QSPR Analysis of HPLC Column Capacity Factors for a Set of High-Energy Materials Using Electronic van der Waals Surface Property Descriptors Computed by Transferable Atom Equivalent Method

CURT M. BRENEMAN* and MARLON RHEM

Department of Chemistry, Cogswell Labs, Rensselaer Polytechnic Institute, Troy, New York 12180

Received 20 November 1995; accepted 23 April 1996

ABSTRACT

The new transferable atom equivalent (TAE) method for rapid molecular electron density reconstruction was used to compute a set of molecular surface property descriptors. These descriptors were then used to construct HPLC column capacity factor PLS models for a series of high-energy materials. The new TAE-derived surface property indices are also available from *ab initio* or semiempirical wave functions, but the speed and accuracy of TAE reconstruction make it the method of choice for obtaining these indices. The new QSPR indices are based upon the extrema, distributions, and surface integrals of the electronic kinetic energy density, the Politzer average local ionization potential (pip), and the electrostatic potential, as well as the rates at which these properties change normal to the 0.002-e/au^3 molecular surface. The distribution of the properties were recorded as surface histograms. While property extrema and surface integral averages proved to be descriptive, the most useful new indices were found to correspond to histogram bin data computed for K and G surface kinetic energy densities.

This article was submitted on the occasion of Professor Richard Bader's 65th birthday.

This article includes Supplementary Material available from the authors upon request or via the Internet at ftp.wiley.com/public/journals/jcc/suppmat/18/182 or http://www.wiley.com/jcc

* Author to whom all correspondence should be addressed. E-mail: breneman@xray.chem.rpi.edu

Journal of Computational Chemistry, Vol. 18, No. 2, 182–197 (1997)
© 1997 by John Wiley & Sons
CCC 0192-8651 / 97 / 020182-16

All-subsets-regression modeling showed that when mixtures of traditional connectivity indices, theoretical linear solvation energy relations (TLSERs), and generalized interaction properties functions (GIPFs) were included with the new indices in the variable sets, the new indices were consistently involved in the best 2% of the capacity factor models. Peak retention time data from two different columns were examined (a Hypersil ''CPS'' cyanoalkylated column and a standard reverse-phase Hypersil ''ODS'' column) using a phosphate buffer mobile phase at pH 3.0. The result were compared to an earlier TLSER correlation analysis of the same data by Lowrey and Famini. The TAE-generated surface property descriptors were shown to provide superior PLS models for both sets of columns and conditions. © 1997 by John Wiley & Sons, Inc.

Introduction

Although a diverse set of molecular descriptors may be found in the QSAR and QSPR literature[1], new descriptors which correlate well with important intermolecular interaction modes are worthy of investigation.[2] While descriptors which are directly related to well-known types of intermolecular interactions are more intuitive, property indices which correlate with less clearly defined nonbonded interaction modes can often lead to models with excellent predictive power. Separate descriptions of electrostatic interactions, van der Waals forces, charge transfer interactions, polarizability, and hydrophobic effects are desirable, but such clear dimensions are not often possible given the kind of molecular information available in a typical QSAR or QSPR study. One approach to this problem has been to use property indices computed from atomic connectivity patterns of molecules, or from spectral[3] and experimental[4] data taken from a training set of molecules, but clean separation of each type of interaction into a set of additive linear models can be elusive in the real world. A well-utilized alternative approach uses molecular connectivity information to construct topological property indices and substructure keys. Connectivity-based two-dimensional (2D) approaches such as these have been heavily used in ''traditional'' QSAR.[5] Hansch and others were able to demonstrate that the use of such indirect property indices was a valid, if not perfect approach to structure/activity analysis.[6] Implicit in such 2D QSAR work is the assumption that characteristic connectivity patterns of atoms result in specific chemical behavior. Some of the early pattern descriptors are simple sums of hydrogen bond donor or acceptor atoms, or the number of rings present in a particular compound. The donor/acceptor atom count is an example of an index type which can be directly related to a specific kind of intermolecular interaction. This approach works to some extent because each substructure is presumed to add an energy increment to the binding. Verloop indices and Balaban indices[7] represent extensions of connectivity indices on which some kind of shape information is encoded. The middle ground between direct and indirect relationships with known modes of interaction is occupied by indices derived from shadow areas or moments of inertia about principle molecular axes. Aside from a tally of each kind of substructure, none of these descriptors include any magnitude or directionality information concerning the implied chemical effects of the substructure on binding.[8] A significant advance in the graph theoretical approach to QSAR was made with the introduction of electrotopological indices, which can be used to designate ''E-states'' of molecules and were designed to describe electronic alterations which accompany changes in connectivity.[9] Keir and Hall pioneered this area of research,[10] and contributed to a significant improvement in 2D descriptors. Keir also described some early efforts to use semiempirical information in 2D QSAR descriptor generation.[11]

Many modern efforts have concentrated on three-dimensional (3D) techniques[12] such as the CoMFA-style field comparison approaches. These field-based methods concentrate on both steric and electrostatic similarities between molecules. The power of 3D QSAR rests in its ability to identify regions of space which are sensitive to changes in shape or electrostatic fields. As with all such approaches, the 3D methods suffer from molecular alignment problems. The hybrid molecular surface methods used in the present work seek to identify intrinsic electronic properties of molecules by evaluating these properties on and near the molecular van der Waals surfaces. This approach combines some of the benefits of 2D methods, in that molecular alignment is not important, while still utiliz-

ing high-quality electron-density-based values of surface properties as molecular descriptors. The concept of using surface property distributions as indicators of molecular behavior is not a new one. Politzer and coworkers have explored the relationships between the distribution of molecular surface properties and a diverse set of spectral[13] and reactivity data.[14] In each of these cases, however, the electrostatic potential was the primary property to be statistically analyzed. Because the electrostatic potential field surrounding a molecule is a good indicator of long range interactions as well as a descriptor of the electron density distribution, this approach has proven to be quite useful. A number of years ago, Hehre pointed out that the pK_a values of carboxylic acids correlated strongly with the most positive electrostatic potential found on their electronically defined van der Waals surfaces.[15] Politzer has taken this approach further, and defined a set of surface property indices known as generalized interaction properties function (or GIPFs) which contain information about the distributions of the STO-nG-generated positive and negative electrostatic potential distributions on molecular surfaces. The GIPF descriptors also include a volume-normalized polarizability term and the extrema of the local average ionization potential of the surface electron density.[16] Politzer's group has successfully used these descriptors to generate correlations of structure with supercritical fluid solubility and boiling point, as well as other physical properties related to intermolecular interactions.[16] Other significant work in the area of theoretically generated descriptors has been pioneered by Famini and coworkers. Their efforts resulted in a set of theoretical linear solvation energy relationships (TLSERs) and associated descriptors which are also based upon MOPAC results.[17] As with Politzer's GIPFs, these descriptors were designed to represent various specific types of intermolecular interactions. TLSERs differ from GIPFs in that they are not determined on the basis of surface electrostatic potentials, but they do contain two ''ionic hydrogen bond'' charge terms to account for the electrostatic components of bonding as well as separate covalent donor and acceptor terms. Among many other successes, TLSERs have been used to create solvation models[18] as well as to predict HPLC column capacity factors.[19]

Theoretically defined QSAR/QSPR descriptors such as those discussed above have a number of advantages over empirically derived indices: They are related to the actual molecular electron density, they are applicable to a wide variety of molecules, and are fairly accessible through semiempirical or small *ab initio* calculations.[20] For those reasons, TLSERs and GIPFs are also favored over many graph theoretical descriptors and substructure keys. One logical improvement for such theoretical descriptors would be the inclusion of electronic properties other than the electrostatic potential and local average ionization potential. Electronic kinetic energy density is one such type of potential descriptor. There are reasons to believe that some of these properties might be available from MOPAC or other semiempirical methods, but since these methods are not calibrated for such output, the usefulness of that approach would be questionable. To obtain accurate molecular electron density distributions which include good values of first and second derivatives, high-quality *ab initio* calculations are required. For even the most modestly sized molecules, however, the time and computer resources required to complete a large number of *ab initio* calculations is prohibitive. Although computational facilities have become much faster, most QSAR/QSPR studies involve dozens or hundreds of sample molecules, rendering large *ab initio* studies impractical. Nevertheless, quantum mechanical molecular descriptors remain attractive.[21] To facilitate the use of density-based descriptors, an alternative approach is required which can generate property-encoded high-quality electron density reconstructions of any molecule.[22] Our approach to this problem has been to use assemblies of transferable atom equivalents (TAEs) as representations of molecular electron densities.[23] Through the use of these TAE reconstructions, we have been able to derive new sets of atomic and molecular descriptors which are based upon properties of the *ab initio*-quality molecular electron density. In addition to some of the new descriptors mentioned in this work, TAE assemblies are also capable of providing all of the traditional 2D and 3D descriptors as well as HF/6-31+G* *ab initio*-equivalent GIPF and TLSER indices.

Transferable Atom Equivalent Method

Transferable atom equivalents (TAEs) are essentially atom-centered electron density fragments with discrete boundaries, each containing a charged nucleus. Additionally, TAEs are equipped to change their shape and their properties slightly in response to new molecular environments. In the terminology of Bader's Theory of Atoms in

Molecules,[24] atomic nuclei are *attractors* which are surrounded by *basins* of electron density bounded by zero-flux interatomic surfaces ($\nabla\rho \cdot N = 0$). Within this approach, the boundaries between one atom and the rest of the molecule may be determined by first locating all of the *bond critical points* which involve the atom of interest, and then developing a set of steepest-descent paths in electron density radiating outward from those points. Bond critical points are defined as those points along bond paths where the electron density reaches its minimum. *Bond paths* are the paths of maximum electron density connecting two atoms. The points found along the steepest descent pathways from each bond critical point define the *zero flux surface* separating two atoms. In the PROAIM and FASTINT programs, positions of these surfaces are stored as an ordered set of distances from the nuclei, with the spherical polar abscissas chosen on the basis of the Gauss quadrature formula. For mathematical reasons, this set of uneven abscissas and compensating weights are very desirable when undertaking 3D integrations of the property densities contained within the basin. In this manner, the Atoms-in-Molecules approach has been used successfully by many groups to explain a diverse set of molecular phenomena.[25]

The evolution of electron density partitioning methodology continued over the past few years during which time Bader pointed out that it was not only very desirable but also theoretically possible to recombine subsets of AIM electron density distributions to produce whole molecules.[26] The transferable atom equivalent technology and the RECON program are therefore a logical and philosophical extension of the Theory of Atoms in Molecules.

To accomplish actual molecular electron density reconstruction, two conditions must be met: (1) there must be a supply of appropriate atom types in a form which can carry electron density properties; and (2) the atomic fragments must be able to make small, self-consistent adjustments to the properties of their base atom types to fit new bonding environments. The goal set forth in point (1) has also been stated by Pichon-Pesme and Lecomte,[27] Bader,[28] and Walker and Mezey,[29] while the even more crucial goal outlined in point (2) has been elusive until now. This is where the greatest point of difference exists between the TAE approach and the methods developed by the aforementioned investigators. Mezey's MEDLA program generates molecular electron densities as mathematical superpositions of local density matrices and is designed to support shape similarity calculations of large molecules. With the exception of an approximation of the molecular electrostatic potential, no other atomic or molecular properties of the density are currently available from that method.[30] Electron densities produced by the MEDLA method are not meant to be queried for energetics or be cognizant of any local polarization effects. Similarly, Pechon-Pesme and Lecomte use sets of atom-centered multipolar functions to reproduce experimental X-ray electron density distributions. While this is not uncommon among the high-resolution X-ray community, Lecomte has also stated that it should be possible to create and use a library of generally transferable atomic multipole models to construct the electron density of any molecule. This type of reconstruction yields results broadly similar to those from Mezey's MEDLA approach, in that there is no attempt to reproduce density-based molecular properties in addition to the density. In contrast to these methods, TAE electron density reconstructions provide not only molecular electron densities, but also electronic kinetic energy information, local ionization potential (pip) distributions,[31] and electrostatic potential data as well as other first- and second-derivative properties of the final, polarized molecular density. The TAE method also keeps track of the additive atomic contributions to all molecular properties. Consequently, the TAE approach is an attractive alternative to other methods of fast molecular electron density reconstruction for any application which would benefit from more information than $\rho(r)$.

As mentioned inpoint (1), an appropriate TAE library must be maintained which contains enough atom types to be representative of many bonding environments. Second, appropriate atom types must be selectable from the library using only molecular connectivity and conformation information. On the basis of these guidelines, a TAE data structure has been defined which makes use of the relationship between atomic shape, atomic properties, and molecular connectivity. Within this definition, the data for each TAE atom type consists of a spherical polar coordinate model of its shape, surface properties, and property derivatives in a predefined ''standard'' orientation. The orientation chosen for the current TAE modeling paradigm is one in which the shortest distance from the nucleus to an interatomic surface is aligned with the $+Z$ axis, with the next shortest distance placed in the XZ plane. This procedure allows each position on the external and interatomic surfaces of an

atom to be uniquely defined.[32] This data structure format also includes a set of property derivatives for each surface point which describes the response of all atomic properties (such as interatomic surface electron densities and total atomic energies) to radial variations of each unique surface position. These derivatives are initially determined analytically from the wave functions of the sample molecules used to produce the TAEs. Together with the RECON pairwise bonding algorithm, this results in a flexible atomic density representation which is capable of slightly altering its properties in order to fit a new environment.

TAE Library

To appreciate our choice of TAE library data structure, one must first become familiar with the kind of information which is available from a FASTINT or PROAIM electron density partitioning calculations. After the molecular connectivity has been determined, the bond critical points are located and the interatomic surfaces are generated. The limits of each atomic basin are determined as an order list of distances from each atomic nucleus. This list contains distance information at each point in a spherical polar coordinate system using Gauss quadrature abscissas in both theta and phi. In a typical FASTINT integration, 64th order quadrature is used in the theta (longitudinal) dimension, and 48th order quadrature is used in the phi (latitude) dimension. This means that the contents of the ordered data list contains 3072 distances beginning near the $+Z$ pole in the XZ plane, proceeding in 48 disks of common phi until almost reaching the $-Z$ pole. IN the FASTINT and PROAIM programs, multiple intersection distances of rays with folded interatomic surfaces are also stored, but this information is not presently utilized in the TAE library. During numerical integration, electron density is sampled at 64-Gauss quadrature abscissas along the length of each ray from its nuclear origin to its endpoint on an interatomic surface. If no interatomic surfaces are encountered in a given direction, integration is normally continued out to 10 au from each nucleus. If the rays are truncated when the electron density has fallen to the 0.002-e/au^3 level, their endpoints define the molecular VDW surface.[33] Within our electronic modeling paradigm, atomic surface properties are determined on this 0.002-e/au^3 isosurface.

The gradient vector field of the electron density of any molecule consists of a set of noncrossing gradient paths which lead away from each nucleus and go asymptotic with each of the interatomic surfaces. In fact, if a gradient pathway is initiated at a bond critical point, it will follow a steepest descent pathway from that critical point and generate part of a zero-flux interatomic surface. Within this paradigm, a hypothetical terminal hydrogen atom basic can be used to exemplify how the spherical polar coordinate system of each AIM atom can be used to define the shape of a TAE. For each Gauss quadrature abscissa, three types of rays must be considered: rays which hit the zero-flux interatomic surface of the atom in question; rays which do not hit any interatomic surfaces but which are truncated at the 0.002-e/au^3 electron density level; and rays which continue on until a distance of 10 au is reached. The first set of rays is the one which describes the interatomic surface shape, whereas the second set of truncated rays serves to describe the electronic van der Waals surface. The third set of rays is used to determine if any new zero-flux surfaces have come into being as a result of the proximity of other atoms.

Examination of the shapes and integrated atomic properties of over 7000 atoms of all types led to the hypothesis that there is a fundamental relationship between interatomic surface shape, atomic volume, and bonding environment. To further examine this postulate, a surface matching program was written to quantitatively compare atomic shapes and molecular positions for several hundred atoms in molecules.[34] The encouraging results from that investigation reinforced our belief that chemically similar atoms ought to have similar shapes and electronic properties—including surface properties.[35] While this does not preclude atoms of different shapes having similar properties, it simply reinforces the chemical intuition that atoms with similar connectivity patterns will behave similarly and will have similar shapes—at least in regions of high electron density near the bonding axes.

TAE Definition and Identification

To determine whether the integrated properties of atoms could be divided into a reasonable number of statistically valid groups, a set of 7250 property-encoded integrated atoms were subjected to cluster analysis methods. Cluster analysis attempts to define ''natural'' groupings of objects by

measuring the similarity between them. The algorithms used in defining TAEs were taken from the agglomerative hierarchical clustering methods available in the Statistical Analysis Software (SAS) package from SAS Institute Inc. Within this method, an initial set of single element cluster groups is defined and the algorithm works toward reducing the number of groups by combining those which have the smallest Euclidean distance between them. To assist the assignment of atoms to elongated or odd-shaped clusters, cluster centers were assigned to unique atomic connectivity patterns. In this manner, members of irregular clusters could be unambiguously assigned to specific TAE atom types which all have the same nearest and next-to-nearest neighbors. It should be noted that the TAE library is a dynamic entity which is subject to continuous upgrades.

The variables used in the TAE cluster analysis were taken from integrated atomic properties as well as scalar and vector van der Waals surface properties and integrated 3D basin properties. The integrated atomic basin properties included electron population, atomic dipole magnitude, total electronic kinetic energy, volume, and the three principal quadrupole components. Surface properties included external surface area, valence region surface area, valence critical point electron denNsities. The following integrated van der Waals surface properties were also used: $\sum(\nabla\rho \cdot N)_i\Delta_i$, $\sum(\nabla K \cdot N)_i\Delta_i$, $\sum(\nabla G \cdot N)_i\Delta_i$, $\sum(K)_i\Delta_i$, and $\sum(G)_i\Delta_i$, where Δ_i is the ith surface element and N is the molecular surface normal vector. These terms are further defined in Table I.

Thus, for an sp^3 carbon atom with four bonds, there exist 18 clustering variables for each element. The variables were standardized prior to cluster analysis, and the 7250 atoms were subdivided into 12 subsets: hydrogens, sp^3 carbons, sp^2 carbons, sp carbons, sp^3 nitrogens, sp^2 nitrogens, sp nitrogens, sp^3 oxygens, sp^2 oxygens, sulfurs, fluorines, and chlorines. Several clustering methods were used on each data set to ascertain the most useful method for the given data. The reliability of each method is based on the connectivity pattern of the clusters produced and the standard deviations from the mean for each property used in clustering.

Within the limitations of our initial atomic property database, we have identified 90 unique TAEs for carbon, hydrogen, nitrogen, oxygen, silicon, sulfur, chlorine, and fluorine atom types. A table of these types and their associated connectivities is available as part of the Supplementary Material.[36] TAEs covering phosphorus and hypervalent sulfur are under development, as are special nitrogen atom types designed for use in photographic dyes and dye couplers.

RECON Algorithm

The RECON program module is employed when TAE atoms are to be combined to generate prop-

TABLE I.
Key to Atomic Properties Used in Current TAE Definition and Reconstruction.

W, X, Y, Z	valence critical point ρ values in e/au^3
A	valence region surface area
B	external surface area (au^2)
N	electron population (electrons)
E	electronic energy (Hartrees)
EP	electrostatic potential surface integral (au)
D	dipole magnitude (au · e)
V	atomic volume within the 0.002-e/au^3 isosurface (au^3)
QAA, QBB, QCC	diagonalized quadrupole components (au^2 · e)
$\nabla\rho \cdot N$	$\sum(\nabla\rho \cdot N)_i\Delta_i$, The rate of change of electron density perpendicular to the atomic surface
$\nabla K \cdot N$	$\sum(\nabla G \cdot N)_i\Delta_i$, the rate of change of K kinetic energy density normal to the atomic surface
$\nabla G \cdot N$	$\sum(\nabla G \cdot N)_i\Delta_i$, the rate of change of G kinetic energy density normal to the atomic surface
K	electronic kinetic energy density $= -(\hbar/4m)N\int d\tau'\{\Psi'\nabla^2\Psi + \Psi\nabla^2\Psi'\}$
G	electronic kinetic energy density $= -(\hbar/2m)N\int d\tau'\,\nabla\Psi'\cdot\nabla\Psi$
K	$\sum(K)_i\Delta_i$, the surface integral of the kinetic energy density derived from K (au · Bohr)[a]
G	$\sum(G)_i\Delta_i$, the surface integral of the kinetic energy density derived from G (au · Bohr)[a]
Δ_i	surface area element

[a] Kinetic energy densities K and G are related by $K = G + L$, where L is the Laplacian of charge density. The integral of L vanishes for topologically defined atoms.

erty-encoded molecular electron density distributions. Input to the RECON program currently requires either a Gaussian-style .COM file or a standard MOPAC.MOP file with *Z*-matrix geometry definitions. In a typical RECON run, atom type selection is first completed on the basis of atomic connectivity, at which point the program positions the first atom at the origin and the second along the $+z$ axis. The third atom is placed in the xz plane, and the fourth and subsequent atoms are situated according to the rotational angles read from the molecular *Z*-matrix. During each atom addition, the local best-fit surface rotation angle is located on the basis of surface shape, after which this atomic pair is considered to be in a combined but ''unmelded'' state. A trial geometry of the entire molecule may be generated in this fashion, or the first two atomic surfaces can be ''melded'' prior to the addition of another atom. Color-coded pre- and postmelding electron density and surface property figures are available in the Supplementary Material to this article. It is interesting to note that, in most cases, only minor changes are seen in the overall shape or color-encoded surface properties before and after the melding process. In some cases, the ''unmelded'' data is sufficient if the present atomic environments are similar to those from which the TAE atoms were derived, but in most cases, interatomic surface melding is desirable. In an earlier version of the melding algorithm, a total energy criterion was used to find the best new interatomic surface for a given atom pair. Within this model, the new surface position was also constrained to satisfy the zero flux condition $\nabla\rho \cdot N = 0$, where N is the normal vector of the new interatomic surface. That version of the surface adjustment algorithm relied on the principle that, within limits, each atom type has a unique ''electronic electronegativity'' which is defined as the ratio of atomic electronic kinetic energy to the integrated atomic population. A simple way to envision that relationship is to note that a more electronegative atom will receive more energetic benefit from an increment of electron density than a less electronegative one. This relationship is highly linear over small perturbations of each surface position within a given atom type, and can be used to generate a set of ''property derivatives'' which describe the sensitivity of each integrated or scalar atomic property to surface point motion. These property derivatives are then used to compute changes in the electronic kinetic energy, atomic electron population, and the surface electron density during the recombination melding process. Surface electronic properties used for generating QSAR/QSPR descriptors are also adjusted in this step. Each original TAE description includes default *ab initio* values for all integrated and surface properties of the atom, but once the ray lengths are changed during recombination, these values are adjusted accordingly.

This original method of atomic surface melding was found to be satisfactory for many molecules and bonding situations, but a more general method related to the atomic virial theorem was sought. The most recent version of the RECON melding algorithm makes use of a conceptual analogy between two unmelded electron density fragments and a pair of nonhomogeneous fluid volumes separated by a membrane. In the case of the separated atomic fragments, each atom assumes its shape as a result of the virial of forces acting upon it.[37] Atomic shape is described by the position of each element of the original set of zero-flux interatomic surfaces. Unless the two atoms to be melded were originally bound to each other, there will be portions of the two interatomic surfaces which will be mismatched when they are brought together by RECON. As pointed out by Mezey, such surface mismatches can initially lead to both undershoot and overshoot of each atomic volume.[38] Rotational optimization about the bond axis locates the local minimum of the surface mismatch. Once the atom pair is positioned, it is first necessary to determine if the bond length needs adjustment. This can be determined by examining the electron density and local electronic ''pressure'' of the two bond-path endpoints: If they are unequal, the surface position (and bond length) must be changed to equalize these values. This can be done as a two-equation/two-unknown problem by using the aforementioned surface property derivatives to extrapolate changes in both properties with respect to ray length. This is where the fluid analogy is helpful. The first equation shown below asserts that the total electronic energy of an atomic basin is equal to the one half the surface virial plus one half the basin virial. The surface virial is in turn related to the surface integral shown in the second equation, where one can use the electronic stress tensor to determine the surface ''pressure.'' The third equation shows the TAE interatomic surface encoding with electronic ''pressure'' property derivatives:

$$E_a(\Omega) = \frac{1}{2}V_B(\Omega) + \frac{1}{2}V_S(\Omega),$$

where V_B is the basin virial and V_S is the surface virial.

$$V_S(\Omega \mid \Omega') = \int dS(\Omega \mid \Omega', \vec{r})\vec{r} \cdot \overleftrightarrow{\sigma} \cdot \vec{n},$$

where $\overleftrightarrow{\sigma}$ is the stress tensor and $\overleftrightarrow{\sigma} \cdot \vec{n}$ is the pressure.

$$\frac{d(\overleftrightarrow{\sigma} \cdot \vec{n})}{dR} = \begin{array}{l}\text{rate of change of pressure normal}\\ \text{to the interatomic surface}\\ \text{with respect to ray length.}\end{array}$$

Once the proper bondlength is determined, the rest of the surface positions are tested for ''pressure'' and electron density equality in a radial pattern and optimized accordingly. Once this procedure has gone through one complete pass, the two original surfaces should not be one. The bond angle is held fixed in the prototype version of the RECON program, but methods of optimizing this parameter are being investigated. One possibility for bond angle optimization arises by examining and optimizing 1,3 zero-flux surface contacts between the geminal atoms.

The entire procedure may be repeated to allow propagation of polarization, because the property derivatives on each interatomic surface are dependent upon the positions of all other atomic surface elements. This is done using a set of *interdependency matrices* which alter the property values and derivative of each ray when any other ray length is changed during a melding operation. The property changes associated with the interatomic surface position changes are encoded into these matrices by comparing large sets of unmelded TAE assemblies with corresponding sets of FASTINT integrated atoms for the same molecules. Approximately 900 such comparisons have been made, and the results have been included in the existing TAE library. As the TAE library continues to expand, more of these comparisons will be performed to calibrate the new interdependency matrices.

QSAR and QSPR Descriptors from TAE Assemblies

The results of TAE calculations on the molecules in Figure 1 are presented in Table II. Within that table, the energies of each TAE assembly are compared to HF/6-31G* and HF/6-31+G* *ab initio* models. As in earlier examples of the TAE approach, errors in total energy for this data set are less than 2 kcal/mol. If one treats the new RECON-generated TAE electron density representation as a surface property-encoded *ab initio* electron density distribution, it becomes apparent that such molecular descriptors as molecular volume and surface area are readily accessible. Due to the structure of the TAE data, surface area and related surface integrals may be easily evaluated by collecting surface points into triads, and then using vector cross-products within the triads to calculate surface area increments and surface normals. These kinds of data can then be condensed into atomic and molecular contributions. Examples of these integrations and sums are illustrated in Table I. An alternative way to work with surface property data is to generate *surface histograms* of the various surface property distributions. These histograms may be generated with respect to each atom in the TAE assembly, or for the molecule as a whole. A standard set of 10 or 20 surface histogram bin sizes and limits have been defined for each surface property, which spans the range of each property that has been observed to date. To check the RECON results, TAE assembly property values were tested against *ab initio* standards using the new MARCH94 program.[39] This program uses GAUSSIAN94.CHK and AIM.WFN file data as input, and is able to compute all of the surface and 3D properties normally associated with TAE electron density assemblies. As a result of its actual use of *ab initio* theory, however, calculations using the MARCH94 method can be quite time consuming, but are practical for molecules of the size used in this study. The MARCH94 program takes its name from the *Marching cube algorithm* which is used to generate the molecular electron density isosurface for use in property encoding. This very rapid routine is followed by an algorithm which treats points on the 0.002-e/au^3 isodensity surface as if they were part of a TAE assembly. The descriptors calculated by RECON were compared to those obtained at the HF/6-31+G*//HF/6-31G* level of theory by Gaussian94/MARCH94 program for all 22 compounds. In every case, the RECON values were found to be within 1% of the MARCH94 results.[40] The molecular descriptor information from RECON are stored in QMF (QSAR Meta File) and EXCEL spraedsheet format. The EXCEL data are available as Supplementary Material. The QMF file format is common to all application programs

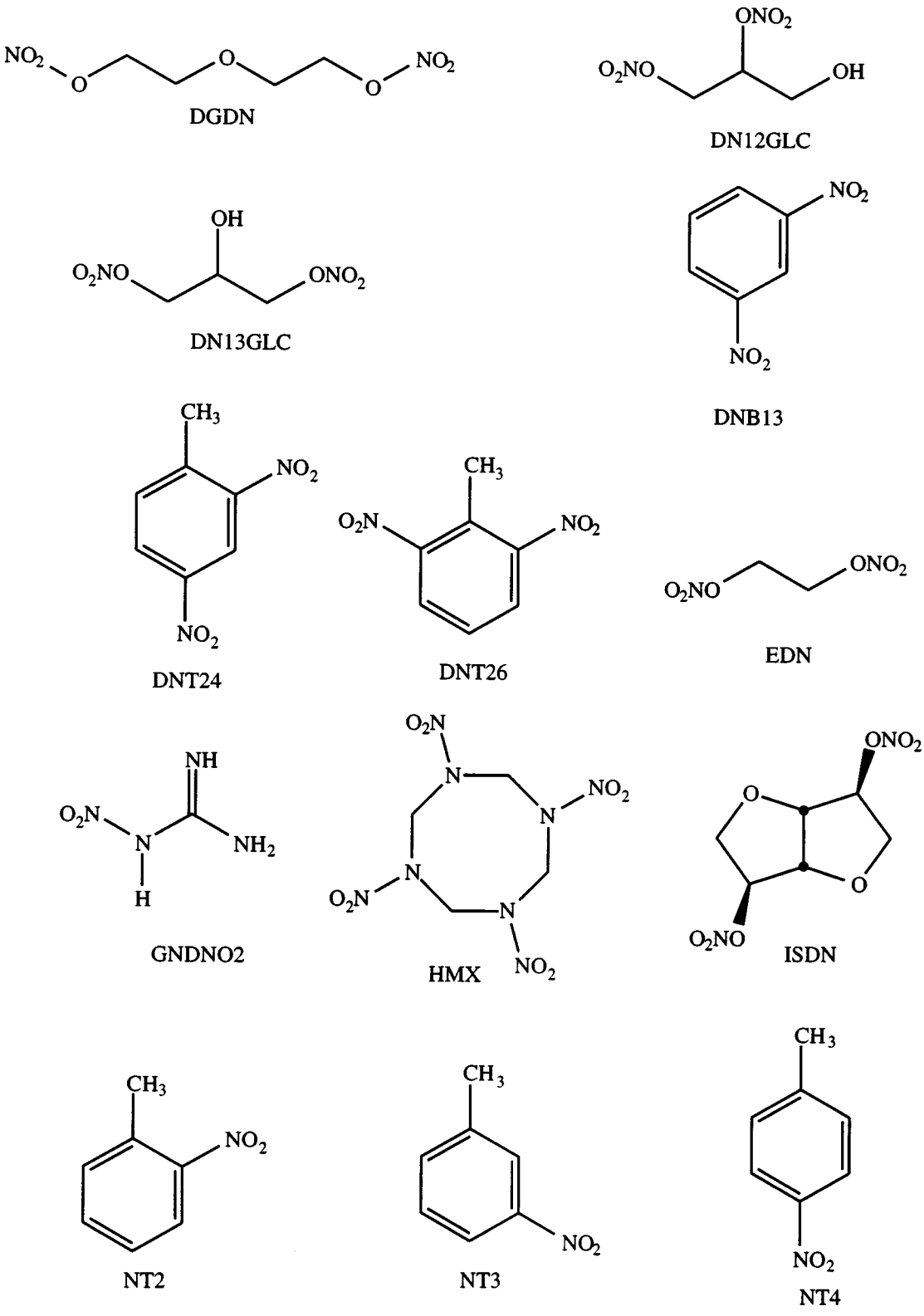

FIGURE 1. The compounds used in the capacity factor data set.

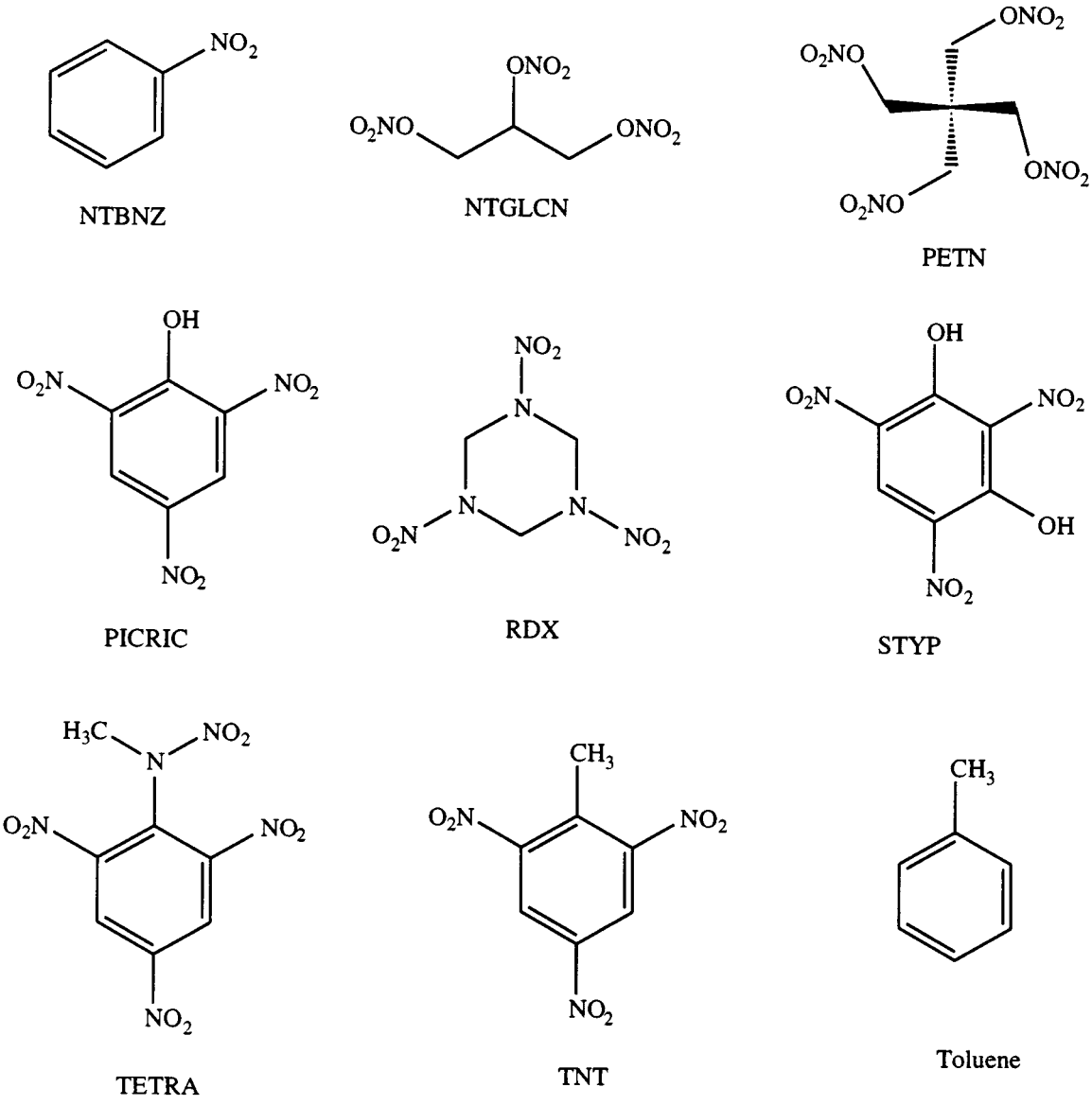

FIGURE 1. (continued)

used in our research group and within the Kodak Scientific Computing Team as a means of rapid data exchange. This standardization allows the statistical analysis programs to use data in the same form as they were generated by the diverse array of QSAR tools in our group. As with any molecular structure file format, the QMF file is capable of carrying geometric data and bond connectivity information. The QMF format is also able to store virtually unlimited amounts of atomic, regional, bond, and molecular level QSAR descriptor information.

Variable Selection and PLS Modeling with TAE Surface Indices

Whenever a large number of descriptors must be evaluated in order to find the best ones to include in a concise QSPR model, it is often useful to employ an automated variable selection technique such as all-possible-subsets (APS) regression. This approach is even more useful when a large number of unfamiliar descriptors make up the variable pool. In the case of TAE-generated

TABLE II.
High-Energy Materials HF/6-31+G*//HF/6-31G* (Hartrees) vs. TAE RECON Energies.

Compound	HF/6-31G*	HF/6-31+G*	TAE/RECON
1,3-Dihydroxy-2,4,6-trinitrobenzene (Styphnic acid, STYP)	−990.80128	**−990.82907**	**−990.82823**
1,2-Dinitroglycerol (DN12GLC)	−749.66684	**−749.68918**	**−749.68612**
1,3-Dinitroglycerol (DN13GLC)	−749.67753	**−749.69919**	**−749.70021**
meta-Dinitrobenzene (DNB13)	−637.64032	**−637.65829**	**−637.65913**
2,4-Dinitrotoluene (DNT24)	−676.67425	**−676.69226**	**−676.69436**
2,6-Dinitrotoluene (DNT26)	−676.66474	**−676.68287**	**−676.68945**
Diglycol dinitrate (DGDN)	−788.70256	**−788.72572**	**−788.72623**
Ethylene glycol dinitrate (EDN)	−635.78875	**−635.80723**	**−635.80814**
Hexahydro-1,3,5-trinitro-1,3,5-triazine (Hexagen, RDX)	−892.50534	**−892.52926**	**−892.52883**
Isosorbide dinitrate (ISDN)	−939.28887	**−939.31363**	**−939.31411**
N-methyl-*N*,2,4,6-tetranitroaniline (Tetryl, TETRA)	−1138.57806	**−1138.6100**	**−1138.6110**
Nitrobenzene (NTBNZ)	−434.17523	**−434.18795**	**−434.18801**
Nitroglycerine (NTGLCN)	−953.10190	**−953.12845**	**−953.12747**
Nitroguanidine (Picrite, GNDNO2)	−407.56437	**−407.57840**	**−407.57935**
meta-Nitrotoluene (NT3)	−473.21259	**−473.22522**	**−473.22499**
ortho-Nitrotoluene (NT2)	−473.20772	**−473.22036**	**−473.22011**
para-Nitrotoluene (NT4)	−473.21363	**−473.22634**	**−473.22535**
Octahydro-1,3,5,7-tetranitro-1,3,5,7-tetrazocine (Octagen, HMX)	−1190.0020	**−1190.04389**	**−1190.0423**
Pentaerythretol tetranitrate (PETN)	−1309.43453	**−1309.47221**	**−1309.47115**
Picric acid (PICRIC)	−915.94982	**−915.97563**	**−915.97487**
Trinitrotoluene (TNT)	−880.12508	**−880.14886**	**−880.14835**

QSPR indices, there were 122 different atomic and molecular descriptors to choose from. To make the selection easier, the APS regression technique was used to find the best five parameter CPS capacity factor models for 21 of the sample compounds excluding DGDN, for which there was no experimental CPS column retention data. All 22 molecules were included in the APS regression calculations for the ODS column data. Five-parameter models were chosen to be comparable to Lowrey's earlier five-parameter TLSER work on this data set. The Microsoft Windows™ program ''PLSPC4'' was utilized for the APS regression and PLS calculations.[41] The results of APS calculations on the CPS and ODS data are presented in Table III. It is interesting to note that the two HPLC columns have quite different primary modes of interaction with substrate molecules in the mobile phase, and are best modeled by different sets of five molecular descriptors. This change in descriptor basis is not surprising: Although it is very unlikely that the value of any single new surface index alone would be directly proportional to only one kind of binding model, it is our expectation that, when taken together, combinations of electron density-based surface descriptors will be able to represent each of the common modes of noncovalent interaction. In Lowrey's earlier TLSER work, the following molecular descriptors were used: volume, volume normalized polarizability, covalent H-bond donor capability, ionic H-bond donor capability, covalent H-bond acceptor capability, and an intercept term. After the two or three outliers were removed, the two most important terms for both columns were found to be the molecular volume and the ionic H-bond acceptor capability, as represented by the highest positive charge on any hydrogen at the MOPAC level of theory. Of lesser importance in both cases was the covalent H-bond donor capability descriptor which was derived from the MOPAC energy difference between the molecular HOMO and water LUMO. The volume normalized polarizability was also included in each model at a lower significance level. For comparison, the TAE regression data for both CPS and ODS columns can be found in Table IV. It should be noted that models based on TAE-derived surface property indices gave superior models for the full data sets: there were no outliers to be removed from the data. The crossvalidated R^2 for the worst case (CPS) with no data removed was found to be 0.989 for TAEs and 0.804 for TLSERs. The cross-validated R^2 for the TAE model of the ODS column capacity factor data was

TABLE III.
All-Possible Subsets Regression Results for CPS and ODS Columns

CPS column regression variable	Appearance frequency in best 260 subsets
DELRHONA3: Del(RHO) · *N* (area 3)	260
SIGA9: surface integral of *G* (area 9)	260
SIGA6: surface integral of *G* (area 6)	180
SIK: surface integral of *K* (molecular)	130
PIPMAX: PIP maximum	100
PIP15: PIP (area 14)	90
SIGA4: surface integral of *G* (area 4)	90
PIV: GIPF electrostatic balance parameter	80
SIGA5: surface integral of *G* (area 5)	60

ODS column regression variable	Appearance frequency in best 264 subsets
PIV: GIPF electrostatic balance parameter	226
SIGA8: surface integral of *G* (area 8)	129
PIP14: PIP (area 14)	110
PIP16: PIP (area 16)	88
SIGA10: surface integral of *G* (area 10)	82
DELRHONA3: Del(RHO) · *N* (area 3)	79
PIP17: PIP (area 17)	79
SIKA9: surface integral of *K* (area 9)	74
Volume: molecular volume in au^3	52

found to be 0.997. It is important to distinguish that these crossvalidated correlation coefficients represent much higher values of R and R^2 as commonly used in least-square regression. In all of the regression models used in this work, standard crossvalidation techniques were used to evaluate the quality of the models. Figures 2 and 3 illustrate the best TAE and TLSER models for the ODS column data. The TLSER ODS data are shown with the outliers plotted on the figure, but not included as variables in the regression procedure. Table V contains all of the experimental and predicted values for the ODS column. TLSER data are also included in those tables both with and without outliers for each model.

All-possible-subsets (APS) regression using the TAE indices and CPS column data indicated that the best model utilized the DelRhoNA3, SIGA9, SIK, SIGA6, and PIPmax descriptors. For the ODS case, the most important variables were found to be PIV, SIGA8, PIP14, PIP16, and SIGA10. It is interesting to note that when the TLSER indices were combined with the TAE indices for APS regression analysis, the top eight variables and top ten models for both CPS and ODS data contained only TAE surface property indices. In the examination of ODS column data by APS, the ninth most important descriptor was found to be molecular volume. Figures 4 and 5 show the best-fit models for TAE CPS and TLSER CPS models. The predicted and experimental CPS column data is shown in Table VI. As in the ODS column case, the TLSER outliers were not included in the model.

Comparison of the TLSER coefficients with those of the TAE indices provide useful information about which physical interaction modes are represented by each of the TAE descriptors. First, because the TAE descriptors are not orthogonal, it is necessary to accept a less clear distinction about how much of each index describes the energy of each mode of noncovalent interaction. More data and analysis will be required before the information content of the new indices can be fully understood. For now, the definition of each index type allows some speculation: DelRhoNA3 represents the height of the third histogram bin in the Del RHO · N data category. The DelRho · N descriptor class tells us how fast the electron density falls off with respect to distance from the surface, where the lower five bins represent the slow fall-off rates consistent with surfaces over π-systems and heteroatoms. Because the value of the histogram bin also signifies the amount of surface area of the molecule which has a certain range of the property, it gives some measure of molecular size. The SIK index is also related to size, but it is actually the whole surface integral of the "K" formulation of the electronic kinetic energy density. Because electronic kinetic energy values are believed to be representative of hydrogen bonding activity (both donor and acceptor, depending upon the values), it may serve as a combination variable describing surface area and potential acceptor/donor activity. The SIGA6 and SIGA9 bins are near the high end

TABLE IV.
Regression Coefficients for Five-Parameter CPS and ODS Capacity Factor TAE PLS Models.

CPS descriptor	Normalized coefficient	Coefficient in original units[a]	ODS descriptor	Normalized coefficient	Coefficient in original units[b]
DELRHONA3	0.6346	0.02	PIV	−1.063	−0.06264
SIGA9	−0.4553	−0.03639	SIGA8	−0.8981	−0.0264
SIK	−0.3769	−0.9843	PIP14	0.7976	0.004382
SIGA6	0.3437	0.000833	PIP16	0.366	0.002387
PIPMAX	−0.2856	−0.8452	SIGA10	0.285	0.06707

[a] Intercept = 0.4341; CPS crossvalidated R^2 = 0.989.
[b] Intercept = 1.465; ODS crossvalidated R^2 = 0.997.

of the surface areas with large "G" kinetic energy densities, K and G electronic kinetic energy densities are often correlated to some degree, so it can be said that this region of the energy spectrum may describe hydrogen bond donor capability. The PIPmax descriptor is actually a GIPF index calculated using TAE methodology. PIPmax is associated with the highest value of the Politzer local ionization potential (PIP) found on the molecular surface. This is believed to describe the hydrophobicity and charge transfer capabilities of the molecule. The ODS column results can be analyzed in a similar fashion: PIV is a GIPF[42] consisting of the average deviation of the surface electrostatic potential. This parameter is usually interpreted as a measure of charge separation within the molecule. The high negative importance of this parameter in the ODS model (Table IV) suggests that molecules with large internal charge separations will not interact favorably with the nonpolar ODS stationary phase. The importance of the SIGA8 descriptor for the hydrophobic ODS column model can be interpreted to mean that hydrogen bond donor capability is detrimental to retention on this column. The high weight placed on PIP14, and to a lesser extent PIP16, appears to show that the molecules with the most affinity for this column would be difficult to ionize—an observation consistent with hydrophobicity. The SIGA10 descriptor involvement appears to be a correction factor for some of the molecules for which the SIGA8 coefficient might be too large. As additional examples of TAE surface descriptor modeling are performed, our understanding of the relationships between these electronic indices and more traditional interaction modes should be enhanced.

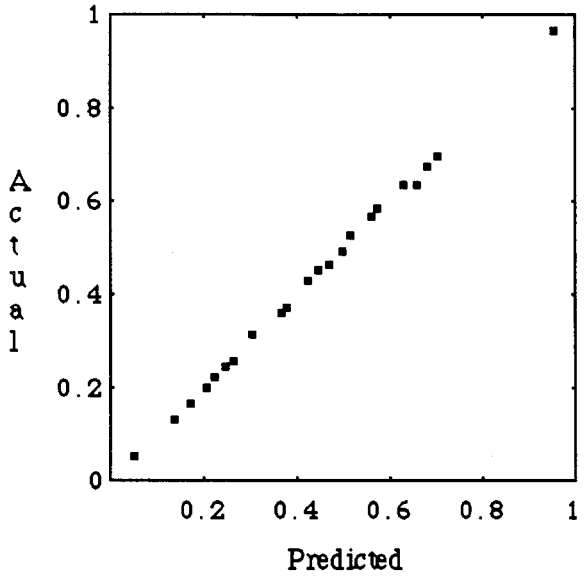

FIGURE 2. Experimental vs. TAE ODS Log(k') capacity factor data set.

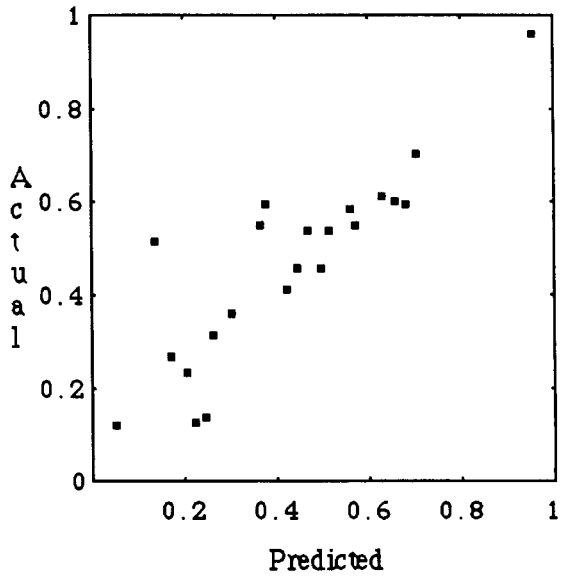

FIGURE 3. Experimental vs. TLSER ODS Log(k') capacity factor data.

TABLE V.
Predicted and Experimental Capacity Factor Data for ODS Column.

Molecule ID	ODS Log k' experimental	ODS k' experimental	ODS TAE Log k'	ODS TAE k'	ODS TLSER Log k' with outliers	ODS TLSER k' with outliers	ODS TLSER Log k', no outliers	ODS TLSER k', no outliers
DGDN	3.64E − 01	1.31E + 00	3.60E − 01	1.29E + 00	4.79E − 01	2.01E + 00	5.47E − 01	2.52E + 00
DN12GLC	2.46E − 01	7.60E − 01	2.43E − 01	7.50E − 01	1.49E − 01	4.09E − 01	1.37E − 01	3.71E − 01
DN13GLC	2.23E − 01	6.70E − 01	2.24E − 01	6.75E − 01	1.37E − 01	3.71E − 01	1.24E − 01	3.30E − 01
DNB13	4.25E − 01	1.66E + 00	4.27E − 01	1.67E + 00	3.94E − 01	1.48E + 00	4.10E − 01	1.57E + 00
DNT24	5.72E − 01	2.73E + 00	5.83E − 01	2.83E + 00	5.16E − 01	2.28E + 00	5.49E − 01	2.54E + 00
DNT26	5.61E − 01	2.64E + 00	5.64E − 01	2.66E + 00	5.45E − 01	2.51E + 00	5.84E − 01	2.84E + 00
EDN	3.03E − 01	1.01E + 00	3.12E − 01	1.05E + 00	3.52E − 01	1.25E + 00	3.62E − 01	1.30E + 00
GNDNO2	5.31E − 02	1.30E − 01	5.21E − 02	1.27E − 01	1.76E − 01	5.00E − 01	1.17E − 01	3.09E − 01
HMX	1.34E − 01	3.60E − 01	1.32E − 01	3.55E − 01	3.58E − 01	1.28E + 00	5.16E − 01	2.28E + 00
ISDN	3.75E − 01	1.37E + 00	3.70E − 01	1.34E + 00	5.24E − 01	2.34E + 00	5.97E − 01	2.95E + 00
NT2	6.28E − 01	3.25E + 00	6.35E − 01	3.32E + 00	5.76E − 01	2.77E + 00	6.13E − 01	3.10E + 00
NT3	6.79E − 01	3.78E + 00	6.75E − 01	3.73E + 00	5.50E − 01	2.55E + 00	5.92E − 01	2.91E + 00
NT4	6.54E − 01	3.51E + 00	6.35E − 01	3.32E + 00	5.63E − 01	2.66E + 00	6.00E − 01	2.98E + 00
NTBNZ	4.68E − 01	1.94E + 00	4.63E − 01	1.90E + 00	5.17E − 01	2.29E + 00	5.37E − 01	2.44E + 00
NTGLCN	4.94E − 01	2.12E + 00	4.93E − 01	2.11E + 00	4.12E − 01	1.58E + 00	4.58E − 01	1.87E + 00
PETN	7.00E − 01	4.01E + 00	6.96E − 01	3.97E + 00	5.94E − 01	2.93E + 00	7.03E − 01	4.05E + 00
PICRIC	2.07E − 01	6.10E − 01	2.01E − 01	5.89E − 01	2.46E − 01	7.62E − 01	2.33E − 01	7.10E − 01
RDX	2.60E − 01	8.20E − 01	2.55E − 01	7.99E − 01	2.12E − 01	6.29E − 01	3.12E − 01	1.05E + 00
STYP	1.70E − 01	4.80E − 01	1.65E − 01	4.62E − 01	2.70E − 01	8.62E − 01	2.70E − 01	8.62E − 01
TETRA	4.43E − 01	1.77E + 00	4.49E − 01	1.81E + 00	3.40E − 01	1.19E + 00	4.56E − 01	1.86E + 00
TNT	5.16E − 01	2.28E + 00	5.23E − 01	2.33E + 00	4.97E − 01	2.14E + 00	5.37E − 01	2.44E + 00
TOLUENE	9.52E − 01	7.95E + 00	9.65E − 01	8.23E + 00	1.02E + 00	9.35E + 00	9.61E − 01	8.14E + 00

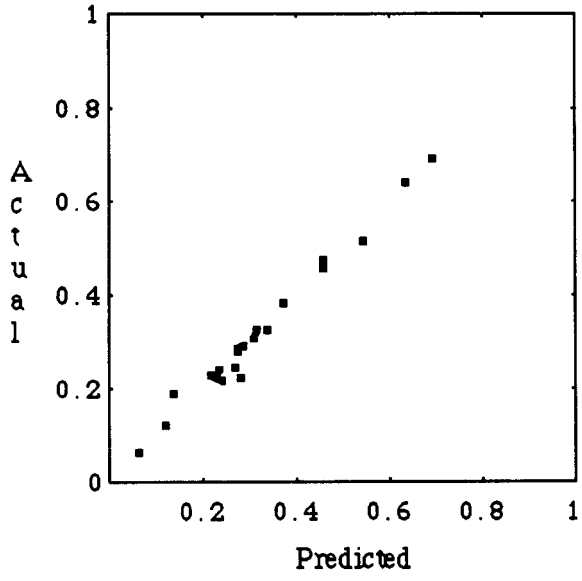

FIGURE 4. Experimental vs. TAE CPS Log(k') capacity factor data.

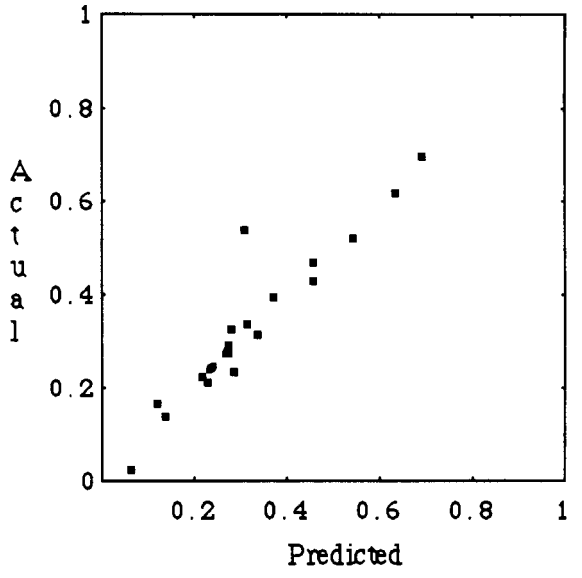

FIGURE 5. Experimental vs. TLSER CPS Log(k') capacity factor data.

TABLE VI.
Predicted and Experimental Capacity Factor Data for CPS Column.

ID	CPS log k'	CPS k'	CPS TAE Log k'	CPS TAE k'	CPS TLSER Log k'	CPS TLSER k'	CPS TLSER Log k', with no outliers	CPS TLSER k' with no outliers
DGDN	Not observed	Not observed	4.52E − 01	1.83E + 00	Not reported	Not reported	Not reported	Not reported
DN12GLC	2.38E − 01	7.30E − 01	2.19E − 01	6.54E − 01	1.99E − 01	5.81E − 01	2.43E − 01	7.50E − 01
DN13GLC	2.18E − 01	6.50E − 01	2.26E − 01	6.84E − 01	1.85E − 01	5.31E − 01	2.20E − 01	6.60E − 01
DNB13	2.83E − 01	9.20E − 01	2.90E − 01	9.49E − 01	2.53E − 01	7.91E − 01	2.33E − 01	7.10E − 01
DNT24	3.37E − 01	1.17E + 00	3.28E − 01	1.13E + 00	3.24E − 01	1.11E + 00	3.14E − 01	1.06E + 00
DNT26	3.12E − 01	1.05E + 00	3.24E − 01	1.11E + 00	3.40E − 01	1.19E + 00	3.36E − 01	1.17E + 00
EDN	2.79E − 01	9.00E − 01	2.25E − 01	6.80E − 01	3.46E − 01	1.22E + 00	3.24E − 01	1.11E + 00
GNDNO2	6.45E − 02	1.60E − 01	6.42E − 02	1.59E − 01	2.90E − 02	6.91E − 02	2.50E − 02	5.93E − 02
HMX	6.36E − 01	3.32E + 00	6.40E − 01	3.37E + 00	5.82E − 01	2.82E + 00	6.17E − 01	3.14E + 00
ISDN	3.08E − 01	1.03E + 00	3.10E − 01	1.04E + 00	4.73E − 01	1.97E + 00	5.38E − 01	2.45E + 00
NT2	2.67E − 01	8.50E − 01	2.47E − 01	7.66E − 01	2.65E − 01	8.41E − 01	2.72E − 01	8.71E − 01
NT3	2.74E − 01	8.80E − 01	2.80E − 01	9.07E − 01	2.65E − 01	8.41E − 01	2.90E − 01	9.50E − 01
NT4	2.72E − 01	8.70E − 01	2.83E − 01	9.17E − 01	2.70E − 01	8.62E − 01	2.76E − 01	8.88E − 01
NTBNZ	2.30E − 01	7.00E − 01	2.23E − 01	6.73E − 01	2.20E − 01	6.60E − 01	2.12E − 01	6.29E − 01
NTGLCN	4.59E − 01	1.88E + 00	4.77E − 01	2.00E + 00	4.49E − 01	1.81E + 00	4.70E − 01	1.95E + 00
PETN	6.91E − 01	3.91E + 00	6.89E − 01	3.88E + 00	6.31E − 01	3.28E + 00	6.95E − 01	3.95E + 00
PICRIC	1.34E − 01	3.60E − 01	1.91E − 01	5.51E − 01	1.61E − 01	4.49E − 01	1.34E − 01	3.61E − 01
RDX	4.56E − 01	1.86E + 00	4.59E − 01	1.88E + 00	4.25E − 01	1.66E + 00	4.31E − 01	1.70E + 00
STYP	1.21E − 01	3.20E − 01	1.18E − 01	3.12E-01	1.79E − 01	5.10E − 01	1.67E − 01	4.69E − 01
TETRA	5.43E − 01	2.49E + 00	5.14E − 01	2.27E + 00	5.11E − 01	2.24E + 00	5.19E − 01	2.30E + 00
TNT	3.71E − 01	1.35E + 00	3.83E − 01	1.41E + 00	4.05E − 01	1.54E + 00	3.94E − 01	1.48E + 00
TOLUENE	2.33E − 01	7.10E − 01	2.40E − 01	7.37E − 01	2.10E − 01	6.22E − 01	2.41E − 01	7.42E − 01

Conclusions

The results presented here have shown that, in addition to being able to match *ab initio* property data to within 1%, the TAE/RECON method is capable of rapidly producing valuable new QSAR/QSPR descriptors which are derived from the electronic properties of molecular van der Waals surfaces. Although these same descriptors can be obtained directly through *ab initio* calculations, the TAE/RECON program is capable of generating them at least 300 times faster than double-zeta *ab initio* for medium-sized molecules. Because the TAE/RECON program has been shown to be practical with molecules up to the size of small proteins, such as FKBP,[43] TAE/RECON QSAR, and QSPR, work is now being undertaken in the areas of DNA, RNA, and carbohydrates.

Acknowledgments

The authors thank Dr. Al Lowrey, from the Naval Research Laboratory, and Dr. George Famini, from the Aberdeen proving grounds, for sharing their experimental data and providing an early copy of their manuscript containing the high-energy material HPLC data and TLSER analysis.

References

1. H. van de Waterbeemd and B. Testa, In *Advances in Drug Research*, Vol. 16, B. Testa, Ed., Academic Press, New York, 1987.
2. C. Hansch, *Accounts Chem. Res.* **26**, 147 (1993).
3. C. J. Cramer, G. R. Famini, and A. H. Lowrey, *Accounts Chem. Res.*, **26**, 599 (1993).

4. T. Fumita, J. Iwasa, and C. Hansch, *J. Am. Chem. Soc.*, **36**, 5175 (1964).
5. For some relevant examples, see H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches*, VCH, Weinheim, 1993, and the references therein.
6. (a) C. Hansch, *Accounts Chem. Res.*, **2**, 232 (1969); (b) T. Fujita, J. Iwasa, and C. Hansch, *J. Am. Chem. Soc.*, **36**, 5175 (1964).
7. A. T. Balaban, I. Motoc, D. Bonchev, and O. Mekenyan, *Topics Curr. Chem.*, **114**, 21 (1983).
8. P. G. Seybold, M. May, and U. A. Bagal, *J. Chem. Educ.*, **64**, 575 (1987).
9. For a discussion of the electrotopological approach, see L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure Activity Analysis*, Research Studies Press, Chichester, UK, 1986.
10. L. B. Kier and L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976.
11. L. B. Kier, *Molecular Orbital Theory in Drug Research*, Academic Press, New York, 1971.
12. For examples and references, see (a) S. M. Green, G. R. Marshall, *Trends Pharmacol. Sci.*, **16**, 285 (1995); (b) K. B. Rhyu, H. C. Patel and A. J. Hopfinger, *J. Chem. Informat. Comput. Sci.*, **35**, 771 (1995).
13. H. Hagelin, J. S. Murray, T. Brinck, M. Berthelot, and P. Politzer, *Can. J. Chem.-Revue Canadienne Chim.*, **73**, 483 (1995).
14. J. S. Murray, K. Paulsen, and P. Politzer, *Proc. Indian Acad. Sci.—Chem. Sci.*, **106**, 267 (1994).
15. Hehre, W. *Spartan* product literature, 1993.
16. J. S. Murray, T. Brinck, P. Lane, K. Paulsen, and P. Politzer, *Theochem*, **307**, 55 (1994).
17. A. H. Lowrey, C. J. Cramer, J. J. Urban, and G. R. Famini, *Comput. Chem.*, **19**, 209 (1995).
18. A. D. Headley, S. D. Starnes, L. Y. Wilson, and G. R. Famini, *J. Org. Chem.*, **59**, 8040 (1994).
19. A. Lowery, private communication, Naval Research Labs.
20. M. G. Ford and D. Livingstone, *J. Quant. Struct.-Act. Rel.*, **9**, 107 (1990).
21. I. Lukovits, *J. Med. Chem.*, **26**, 1104 (1983).
22. C. M. Breneman and L. W. Weber, In *The Application of Charge Density Research to Chemistry and Drug Design* (NATO ASI Series), G. A. Jeffrey, J. F. Piniella, Eds., Plenum Press, 1991.
23. C. M. Breneman, T. R. Thompson, M. Rhem, and M. Dung, *Comput. Chem.*, **19**, 161 (1995).
24. R. F. W. Bader, *Atoms in Molecules: A Quantum Theory*, Oxford Press, Oxford, UK, 1990.
25. For example, see K. B. Wiberg and C. M. Greneman, *J. Am. Chem. Soc.*, **114**, 831 (1992).
26. R. F. W. Bader and P. Becker, *Chem. Phys. Lett.*, **148**, 452 (1988).
27. V. Pichon-Pesme, C. Lecomte, R. Wiest, and M. Benard, *J. Am. Chem. Soc.*, **114**, 2713 (1992).
28. R. F. W. Bader, *Accounts Chem. Res.*, **18**, 9 (1985).
29. P. D. Walker and P. G. Mezey, *J. Am. Chem. Soc.*, **115**, 12423 (1993).
30. Techniques for extracting molecular electrostatic potential information from MELDA assemblies are said to be undergoing development, and may be included in that package by the time this article is published.
31. The local ionization potential is an energy-weighted electron density property first described by Professor Peter Politzer. Extrema of this function were used in prior QSAR/QSPR studies by the Politzer group, and in that role the IP parameter was included as one of the GIPFs (generalized interaction properties functions). See P. Sjoberg, J. S. Murray, T. Brinck, and P. Politzer, *Can. J. Chem.*, **68**, 1440 (1990).
32. The importance of this standard orientation will become clear as the recombination algorithm is discussed.
33. R. F. W. Bader, M. T. Carroll, J. R. Cheeseman, and C. Chang, *J. Am. Chem. Soc.*, **109**, 7968 (1987).
34. The BRILLIG program, T. R. Thompson, Rensselaer Polytechnic Institute, Troy, NY, 1992.
35. T. R. Thompson, Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, NY, 1993.
36. Supplementary Material is available from the author, or via the Internet (Wiley site).
37. R. F. W. Bader, *Atoms in Molecules: A Quantum Theory*, Oxford Press, Oxford, UK, 1990.
38. P. D. Walker, P. G. Mezey, *J. Comput. Chem.*, **16**, 1238 (1995).
39. The MARCH94 program was designed locally and written by Marlon Rhem, Chris Whitehead, and Curt Breneman.
40. A detailed comparison of TAE/RECON data to that obtained form the MARCH94 *ab initio* program is the subject of another study. The results of TAE/RECON reconstructions were compared to MARCH94 calculations for a set of 110 molecules in ref. 23.
41. The PLSPC4 program package was written in Microsoft Windows™ APL by Dr. Charles Heckler of Eastman Kodak, Inc. Among many other analytical and graphical capabilities, this versatile program includes partial least-squares analysis (PLS) as well as principle component analysis (PCA) and all-possible subsets (APS) regression. The PLSPC program is equipped to read QSAR Meta File (QMF) files, EXCEL worksheets, free-format text files, and its own proprietary archive format. The PLSPC program is a limited-distribution product that was made available to members of Kodak Scientific Computing group. Future public release of the program is possible.
42. Since Politzer's GIPF values are usually calculated on the 0.001-e/au^3 isosurface, and arise from STO-5G or 6G wave functions, they are not directly comparable with the GIPF values calculated using TAEs. Since they are meant to be proportional to various molecular properties, this difference does not matter as long as GIPF indices from the two sources are not mixed in a given model. TAE-derived GIPFs are derived at a HF/6-31+G* equivalent level of theory.
43. C. M. Breneman, M. Rhem, T. R. Thompson, and M. H. Dung, In the ACS Symposium volume *Modeling the Hydrogen Bond*, D. Smith, Ed., August 1993, pp. 152–174.