppose that the low momentum end of the $D_2(k)$ distribution will contain much of the information required for generating useful descriptors. It was found by experimentation that four descriptors were generally sufficient to characterize the low momentum end of the distributions and that the values of these descriptors for a given molecule could usefully be set equal to the coefficients obtained by curve-fitting a simple polynomial $d(\nu;k)$ to $D_2(k)$ between the origin and the peak of the distribution.

$$\mathrm{d}(\nu;k) = \sum_j \nu_j k^j \quad \text{for } j = 1, 2, 3, 4 \tag{17}$$

Multiple regression analysis was used to determine the polynomial coefficients and their associated errors. The values of the descriptors for the molecules used in this work are given in Supporting Information part A.

APPLICATIONS

QSAR and QSPR methods involve expressing a suitable function of the chosen activity/property X in terms of m predictor variables (descriptors) ν_1, ν_2, ..., ν_m and $n+1$ parameters b_0, b_1, ..., b_n. For a chosen X, the descriptors have different values for different molecules, whereas the parameters are the same for all molecules. The parameters are obtained as a best-fit to known values of X for a collection of molecules. The integrity of this procedure requires that the number of molecules used to determine the parameters be considerably larger than the number of fitted parameters. The simplest approach (used in this work) is to use multiple linear regression to determine the parameters. Thus, with four descriptors, we obtain five parameters as linear coefficients in a relation of the form:

$$\mathrm{f}(X) = b_0 + b_1\nu_1 + b_2\nu_2 + b_3\nu_3 + b_4\nu_4 \tag{18}$$

McCoy and Sykes[12] showed how this approach can be used to determine the polarizability α of a molecule. Using published experimental data[22] for a predictor set of 30 structurely diverse molecules [One molecule (CS_2) was omitted from the analysis, because it appeared to be an outlier and because there was sufficient reason, based on conflicting published data, to challenge its reported observed value as being too high.], multiple regression analysis yielded the following relation for the calculated volume polarizability:

$$\alpha/10^{-30}m^3 = -0.964 + 0.184\nu_1 + 0.206\nu_2 + 0.111\nu_3 + 0.055\nu_4 \tag{19}$$

Agreement between observed (experimental) and calculated values of α was very good, with an R^2 value (coefficient of determination) equal to 0.992. In comparison, the Gaussian 94 program tended to underestimate α and the calculated values correlated less well with the observed values ($R^2 = 0.981$).

MOLECULAR VOLUME

It is reasonable to suppose that certain activities/properties will depend to a significant extent upon molecular volume. The volume of a molecule is however difficult to specify uniquely inasmuch as the r-space perimeter of a molecule is not sharply delineated. Volume polarizability is generally regarded as a useful measure of molecular volume. Another useful volume descriptor for QSAR/QSPR studies is "McGowan's characteristic volume" (V).[23]

Using reported values of McGowan's volume[7] for a randomly chosen predictor set of 54 structurally diverse

Table 3. Summary of Results Obtained from the McGowan's Characteristic Volume Regression Model[a]

b_0	b_1	b_2	b_3	b_4
0.0363	0.0274	0.0173	0.0124	0.0099
0.0044	0.0015	0.0004	0.0007	0.0007

[a] The first row shows the parameters and the second row their associated standard errors.

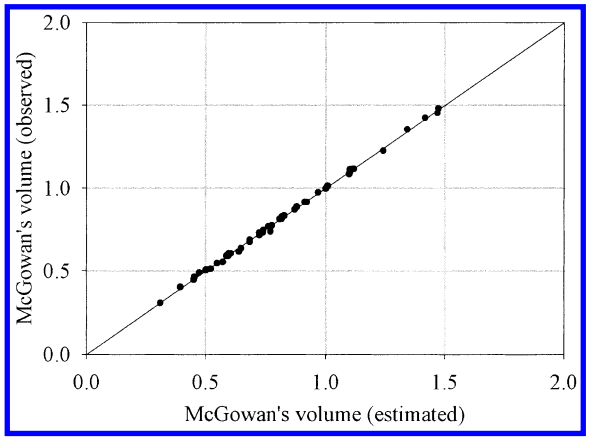

Figure 3. Plot of observed versus calculated McGowan's volume.

molecules, multiple regression analysis yielded the coefficients shown in Table 3. A scatter plot of observed versus calculated values is given in Figure 3. The agreement between observed (experimental) and calculated values of V is remarkably good ($R^2 = 0.999$).

GAS CHROMATOGRAPHY RETENTION TIMES

"Predicting the retention characteristics of a solute molecule given only its chemical structure and the experimental chromatographic conditions remains one of the grand challenges in separation science".[24] The earlier work in this area sought correlations for relatively narrow classes of compounds such as aromatic hydrocarbons,[25] substituted pyrazines,[26] stimulants and narcotics,[27] and anabolic steroids.[28] The first reported correlation that included a wide variety of organic molecules across many classes of compounds was that of Katritzky et al.[2] In the latter study, a mix of topological and quantum-mechanical descriptors was used to correlate the retention times of 152 diverse structures. The best model obtained was a six parameter linear model (with $R^2 = 0.959$). The data set of Katritzky et al. was subsequently reevaluated using a six parameter linear model to yield an R^2 value of 0.977.[29]

In the present study, a subset of 23 molecules from the Katritzky database was randomly selected for analysis. The subset included a variety of compounds (aliphatics, carbocyclics, heterocyclics, alcohols, phenols, ketones, amines, and aldehydes). Using the five parameter linear model (eq 18), good agreement was obtained between observed and calculated retention times ($R^2 = 0.957$). Noting that molecular weight M had been generally regarded by previous workers as an important descriptor for predicting retention times, it was decided to include it in the present analysis. The model thus became

$$\tau_{ret} = b_0 + b_1\nu_1 + b_2\nu_2 + b_3\nu_3 + b_4\nu_4 + b_5M \quad (20)$$

Table 4. Summary of Results Obtained from the GC Retention Time Regression Model[a]

b_0	b_1	b_2	b_3	b_4	b_5
−3.153	−0.383	0.195	0.344	0.385	0.082
0.536	0.152	0.105	0.128	0.141	0.028

[a] The first row shows the parameters and the second row their associated standard errors.

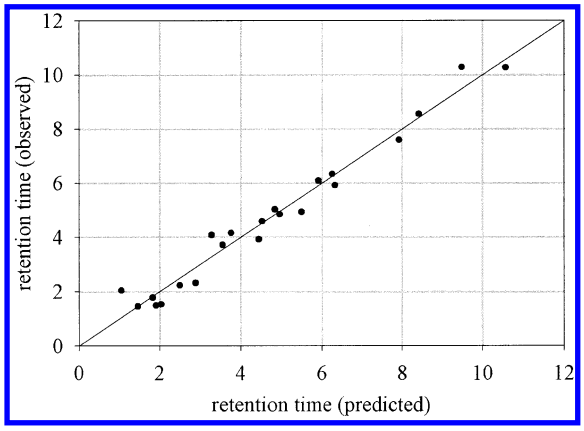

Figure 4. Plot of experimental versus calculated gas chromatographic retention time.

The inclusion of M as an additional descriptor in the linear model significantly improved the correlation ($R^2 = 0.971$). The six parameters obtained for the subset of 23 molecules are given in Table 4. While it is apparent that the precision obtained is comparable to that obtained by Katritzky et al.,[2] a more carefully considered comparison is not possible since only a subset of the 152 compounds in the original Katritzky database was used in the present analysis. A scatter plot of observed versus calculated values of the retention time is given in Figure 4.

GAS−HEXADECANE PARTITION COEFFICIENTS

The use of partition coefficients, which quantify the distribution of a solute between gas and liquid phases or between two immiscible liquids, is ubiquitous in medical and pharmacological research and has its origin in the well-known correlation between anesthetic potency and the olive oil−air partition coefficient.[30,31] Partition coefficients have since found many uses in predicting activities/properties such as proton binding, receptor affinity, and pharmacological activity.[32]

General equations have been developed to predict the effects of solutes on physicochemical and biochemical phenomena.[33] In these equations the activity/property of concern is linearly dependent upon several solute properties including the logarithm of the hexadecane−gas partition coefficient (L^{16}). In this study, a subset of 63 molecules was chosen from published data.[6,33−34] A scatter plot of observed versus calculated values of log L^{16} is shown in Figure 5. The correlation is generally good except for four molecules which are obvious outliers. The outliers (carbon tetrafluoride, sulfur hexafluoride, and the gaseous anesthetics isoflurane and methoxyflurane) each contain fluorine atoms and are indeed the only fluorine containing molecules in the data set. Furthermore, the discrepancy between observed and calculated values of log L^{16} increases with the number of

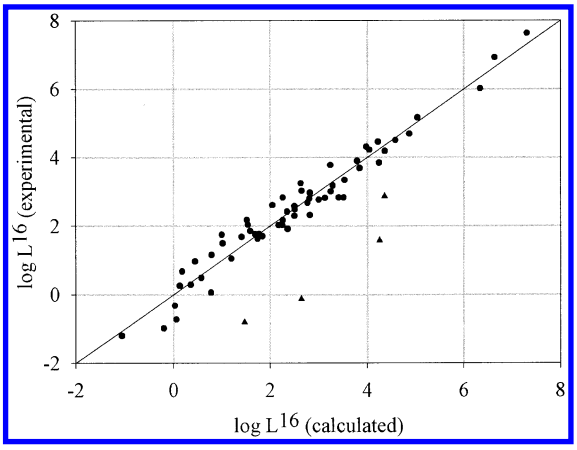

Figure 5. Plot of experimental versus calculated gas−hexadecane partition coefficient (as log L^{16}). The fluorine containing outlier molecules are represented as triangles.

Table 5. Summary of Results Obtained from the Gas−Hexadecane Regression Model[a]

b_0	b_1	b_2	b_3	b_4
−0.851	−0.181	0.130	0.145	0.136
0.127	0.052	0.017	0.026	0.032

[a] The first row shows the parameters and the second row their associated standard errors.

fluorines in the molecule. The coefficients obtained from a five-parameter linear model (eq 18) for the set of 59 non-fluorine containing molecules are given in Table 5. The analysis yielded an R^2 value of 0.954 for this data set.

In as much as fluorine atoms have had a troublesome history in ab initio work it is tempting to attribute the poor predictions to inadequacies in the wave functions. However, despite the well-documented difficulties with fluorine compounds using HF and post-HF methods, there is mounting evidence from electron momentum spectroscopy that the DFT method (used in this work) can reasonably be expected to give a consistently good account of the properties of fluorine containing molecules.[35] Furthermore the data sets used in our work to predict other molecular properties (such as polarizability, hyperpolarizability, and diamagnetic susceptibility) included fluorine-containing compounds, and no anomolous behavior of fluorine compounds as regards these other properties was apparent. We are regrettably not in a position to offer a satisfactory explanation for the anomolous behavior of fluorine compounds in respect of the gas−hexadecane partition coefficient.

HYPERPOLARIZABILITY

The dipole moment of a molecule in an electric field can be developed as a Taylor series in the applied field strength. The constant term is the permanent dipole moment. The linear term determines the mean polarizability α. The quadratic and cubic nonlinear terms determine the first and second mean hyperpolarizabilities β and γ. The ability to estimate hyperpolarizability can be expected to facilitate the development of practical devices with suitable nonlinear properties for optical harmonic generation and signal processing.[36]

Although the hyperpolarizability can be calculated for small molecules (albeit with great difficulty), it would nevertheless be desirable to have available a reliable structure−property relationship to estimate it at a lower cost. Previous attempts have been made to correlate hyperpolarizability with parameters such as HOMO−LUMO gaps and (for aromatics) resonance energies.[37] These methods have met with only limited success. A possible reason for limited success is the lack of a large and truly homogeneous data set. The data that are available have, for the most part, been determined by computation and originate in different laboratories using different ab initio methods, with different basis sets and different optimization requirements. The choice of method in particular can have a profound effect on calculated values of the hyperpolarizability.[38]

Table 6. Summary of Results Obtained from the ESHG/EFISH Second Hyperpolarizability Regression Model[a]

b_0	b_3	b_4	x
8.050	−2.140	−3.518	2.000
5.034	0.166	0.295	

[a] The first row shows the parameters and the second row their associated standard errors.

McCoy and Sykes[12] have outlined a QSPR method, similar to that described in the present article, to estimate the mean polarizability and the mean second hyperpolarizability in the static limit. For the hyperpolarizability estimates, the data set consisted of 22 diverse but relatively small molecules whose electronic second hyperpolarizabilities had been previously calculated using accurate post-Hartree−Fock methods of computation. The analysis yielded an R^2 value of 0.972 for the data set.

Experimental methods for measuring hyperpolarizabilities include static electric field induced second harmonic generation (ESHG/EFISH), dc Kerr effect, coherent anti-Stokes Raman scattering (CARS), and third harmonic generation (THG).[36] The measured hyperpolarizability of a molecule necessarily includes significant vibrational and rotational contributions in addition to the pure electronic contribution. Using a data set of ESHG/EFISH values for 25 molecules, the following model was found to generate the most reliable estimate of the experimental second hyperpolarizability.

$$X(\boldsymbol{b}; \nu) = (b_0 + b_3\nu_3 + b_4\nu_4)^x \qquad (21)$$

The values of the parameters are given in Table 6, and a scatter plot of given versus predicted values of the hyperpolarizability is shown in Figure 6. In determining this relationship, it became apparent that the descriptor variables ν_1 and ν_2 have very little predictive value. The additional parameter x was included to allow for nonlinearity in the relationship between the descriptor variables and the hyperpolarizability. The analysis yielded an R^2 value of 0.921 which suggests that the method should only be used to obtain a crude estimate of the experimental (ESHG/EFISH) second hyperpolarizability. The method is evidently more reliable for estimating the pure electronic contribution to hyperpolarizability.

TADPOLE NARCOSIS AND ANESTHESIA

The theory of general anesthesia is sketchy. It has its origins in the early work of Meyer[30] and Overton[31] who studied the narcotic activity of aqueous solutes toward the

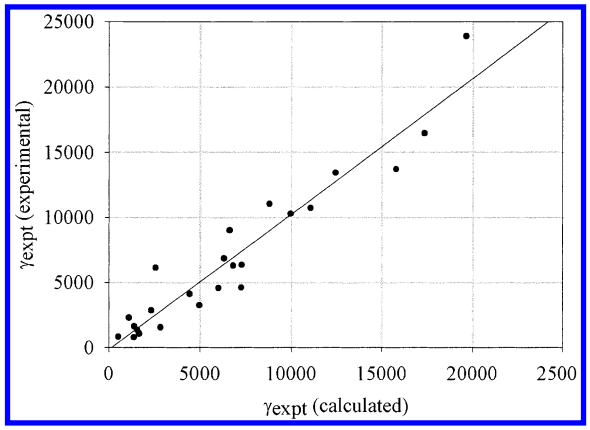

Figure 6. Plot of experimental EFISH/ESHG against calculated second hyperpolarizability. Atomic units are used throughout.

tadpole. Many anesthetic agents obey the well-known Meyer−Overton rule which, in its original form, states that the narcotic concentration C_{nar} (the minimum concentration required for narcosis) is inversely proportional to the olive oil−water partition coefficient P_{ow}. This rule has subsequently been extended to other biological activities and has spawned the so named "lipoid" theories of anesthesia. The focus of contemporary theories of anesthesia however is to implicate a specific protein receptor rather than cell membrane lipids.[39] It has for example been suggested[7] that the anesthetic binding site on the primary protein target is a "large or flexible pocket, extending inward from the external water-facing surface, that is rather aqueous and of limited hydrophobicity".

Although a broad range of compounds is known to comply with the Meyer−Overton law, a number of compounds show an activity that is considerably different to that predicted. Indeed, some halogenated hydrocarbons that are expected to be potent anesthetics exhibit no anesthetic activity at all.[40] These inconsistencies in the Meyer−Overton predictions have led to the investigation of other descriptors by various workers. Abraham and Rafols[7] studied tadpole narcosis using a database containing 89 compounds and a linear model based on five physicochemical descriptors to predict log($1/C_{nar}$), the narcotic activity. The regression analysis yielded an R^2 value of 0.948 (with five of the 89 compounds, judged to be outliers, omitted from the analysis). By comparison a regression analysis of the linear relationship between the narcotic activity and the logarithm of the octanol−water partition coefficient yielded an R^2 value of 0.849.

In the present work, a structurally diverse subset of 54 compounds drawn from the compounds studied by Abraham and Rafols was used to determine the parameters in the linear model

$$\log(1/C_{nar}) = b_0 + b_1\nu_1 + b_2\nu_2 + b_3\nu_3 + b_4\nu_4 + b_5M \quad (22)$$

The inclusion of the molecular weight M as a descriptor led to a significant improvement in the goodness of fit. Regression analysis yielded an R^2 value of 0.931 (with two molecules judged to be outliers removed from the analysis). With M omitted from the model an R^2 value of 0.897 is obtained.

It is not possible to comment assertively on the relative merits of the present method and that used by Abraham and Rafols since the data sets are different in size. It would appear however that, if such a comparison could be made, it would most likely be favorable. The values of the parameters obtained from the present analysis are given in Table 7, and a scatter plot of observed and versus predicted values of log($1/C_{nar}$) is shown in Figure 7. The units for C_{nar} are mol dm^{-3}.

Table 7. Summary of Results Obtained from the Tadpole Narcosis Regression Model[a]

b_0	b_1	b_2	b_3	b_4	b_5
−1.192	−0.194	−0.078	−0.120	−0.189	0.023
0.164	0.060	0.023	0.034	0.034	0.005

[a] The first row shows the parameters and the second row their associated standard errors.

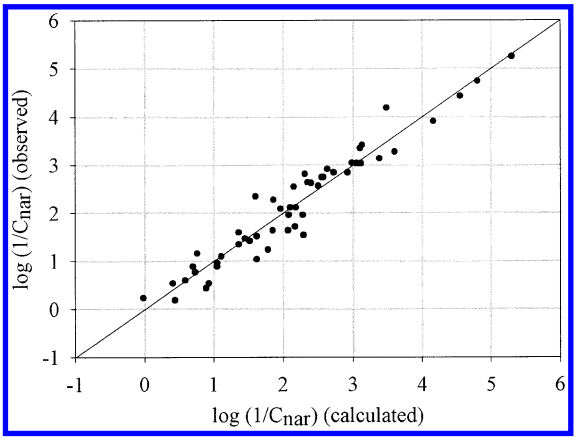

Figure 7. Plot of experimental versus calculated tadpole narcotic concentrations (as log($1/C_{nar}$)).

CONCLUSION

In this article we have described a new method for predicting structure-related properties and activities of chemical compounds. The method compares favorably with conventional approaches using experimentally determined predictor variables. The method has one main advantage over conventional methods however. Once the model has been refined and its parameters determined for a large and sufficiently diverse database of compounds, the predicted properties of a new compound can be obtained more expeditiously and at a lower cost than is possible using conventional methods which rely upon laboratory measurements or disagreeable experiments on sentient organisms.

In the examples given, subsets of compounds selected from larger databases were used. Regrettably, this was necessary in order to limit costs for what is essentially a pilot study for a new approach to QSAR/QSPR research. The subsets included all compounds that were common to our own database and to the much larger published databases with which we wished to make comparisons. We consider that each of the subsets used comprises a more or less random collection of compounds, but we have no way of knowing for certain whether our sampling introduced an unintentional bias. Hence we are reluctant to make claims that cannot be substantiated by formal statistical analyses. We do however believe that it has been clearly demonstrated that the proposed new method has considerable merit. What would be useful in a future study would be to compare the approach advocated by this article with other published analyses using a standard QSAR training set of molecules. Such investiga-

tions would ultimately judge the wider applicability of the model and determine applications where the momentum space approach is not profitable.

In the future, it is clear that the method can be further refined. The use of different computational methods, different model chemistries, and different basis sets may offer advantages. Different ways of parametrizing the momentum space wave function can be trialled. Larger databases should be developed to enable more discerning comparisons with conventional methods. The concurrent use of momentum space quantum-mechanical descriptors and other conventional descriptors is likely to be fruitful in QSAR/QSPR investigations.

ACKNOWLEDGMENT

We acknowledge, with gratitude, financial support from Flinders University, an Australian Postgraduate Award granted to M.J.S. and the South Australian Centre for Parallel Computing for the use of their computing facilities.

Supporting Information Available: Table of quantum-mechanical descriptors obtained from the radial distribution function in momentum space (part A) and subsets of molecules used in this study to determine QSAR/QSPR parameters (part B). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hansch, C.; Fujita, T. ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.
(2) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. Prediction of gas chromatographic retention times and response factors using a general quantitative structure−property relationship treatment. *Anal. Chem.* **1994**, *66*, 1799−1807.
(3) Abraham, M. H.; McGowan, J. C. The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography. *Chromatographia* **1987**, *23*, 243−246.
(4) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(5) Cronce, D. T.; Famini, G. R.; De Soto, J. A.; Wilson, L. Y. Using theoretical descriptors in quantitative structure−property relationships: some distribution equilibria. *J. Chem. Soc., Perkin Trans. 2* **1998**, 1293−1301.
(6) Abraham, M. H.; Grellier, P. L.; McGill, R. A. Determination of olive oil-gas and hexadecane-gas partition coefficients, and calculation of the corresponding olive-oil water and hexadecane-water partition coefficients. *J. Chem. Soc., Perkin Trans. 2* **1987**, 797−803.
(7) Abraham, M. H.; Rafols, C. Factors that influence tadpole narcosis. An LFER analysis. *J. Chem. Soc., Perkin Trans. 2* **1995**, 1843−1851.
(8) Gratton, J. A.; Abraham, M. H.; Bradbury, M. W.; Chada, H. S. Molecular factors influencing drug transfer across the blood-brain barrier. *J. Pharm. Pharmacol.* **1997**, *49*, 1211−1216.
(9) Cooper, D. L.; Mort, K. A.; Allan, N. L.; Kinchington, D.; McGuigan, C. Molecular similarity of anti-HIV phospholipids. *J. Am. Chem. Soc.* **1993**, *115*, 12615−12616.
(10) Cooper, D. L.; Allan, N. L. *Molecular similarity and reactivity: from quantum chemical to phenomenological approaches*; Kluwer: 1995; pp 31−55.
(11) Measures, P. T.; Mort, K. A.; Allan, N. L.; Cooper, D. L. Applications of momentum space similarity. *J. Comput.-Aided. Mol. Des.* **1995**, *9*, 331−340.
(12) McCoy, E. F.; Sykes, M. J. The estimation of molecular properties using momentum-space wave functions. *Chem. Phys. Lett.* **1999**, *313*, 707−712.
(13) Frisch, M. J. et al. *Gaussian 94* (*Revision D.4*); Gaussian Inc.: Pittsburgh, 1995.
(14) Champeney, D. C. *Fourier transforms and their physical applications*; Academic Press: London, 1973.
(15) Press: W. H,; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical recipes in FORTRAN: the art of scientific computing*, 2nd ed.; Cambridge University Press: Cambridge, 1992.
(16) Hehre, W. J.; Radom, L.; Schleyer, P.v. R.; Pople, J. A. *Ab Initio Molecular Orbital Theory*; Wiley: New York, 1986.
(17) Stowasser, R.; Hoffmann, R. What do the Kohn−Sham orbitals and eigenvalues mean? *J. Am. Chem. Soc.* **1999**, *121*, 3414−3420.
(18) Cooper, D. L. personal communication, 2000.
(19) Duffy, P.; Chong, D. P.; Casida, M. E.; Salahub, D. R. Assessment of Kohn−Sham density-functional orbitals as approximate Dyson orbitals for the calculation of electron-momentum-spectroscopy scattering cross sections. *Phys. Rev. A* **1994**, *50*, 4707−4728.
(20) McCarthy, I. E.; Weigold, E. (e,2e) spectroscopy. *Phys. Rep.* **1976**, *27C*, 275−371.
(21) Adcock, W.; Brunger, M. J.; Clark, C. I.; McCarthy, I. E.; Michalewicz, M. T.; von Niessen, W.; Weigold, E.; Winkler, D. A. Theoretical and experimental investigation into the complete valence electronic structure of [1.1.1] propellane. *J. Am. Chem. Soc.* **1997**, *119*, 2896−2904.
(22) Rhee, C. H.; Metzger, R. M.; Wiygul, F. M. Atom-in-molecule polarizabilities. *J. Chem. Phys.* **1982**, *77*, 899−915.
(23) Mellors, A.; McGowan, J. C. Uses of molecular volume in biochemical pharmacology. *Biochem. Pharmac.* **1985**, *34*, 2413−2416.
(24) Martin, M. G.; Siepmann, J. I.; Schure, M. R. Simulating retention in gas−liquid chromatography. *J. Phys. Chem. B* **1999**, *103*, 11191−11195.
(25) Whalen-Pedersen, E. K.; Jurs, P. C. Calculation of linear temperature programmed capillary gas chromatographic retention indexes of polycyclic aromatic compounds. *Anal. Chem.* **1981**, *53*, 2184−2187.
(26) Stanton, D. T.; Jurs P. C.; Computer-assisted prediction of gas chromatographic retention indexes of pyrazines. *Anal. Chem.* **1989**, *61*, 1328−1332.
(27) Georgakopoulos, C. G.; Kiburis, J. C.; Jurs, P. C. Prediction of gas chromatographic relative retention times of stimulants and narcotics. *Anal. Chem.* **1991**, *63*, 2021−2024.
(28) Georgakopoulos, C. G.; Tsika, O. G.; Kiburis, J. C.; Jurs, P. C. Prediction of gas chromatographic relative retention times of anabolic steroids. *Anal. Chem.* **1991**, *63*, 2025−2028.
(29) Lucic, B.; Trinajstic, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A new efficient approach for variable selection based on multiregression: Prediction of gas chromatographic retention times and response factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610−621.
(30) Meyer, H. Zur theorie der Alkoholnarkose, welche Eigenscaft der anaesthetica bedingt ihre narkotische Wirkung. *Arch. Exp. Pathol. Pharmakol.* **1899**, *42*, 109−118.
(31) Overton, E. *Studien uber die Narkose Zugleich ein Beitrag zur Allgemeine Pharmakologie.* Jens, Verlag von Gustav Fischer: Germany, 1901.
(32) Bodor, N.; Gabanyi, Z.; Wong, C,-K. A new method for the estimation of partition coefficients. *J. Am. Chem. Soc.* **1989**, *111*, 3783−3786.
(33) Abraham, M. H.; Fuchs, R. Correlation and prediction of gas−liquid partition coefficients in hexadecane and olive oil. *J. Chem. Soc., Perkin Trans. 2* **1988**, 523−527.
(34) Abraham, M. H.; Whiting, G. S.; Fuchs, R.; Chambers, E. J. Thermodynamics of solute transfer from water to hexadecane. *J. Chem. Soc., Perkin Trans. 2* **1990**, 291−300.
(35) Zheng, Y.; Brion, C. E.; Brunger, M. J.; Zhao, K.; Grisogono, A. M.; Braidwood, S.; Weigold, E.; Chakravorty, S. J.; Davidson, E. R.; Sgamellotti, A.; Vonniessen, W. Orbital momentum profiles and binding energy spectra for the complete valence shell of molecular fluorine. *Chem. Phys.* **1996**, *212*, 269−300.
(36) Shelton, D. P.; Rice, J. E. Measurements and calculations of the hyperpolarizabilities of atoms and small molecules in the gas phase. *Chem. Rev.* **1994**, *94*, 3−29.
(37) Lu, Y.-J.; Lee, S.-L. Semiempirical calculations of the nonlinear optical properties of polycyclic aromatic compounds. *Chem. Phys.* **1994**, *179*, 431−444.
(38) Maroulis, G. A systematic study of basis set, electron correlation, and geometry effects on the electric multipole moments, polarizability and hyperpolarizability of HCl. *J. Chem. Phys.* **1998**, *108*, 5432−5448.
(39) Franks, N. P.; Lieb, W. R. Molecular and cellular mechanisms of general anesthesia. *Nature* **1994**, *367*, 607−614.
(40) Chipot, C.; Wilson, M. A.; Pohorille, A. Interactions of anesthetics with the water-hexane interface. A molecular dynamics study. *J. Phys. Chem. B* **1997**, *101*, 782−791.

CI025597B