Distribution-Based Descriptors of the Molecular Shape

Yegor Zyrianov*

Department of Chemistry, Purdue University, 560 Oval Drive, West Lafayette, Indiana 47907-2084

Received January 5, 2005

A rational design of economically cost-effective chemical libraries as well as successful data mining during a process of drug discovery employs a vast array of the molecular descriptors. Despite the huge importance of this area of the research there is still a need for the further development of the simple, intuitive, easily calculable, specific and size-invariant parameters of the molecular shape. Here we present ab initio calculation of the molecular volumes and expectations for the molecular areas of projection. These molecular size parameters were used as a basis to define a group of novel descriptors of the molecular shape. A set of molecular descriptors was developed: ovality, roughness, size-corrected parameters, and the parameters derived from the higher central momenta of the distributions of the size descriptors—skewness and kurtosis. The rationale for the construction of the descriptors was first to calculate the descriptors of the molecular size along many directions in the space and second to use the statistical parameters of the distribution of those descriptors as the shape descriptors. The size descriptors well suited for the above purpose and discussed in this paper are generalized molecular radii and the molecular areas of projection. Molecular volume and projection area-derived descriptors were calculated and their applicability as shape descriptors was illustrated using exploratory methods (factor analysis and hierarchical cluster analysis). The shape descriptors appear to be promising in their ability to discriminate and classify the molecular shapes e.g. spheroids, disklike, rodlike, starlike, crosslike, anglelike, etc.

1. INTRODUCTION

The modern methods of combinatorial chemistry allowed producing libraries of gigantic sizes. The problem encountered by researchers in the past decade is that the increase in the sizes of the libraries does not always guarantee the diversity represented by the library.[1] Parallel to the synthetic solutions, such as "diversity-oriented synthesis", sets of common molecular descriptors have been used or proposed to ensure the libraries' molecular diversity.[2] Still, certain aspects of the molecular diversity or their numerical measures have not been formulated, quantified, or promoted to the level of the measurement standard in the field. One of the underemployed properties is a basic property of a molecular shape.

Interestingly, a description of the molecular shape has actually attracted a substantial amount of a research effort in the past decade or more due to its importance in the understanding, search, and prediction of molecular interaction and protein inhibition, particularly in the area of drug design.[3−5] Simultaneously, the development of computational technologies allows the solution of the more and more challenging computational problems. In addition to the mainstream application of the shape descriptors in drug design these descriptors are of a huge interest for such areas as chromatography and adsorption processes in general, understanding of the fundamental and applied physical-chemical properties of the matter, such as equations of state, prediction, and description of certain higher organizations of the matter (bilayers, liposomes, and nanostructures of various nature, etc.). For a short outline of the research in the area of molecular shape it is worth mentioning the relevant works on topological molecular connectivity indices (Wiener index,[6] Hosoya's parameter,[7] molecular complexity index,[8] and Randic index[9]), Shape Group Methods (SGM),[5] Continuous Symmetry (Chirality) Measure (CSM,[10] CCM[11]), Weighted Holistic Invariant Molecular descriptors (WHIM);[12] quite different, yet adjacent to the field of molecular shape are Comparative Molecular Field Analysis (CoMFA) and other grid-based methods of the molecular alignment and the calculations of molecular similarities.

It is a crucial step in any QSAR, QSPR, or virtual screening procedure to reduce the initial amount of information about the molecules to be investigated to some small set of molecular descriptors. The question always remains what are the essential and fundamental descriptors of the molecule? Additionally, as a matter of general scientific interest, these descriptors ought to have clear physical, chemical, or geometrical interpretation. Overloading of the input data with information may lead to the results distant from the ones desired. As an extreme example, if to take a matrix containing electronic properties (charge density, potential) at the points of a dense grid surrounding and penetrating a molecule (which, by itself, is quite a complete and informative description of the molecule!), and try to directly find a linear relationship predicting some property of the molecule, the outcome will be nonsensical. Even in purely 3D approaches the implicit goal is to reduce the whole continuum of the information on the molecule, discard the irrelevant data on the areas not essential for the molecular interaction, and extract the data on the critical molecular locales.

* Corresponding author e-mail: zyrianov@purdue.edu.

10.1021/ci050005l CCC: $30.25
Published on Web 04/16/2005

Among the other basic reasons for using simple and concise size and shape descriptors is their ability to discard the molecular candidates before the thorough and expensive synthesis and activity studies have been undertaken. For example, if a molecule does not fit the active site of an enzyme because of its size and shape parameters, it should not be considered further, and even no other computed descriptors, such as the ones taking into account the electrostatic properties of the molecule, are needed to make this decision. Unfortunately, in many cases "the receptor site" cannot be described in simple terms of "width" and "depth", and, in addition to that, there is still a lack of simple and informative descriptors of the molecular shape.

Important characteristics of a molecular descriptor are (a) the source of it, particularly, obtained empirically (including, for example, those computed using models of molecules made of van der Waals atomic spheres or from measurements of certain physical chemical properties) vs calculated semiempirically or ab initio on the basis of quantum-mechanical principles and also (b) the level of the description of a molecule, either based entirely on the atomic connectivity (possibly with the addition of the interatomic distance information) or invoking 3D information (possibly with the mapping some relevant continuous property data). A noticeable reduction of computational costs shifts the attention of the research community to the second choices in the both dichotomies above.

2. DESCRIPTORS OF MOLECULAR SIZE

The notion of molecular volume has many variations of its definition and the procedures for calculation.[13−17] For more comprehensive bibliography on molecular volumes see ref 3.

The source of the data for the molecular size computations is either experimentally derived van der Waals atomic radii[13−15] or calculated quantum-mechanically.[16,17] The major objection against the first approach is that it is intrinsically approximational. This disadvantage is completely overcome by the fact that this approximation is sufficient for most practical purposes and the calculations are much faster than in the case of semiempirical and ab initio calculations. There is another problem about quantum-mechanical calculations. An output of calculations in this case is always a certain continuous function defined, in principle, on the infinite interval of a space. Volume and size are finite parameters of a molecule. Trying to link the quantum-mechanical output calculations with some metric property of the molecule one must unavoidably invoke certain data on the bulk substance to define the cutoff of values of quantum-mechanical function which would allow making the linkage mentioned.

Previously, cutoff values of electronic densities (ED) in the range $0.001 \div 0.005$ e·Bohr^{-3} were used on different occasions.[16−21] The main approaches utilized by the researchers were either to take the contour enclosing a certain arbitrary amount of charge (98%) or to make the linkage between the ab initio and some empirical measure of the molecular volume. The most common empirical references for the molecular volumes were the volumes estimated on the basis of van der Waals (vdW) radii obtained from crystallographic studies,[22] or effective molecular radii from equations of state (ES),[23] or equations describing other bulk properties of substances (e.g. viscosity, molecular scattering).[24] In parametric equations, size parameters could be obtained from fitting the coefficients in EOS with various expressions for molecular potential (Lennard-Jones, Stockmayer, etc.) or without invocation of any assumptions about potential but using some other theoretical assumptions on the nature and physical meaning of coefficients in the empirical equation.[24]

In the situation, when there is no definitive cutoff of the electronic properties relating them to the empirical definition of the molecular volume and even the basic molecular property of size is of necessarily approximative nature, an attempt to do more explorations in this area was undertaken in order to make certain refinement of knowledge on the acceptable EDC (Electronic Density Cutoff) intervals. The following strategy was applied to estimate the values of EDC. For each of the molecules from the data set a set of volumes enclosed by the electronic density isosurfaces for varying values of ED was calculated. An assumption used for further statistical study of the data is that at a certain "cutoff" value of ED, common for any molecule from the data set, the empirically defined molecular volume (either as probe-defined molecular volume, i.e. Connolly volume, or EOS-derived molecular volume) will be approximately equal to the ED-defined volume.

Connolly solvent-excluded volume, referred to above[14] as one of the definitions of molecular volume, illustrates another important basis for dichotomy of the definitions is the apparently "intrinsic" metric properties vs the ones defined as directly dependable on the molecule's environment, i.e. solvent. In the latter case, the medium surrounding the molecule is taken into account the way that the space not accessible by the solvent is allocated to the molecule. The truth is that molecules can never be considered as isolated entities in processes of molecular interactions such as for example collision or protein inhibition. In this regard an "intrinsic" molecular size (volume, etc.) is always an approximation in an attempt to describe molecules. On the other hand, using a predefined probe is also associated with rough estimations because the probe is assumed to be a spherical body, and, in the cases where the shape of molecules cannot be disregarded, the whole notion of "probe" loses its original sense.

Another molecular parameter, area(s) of projection, received much less attention as a molecular descriptor. This is probably because of the general understanding of the area as a single univariate parameter. Also, considered this way, areas of projection become not invariant depending on the orientation of the planes to which projections are to be done.[25] Various arbitrary solutions were proposed to overcome this problem, e.g. to use nonorthogonal planes defined by the axes chosen the way that first axis x is defined by the direction of the maximal length of the molecule, the last axis z is defined by the direction of the minimal thickness of the molecule, and the third axis y is approximately orthogonal to the other axes.[26] Yet, a more common approach is to use orthogonal planes of projection. The three areas of projection, in this case, are called "shadow indices". Shadow indices have become quite commonly used in QSAR/QSPR studies[27] and modeling of adsorption,[26,28] and the calculation of these parameters has been included in software packages.[29] The same problem of ED cutoff, as described above for the case

Table 1. Descriptors of Molecular Size: Molecular Radius, Surface-Averaged Radius, Volume-Averaged Radius, Area of Projection, Surface Area, and Volume

dimensionality of the size parameter			
1	2	2	3
$r_F = \frac{1}{4\pi}\int\int_S \frac{\mathbf{n}}{r}\cdot d\mathbf{S}$	$A = \frac{1}{S_{sph}}\cdot\int\int_{S_{sph}} A\cdot dS_{sph} =$	$S = \int\int_S dS$	$V = \int\int\int_{V:D\geq D_0} dV$
$r_S = \frac{1}{S}\int\int_S r\cdot dS$	$= \frac{1}{S}\int\int_S \mathbf{A}\cdot d\mathbf{S}$		
$r_V = \frac{1}{V}\int\int\int_V r\cdot dV$			

of molecular volumes, exists also in the ab initio calculation of areas of projection.

Finally, three linear dimensions give the simplest presentation of the size of a solid.[25,29] To make it invariant to the position of the solid in the system of coordinate, the direction of the maximal span may be chosen as the first axis, and the next axis is orthogonal to the first and again directed in the direction of the maximal span. The third axis is perpendicular to the first two. To reduce the influence of the "outlier" atoms in the molecule, three eigenvalues of either atomic coordinates or momenta of inertia were used in the shape description.[3] These three one-dimensional descriptors as well as the ratios of these measurements are certainly useful for many computational applications. Still, there are an infinite number of solids with different shapes and same combination of these three numbers.

Once the parameters of molecular size (volume V, surface area S, and area of projection A) are defined in a certain way, an additional group of molecular size descriptors may be introduced, that are the radii of spheres with the same values of V, S, and A ($r_{(V)}$, $r_{(S)}$, and $r_{(A)}$).

Although the shape is some characteristic of the object unrelated to its size, the common understanding of it still implies those metric relationships. In fact, even topological methods, when used to describe the shape, invoke metric (in a physical or geometrical sense) relationships.[3,5] Here we present the results of the exploration of nontopological approach.

3. DESCRIPTORS OF MOLECULAR SIZE: ORIENTATION-DEPENDENT AND AVERAGED PARAMETERS

If one attempts to approach the characterization of the shape of a solid from the common point of view, these questions should be answered: how much the solid is different when observed from the different points of view? how different is it from a sphere? is it elongated? is the surface of it smooth or rough?

The measures of linear span of a molecule along the three directions as described in the previous section may give some representation the size and shape of the molecule. Generally, an infinite number of directions of measurement are needed to give a complete description of the shape of a solid. Another approach is to use an expectation value of radius averaged in a certain way. Let r_0 be a center of the molecule also defined in a certain way, for example, as a center of mass assuming a uniform distribution of mass within the boundaries of the molecule; and r_i − a vector from r_0 to point i on the surface of a molecule. If vectors r_i are being chosen randomly and with probability density uniformly distributed along all directions in the space, expectation of $|r|$ is defined probabilistically as average length of of r_i's determined in the infinite number of experiments. This may be written as a limit:[30]

$$r = \lim_{n\to\infty}\frac{1}{n}\sum_i |r_i| \tag{1}$$

In a common notation, r is defined in Table 1 as a surface integral r_F. If the surface of a molecule is divided into n nonintersecting smooth surfaces ΔS_i, and r_i − a vector connecting the center r_0 and point i that belongs to ΔS_i then

$$r_F = \lim_{\alpha_i\to 0}\sum_i r_i\cdot\frac{\Delta\Omega}{4\pi} = \lim_{\alpha_i\to 0}\sum_i r_i\cdot\frac{\Delta S_r}{S_r} = \lim_{\alpha_i\to 0}\sum_i r_i\cdot\frac{\Delta S_r}{4\pi r_i^2} =$$
$$\frac{1}{4\pi}\lim_{\alpha_i\to 0}\sum_i \frac{1}{r_i}\cdot\Delta S_r = \frac{1}{4\pi}\int\int_S \frac{1}{r}\,dS_r = \frac{1}{4\pi}\int\int_S \frac{\mathbf{n}_r}{r}\,d\mathbf{S} \tag{2}$$

where ΔS_r is an area of a projection of ΔS_i on a sphere with radius r_i, $\Delta\Omega_i$ is a solid angle occupied by r_i, α_{max} − a value of the maximal angle enclosed by any of the solid angles $\Delta\Omega_i$, $\mathbf{n}_r$ − a vector function defined at point i as vector $\mathbf{r}_i$ normalized by its length ($r_i^{-1}\cdot\mathbf{r}_i$). Here, the condition of the uniform distribution of the directions of r_i is substituted by a request for α_{max} to approach zero and r_F has a geometrical meaning of the average radius weighted by the solid angles allocated to each radius r_i. Equivalently to the above, the radius r_F could be defined as a flux of a vector function $r^{-1}\cdot\mathbf{n}$ (and this explains subscript F).

Analogously, in the case if random vectors r_i are chosen the way that their ends are distributed uniformly on the surface of the molecule, the molecular radius is defined as a surface integral of a scalar function. The scalar function is equal to the length of radius-vectors, and the integral should be normalized by the value of the surface area S of the molecule (r_S, Table 1).

Next, for the sake of logical completeness, parameter r_V is introduced as the expectation of a molecular radius calculated by averaging the radius-vectors within all the volume of the molecule (Table 1). This parameter is maybe of less importance because, first, unlike the previous two definitions of molecular radius, the geometrical meaning of it is different from an "average" molecular radius. It is indeed an average distance of a point within the boundaries of the

molecule from the center of this molecule. Second, this descriptor will be highly weighted by the parts of the molecule with higher volume.

These definitions of molecular radius assume the distribution of masses on the surface of the molecule or within it is not important for the consideration of the shape of the solid; therefore no weighting on mass distribution was given in the definitions. Center of the molecule is proposed to be calculated with an assumption of uniform distribution of mass to avoid direct dependence of the calculated shape descriptors on the chemical composition of a molecule.

It is not possible to give any preference for a single descriptor r_F, r_S, or r_V before the consideration of the actual data and the task. One can envisage the need for several versions of the definition for the molecular radius depending on what aspect of a molecular size or shape is more important in a particular type of a research problem. For example, the parts of the molecule with higher surface area will contribute more in the value of r_S comparing to r_F. On the other hand, the parts of the molecule with higher volume will contribute more in the value of r_V comparing to r_F. So, the ratios of these parameters could be of certain value in characterizing the roughness of the surface or shape of the molecules (vide infra).

Next an important molecular descriptor dependent on the orientation of a molecule is the molecular area of projection A. If the areas of projections of a molecule on n randomly oriented planes were measured, expectation A is defined as a limit[30]

$$A = \lim_{n\to\infty} \sum_i \frac{1}{n} A_i \tag{3}$$

where A_i is a value of the area of projection on the ith plane. Let this plane be perpendicular to r_i. Similarly to the expectation of a molecular radius, (3) is equivalent to surface integrals (4) and (5) (Table 1):

$$A = \lim_{\alpha_{max}\to 0} \sum_i A_i \cdot \frac{\Delta\Omega_i}{4\pi} = \lim_{\alpha_{max}\to 0} \sum_i A_i \cdot \frac{\Delta S_r}{4\pi r_i^2} = \frac{1}{4\pi} \int\int_S \frac{1}{r^2} \mathbf{A} d\mathbf{S} \tag{4}$$

$$A = \lim_{\alpha_{max}\to 0} \sum_i A_{sph,i} \cdot \frac{\Delta S_{sph,i}}{S_{sph}} = \frac{1}{4\pi} \int\int_{S_{sph}} \mathbf{A}_{sph} \cdot d\mathbf{S}_{sph} = \frac{1}{4\pi} \int\int_{S_{sph}} A_{sph} \cdot dS_{sph} \tag{5}$$

Here, $\mathbf{A}$ is a vector function defined at point i as a vector with the same direction as r_i and with an absolute value equal to A_i; S_{sph} − surface area of an auxiliary sphere of unit radius employed to build a set of directions of projections, it is divided into nonoverlapping pieces with areas $\Delta S_{sph,i}$; $A_{sph,i}$ and $\mathbf{A}_{sph,i}$ in (5) are the areas defined by random unit vectors r_i from the center of the sphere to point i on piece i. Formula (5) allows easier numerical implementation.

To calculate the average values of r and A numerically, sets of values of r_i and A_i should be calculated. Once those become available, they present even more valuable and unique data on a molecule by providing a basis for construction of inherent shape descriptors.

This work is far from being a complete calculational exploration of the descriptors discussed. Here we describe calculation and some statistical analysis of only descriptors derived from molecular volumes and projection areas. Molecular radius derived descriptors will be presented in a subsequent publication.

4. DISTRIBUTION-BASED SHAPE CHARACTERIZATION

The only finite solid for which the distance from the center to any point on the surface is equal to r is a sphere. For all other solids there is generally a deviation from the average value r. Similarly, areas of projection depend on the orientation of the solid for all nonspherical solids.

The distribution of the calculated parameters r and A for a single molecule can be described using standard statistical parameters: variance (or, more conveniently, standard deviation), skewness, kurtosis, and possibly other, higher central moments of distribution.[30] Ideally, the more moments are employed for the description of the distribution, the better and more complete description of the original distribution can be obtained. In practice, kurtosis was the highest order central moment-derived parameter employed in this study. A general hypothesis underlying our research is that as much as the distribution of the size parameters is specific for a particular shape of a molecule, the parameters of these distributions deliver a concise parametrization of the original shape. The summary of the descriptors is presented in Table 2. Additionally, it includes a nondistribution parameter of "roughness", which is readily derived from the calculated size parameters.

These functions do not provide the possibility to reconstruct the original shape unambiguously (vide infra). Still, they are sufficiently specific to allow discrimination of the great number of cases. In this regard, one may compare them to the spectroscopic study of substances in a way that they provide fingerprint data on the individual compound.

4.1. Roughness. A concept of molecular roughness is not new. Previously, it was measured empirically from SAXS (small-angle X-ray scattering) and gas adsorption studies.[31] In addition to roughness, several other related and overlapping descriptors exist in the literature (ovality, sphericity, compactness, fractality).

Roughness, as used in this paper, is a parameter to characterize both surface and shape of a solid, how much the shape of a solid different from the shape of a solid with the same volume and with minimal surface area or area of projection. Assuming a surface of an ideal sphere to have a roughness equal to unit, roughness could be measured as a ratio of the surface areas of the given solid and a solid of the same volume and "absolutely" smooth surface (i.e. a sphere).[3,32] In this case, the value of roughness is proportional to a square of the ratio of two linear measurements ($\sim$(*length*/*length*)^2). Alternatively, it is simpler to define roughness as ratio of radii of the sphere with the same surface area and the one with the same volume (see ρ_S in Table 2).

In general, one needs to have two parameters of different metrics (usually $\sim length^3$ and $\sim length^2$ or $\sim length^3$ and

Table 2. Descriptors of Molecular Shape[a]

	underlying parameter of molecular size		
	radius	area of projection	surface area
ovality	$O_r = \frac{s_r}{r_x}$	$O_A = \frac{s_A}{A}$	
radius-corrected ovality	$O_r^{(rc)} = \frac{1}{r_x}\frac{s_r}{r_x}$	$O_x^{(rc)} = \frac{1}{r_x}\frac{s_A}{A}$	
roughness	$\rho_r = \frac{r_F}{r_{(V)}}$	$\rho_A = \frac{r_{(A)}}{r_{(V)}}$	$\rho_s = \frac{r_{(S)}}{r_{(V)}}$
	$\rho_{SF} = \frac{r_S}{r_F}$		
	$\rho_{VF} = \frac{r_F}{4/3\, r_V}$		
surface-corrected roughness	$\rho_r^{(sc)} = \frac{1}{r_x^2}\left(\frac{r_x}{r_V} - 1\right)$	$\rho_A^{(sc)} = \frac{1}{r_x^2}\left(\frac{r_{(A)}}{r_{(V)}} - 1\right)$	$\rho_S^{(sc)} = \frac{1}{r_x^2}\left(\frac{r_{(S)}}{r_{(V)}} - 1\right)$
skewness[b]	$\gamma_{1r} = \frac{1}{N}\sum_i \left(\frac{r_i - r}{s_r}\right)^3$	$\gamma_{1A} = \frac{1}{N}\sum_i \left(\frac{A_i - A}{s_A}\right)^3$	
kurtosis[b]	$\gamma_{2r} = \frac{1}{N}\sum_i \left(\frac{r_i - r}{s_r}\right)^4 - 3$	$\gamma_{2A} = \frac{1}{N}\sum_i \left(\frac{A_i - A}{s_A}\right)^4 - 3$	

[a] r_x − either equivalent of the molecular radius r_F, r_S, r_V, $r_{(V)}$, $r_{(S)}$, or $r_{(A)}$ could be employed for the size correction definitions depending on the type of data and the purposes of the molecular data analysis. In this paper, $r_{(V)}$ was used for the calculation of O_A and $r_{(A)}$ was used for the calculation $\rho_A^{(sc)}$. [b] For simplicity, nonweighted versions of definitions of skewness and kurtosis are given here. In practice, calculations of standard deviations s_r and s_A, skewness and kurtosis of those should take into account some weighting coefficients arising from their definitions as surface integrals and from the numerical approximations used for the calculations. See Supporting Information for the full versions of the definitions.

$\sim length^1$, e.g. volume and surface area, or volume and radius) to define one of the generalized measures of roughness. Having calculated, for example, volume and area of projection, another definition of molecular roughness, or the deviation of the molecular shape from the ideal spherical solid, can be introduced, ρ_A

$$\rho_A = \frac{r_{(A)}}{r_{(V)}} \tag{6}$$

where ρ_A − roughness of projection area, or roughness of molecular shape, $r_{(A)}$ − radius of sphere with the same value of area of projection, and $r_{(V)}$ − radius of sphere with the same volume. In this case, the molecular shape is understood as an averaged contour of a molecular shade, and the descriptor itself may be called roughness of shade. The interval for the values of ρ_A is 1 (for ideal sphere) and up (for all other shapes). In principle, one may encounter with a certain correlation between the values of roughness and the degree of elongation of similar types of solids (e.g. cylinders of different lengths). Parameter equivalent to ρ_S[2] was introduced earlier.[3,32] It became available in CambridgeSoft's ChemOffice as a standard property function and was used in several QSAR studies. Although named "ovality"[32] or "degree of sphericity",[3] it does not specifically characterize ovality or elongation of a solid (vide infra) because there are examples of the solids with high roughness and low ovality (e.g. a sphere with many holes or a star-shaped solid).

Fractality dimension was quite commonly used in characterization of proteins, polymers, and many flat surfaces.[33] There are well developed methods for experimental evaluation of fractality; the descriptor was proven to present a valuable representation of the macromolecular structure, but it is doubtful if the same is true for its application for comparatively small molecules because of a particular conceptual disadvantages of this parameter: rough surface may or may not be fractal. In the case of molecular surfaces, fractality is usually constant only at a certain range of probe ("yardstick") sizes. Due to this, molecular surface generally is not fractal.[34]

A more abstract approach was developed to express the roughness as a spectrum obtained from processing the information on the curvature properties (such as lengths of curves between two inflection points) of the 2D contours obtained by parallel slicing of a molecule.[35] In our opinion, a limitation of this method is the overload of numerical information output. On one hand, the roughness spectra produced by the method do not contain much of the initial structural information, so they cannot be used for many docking and alignment calculations. On the other hand, the spectral representation may require further processing to compress the data even more for many other, lower-level types of studies, such as QSAR/QSPR studies or to use them as a filter in data mining and library design.

Finally, if the expectations of radius r_F, r_S, and r_V, are available, their ratios may give some interesting extra information on the shape of a molecule such as, for example, whether the molecular surface is rougher on the periphery of the molecule or in the internal areas (ρ_{SF}); and whether the volume is concentrated mostly near the center of a molecule or on its periphery (ρ_{VF}, a coefficient 4/3 is introduced to make the value of this parameter equal to a unit for a sphere).

4.2. Ovality. Ovality is the parameter to characterize the deviation of a shape of a solid from the spherical shape to form a more oval or elongated shape. The simplest definition of a quantitative measure of it used in the industry is the

difference between the largest and smallest outside diameter on a cross-section of the solid given in percents to the nominal value of a diameter.[36] If many different diameters of the same solid are considered, the equivalent parameter will be a standard deviation divided by the mean value (i.e. coefficient of variation). Defined this way are descriptors O_r and O_A (Table 2). Depending on the character of molecular interaction either descriptor may be of greater convenience and applicability. For example, if molecules are considered as moving and colliding rigid bodies with negligent intermolecular attraction forces, then projection area derived descriptors will make sense to use. Otherwise, if a molecule is able or has enough time to fit and align itself in the field created by another molecule (e.g. an inhibitor in the vicinity of the receptor's active site), then the O_r parameter will be of better use.

Asphericity is another descriptor used for characterization of the same molecular property

$$A_{3d} = \frac{1}{2}\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_3)^2 + (\lambda_2 - \lambda_3)^2}{(\lambda_1 + \lambda_2 + \lambda_3)^2} \quad (7)$$

where A_{3d} is asphericity of a three-dimensional solid, λ_1-λ_3 − eigenvalues of radius of gyration tensor. Interestingly, ovality is in fact a certain generalization of asphericity. For the case when ovality (let us call it here O_λ) is calculated on the basis of the above three eigenvalues (which are in turn proportional to the squares of molecular radii r_i measured in the directions of three mass eigenvectors, or principal moments of inertia, $i = 1,2,3$), the following is true:

$$A_{3d} = \frac{1}{2} O_\lambda^2 \quad (8)$$

To avoid possible confusion, the term *asphericity* is not used for the ovality descriptors defined in Table 2.

4.3. Size-Corrected Descriptors. If the goal is to produce purely size-invariant molecular descriptors, it cannot be achieved using the above descriptors. Upon the calculation of values of roughness and ovality in the course of this study an existence of a significant correlation of these parameters and the pure size descriptors (A, V, $r_{(A)}$, $r_{(V)}$) was found. It appears that the correlation is only a seeming phenomenon. A feature of molecular shapes is that there is a greater possibility for the variation of those shapes for the molecules of bigger size. Unlike a piece of clay, a molecule of a limited size is limited in its possible maximal length, for example, and, therefore, the molecule is also limited in the number of possible shapes it can make. In the extreme case, monatomic molecules may have only one shape − spherical (or nearly spherical). This makes the range of possible values of ovality and roughness dependable on the overall size of the molecule. For purely speculative purposes, the variation by size can be removed artificially. To avoid dependency on the data set used, the variation was not eliminated completely using statistical methods but a simple division by proper size parameters was done. Thus, radius-corrected ovality was introduced as

$$O_r^{(rc)} = O_r / r_x \quad (9)$$

where r_x is a parameter equivalent to the molecular radius, that is either r_F, r_S, r_V, $r_{(V)}$, $r_{(S)}$, or $r_{(A)}$. Analogously, descriptor $O_A^{(rc)}$ is produced (see Table 2).

For the case of roughness descriptors, the dependence of its upper limit on the surface area of the molecule exists due to the same reasons. Additionally, an a priori intercept of the upper limit of roughness values for molecules is 1 (see above). Taking this into account, surface corrected roughness is introduced here as

$$\rho_r^{(sc)} = (\rho_r - 1)/r_x^2 \quad (10)$$

Descriptors $\rho_A^{(sc)}$ and $\rho_S^{(sc)}$ are defined analogously. In general, a removal of size dependence may not be always something needed in the diversity analysis. For example, depending on the purposes of the analysis, two long cylinders of the same diameter and different lengths may or may not be considered as having two different shapes. If size-corrected descriptors are used, these cylinders will be identified as the same shapes.

4.4. Skewness and Kurtosis. Distributions of r and A are dependable on the shape of a molecule. Common statistical parameters to characterize and compare distributions are skewness and kurtosis[30] (see Table 2 and Supporting Information for defining formulas). Skewness measures the degree of asymmetry of the distribution, or how one tail of the distribution is higher than the other one. This asymmetry is indicative of certain geometrical features of a molecule. In case of the distributions of areas of projections, a long narrow cylinder, or a rodlike molecule, will have a distribution skewed to smaller values of A ($\gamma_1 < 0$). A disk, since its lower limit of A is higher than that of a rod, will have a distribution squeezed to the right ($\gamma_1 > 0$) (Figure 1).

Kurtosis, on the other hand, measures the degree of flatness of the distribution, how strong is the concentration of the values in the proximity of their mean value. Here again a dependence on the structural features exists. A spherical solid has the same projection area in any direction which causes distribution to become very narrow ($\gamma_2 > 0$). A star-shaped solid will have a much higher degree of variation of the areas of projection ($\gamma_2 < 0$) (Figure 1).

Earlier, momentum statistics parameters were applied for the molecular dynamics problems to characterize the conformational rigidity of a molecule.[38]

5. THEORETICAL CALCULATIONS

Molecular volumes were calculated using Gaussian 98W (Version 5.4) and Gaussian 03W (Version 6.0) programs. Molecular structures were optimized at the Hartree−Fock level of theory with the 6-31++G(d,p) basis set.[39] Orbital data from the scratch-file output were converted to cube files which contained the data on SCF electronic densities. The data are the electronic densities at each point of the grid within a cube surrounding a molecule. The frequency of the points of this grid was requested to be ≥ 250 points per side of the cube. Connolly solvent-excluded volumes were obtained using CS Chem3D Ultra (version 7.0.0, 2001) software by CambidgeSoft (Cambridge, MA) with a probe radius set to zero.

Further calculations were done using the SAS System for Windows (Release 8.02) by SAS Institute, Inc. (Cary, NC)

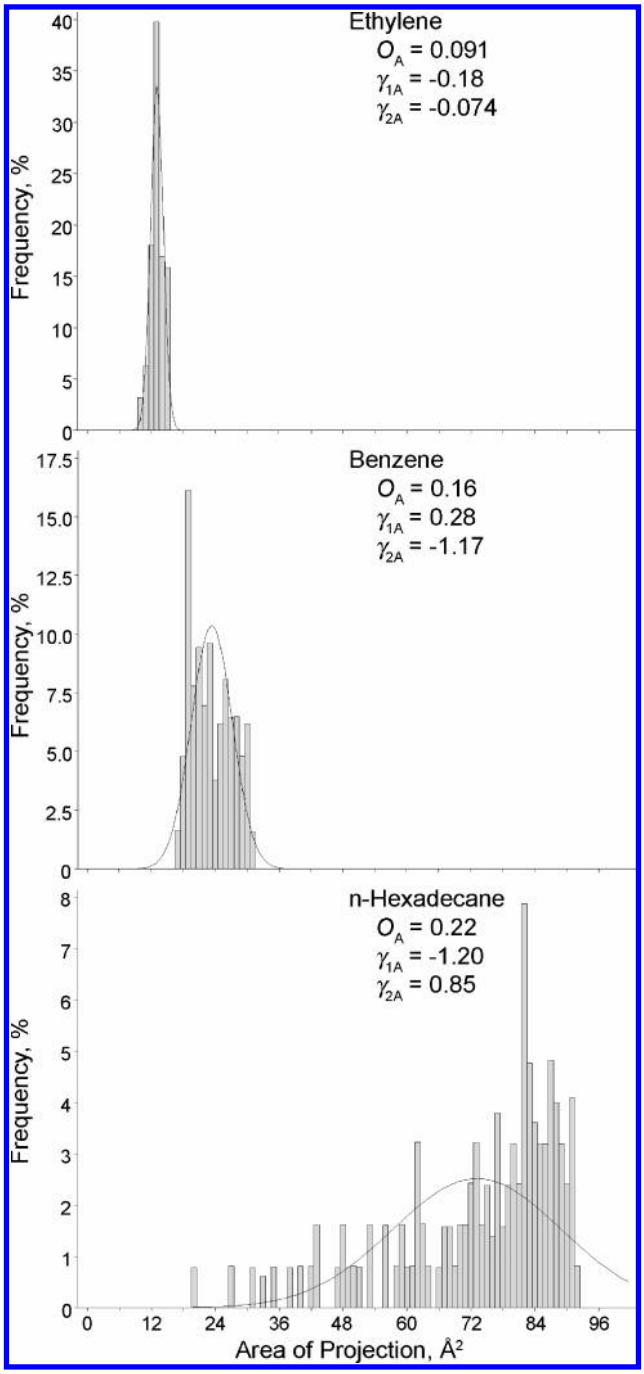

Figure 1. Histograms of distributions of molecular areas of projection along 126 nonparallel directions. Fitted normal curves are overlaid for reference. A molecule of benzene displays positive skewness (γ_{1A}) and negative kurtosis (γ_{2A}) of molecular areas of projection; hexadecane has negative skewness and positive kurtosis. Molecules of ethylene reveal the distribution of areas most close to normal among the three molecules and the lowest value of ovality O_A (geometrically equivalent to the relative horizontal spread of values on the histograms).

as well as standard software packages by Microsoft Corporation. Confidence limits calculated with $\alpha = 0.05$.

5.1. Data Set. A set of 50 molecules from an earlier report[23] was used in this study.[40] The advantage of the set is that it encompasses molecules of diverse shapes: spherical, rodlike, starlike, disklike, and others. The same publication[23] also provided the data on the ES-derived molecular radii. The initial set of 57 molecules was reduced by removing 7 molecules (one salt and the molecules containing 5- and higher period elemements) in order to maintain a uniform procedure of the quantum-mechanical calculations.

5.2. Calculation of Molecular Volumes and Areas of Projection. Data on the electronic densities at the points of a 3D grid enclosing a molecule could be produced using Gaussian software (*cubegen* utility) in a form of so-called cube file. Following the usage given in the above software, these data will be further called *cube*, although the underlying grid is usually rectangular. From this output, molecular volumes and areas of projection were calculated.

Volume was calculated as a product

$$V = V_{\text{cell}} N\,(D \geq D_{\text{cutoff}}) \tag{11}$$

where V_{cell} − volume of one cell forming the grid of a cube; $N(D \geq D_{\text{cutoff}})$ − number of cells of a cube with electronic density values more or equal to a given value of electronic density D_{cutoff}. Formula (11) is an approximation and its accuracy depends on the number of points of a grid per side of a grid (length of each side is chosen by the software). A density of 250 points per side was requested for the generation of cubes for all compounds studied. The software adjusts the actual number of points per side, but the total number of points in the grid remained about the same ($\sim 1.6 \cdot 10^7$ points per cube).

The expected area of projection of the molecule was calculated using the approximation based on (5):

$$A = \sum_i A_{\text{cell},i} N_i\,(D \geq D_{\text{cutoff}}) \frac{\Delta S_{\text{sph},i}}{S_{\text{sph}}} \tag{12}$$

Projection of the original 3D grid along one particular direction i produces 2D grid. Linear dimensions of cells in a newly formed grid are smaller than the dimensions of the original 3D cells. Areas of new cells are $A_{\text{cell},i}$ and the total area along the ith direction of projection is equal to the product $A_{\text{cell},i} \cdot N_i(D \geq D_{\text{cutoff}})$. The method of choosing projection directions i and weighing coefficients $\Delta S_{\text{sph},i}/S_{\text{sph}}$ is based on the tessellation of an auxiliary unit sphere using spherical projection of 252-hedron.[41] Example projections of an electronic density of a molecule of isopentane are presented in Figure 2.

6. RESULTS

6.1. Estimation of Electronic Density Cutoff. It is a common approach to compare the values of size obtained using different methods in order to verify and confirm the validity of those methods.[13] This approach was also used to find an appropriate EDC value. As it was pointed out above, values in the range $0.001 \div 0.005$ e·Bohr^{-3} were used previously to define ED contours in the calculations of size descriptors ab initio.[16−20] Some earlier findings are summarized in Table 3.

The rationale employed here to estimate EDC values follows. In general, one cannot expect the exact equality between ED-derived volumes (V_{ED}), for example, and the volumes obtained using some empirical method (V_{exper}). As an approximation and with no other available method to define the absolute values of molecular volumes, we just compare V_{ED} and V_{exper} at varying values of ED cutoffs and

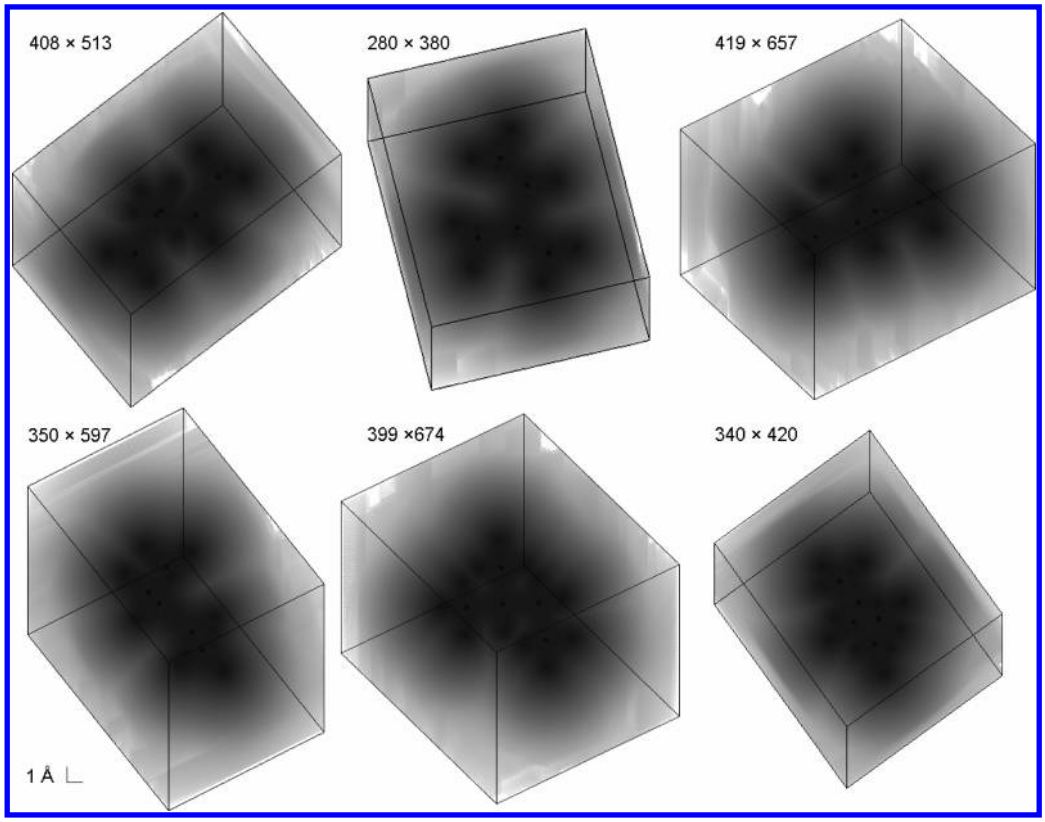

Figure 2. Example projections of electronic density values of isopentane. All outer planes of the original grid with values of ED < 10^{-4} e·Bohr^{-3} were removed. Numbers show the actual pixel sizes of the projections. Intensities of gray are proportional to the logarithm of ED.

Table 3. Electronic Density Cutoff Values Used To Define the Isodensity Contour in ab Initio Calculations of the Molecular Size Parameters and the Criteria Applied to Determine Those Values

EDC value, e·Bohr^{-3}	criteria	ref
0.002	Comparison of one-electron density distributions for first row homonuclear diatomic molecules with molecular sizes derived from Lennard-Jones potential expressions and with van der Waals radii obtained from crystallographic data.	18
0.001	Comparison with the size derived from fitting second virial coefficient and viscosity data with Lennard-Jones (6−12) potential for methane, inert gases, and nonpolar molecules. EDC for polar molecules is said to have higher value than 0.002 e·Bohr^{-3}. 6-31G(d,p) basis set was used for calculations.	16
0.0024 ÷ 0.0046	Comparison of radii of "invariant ED cores" of carbon atom with radii obtained using all major empirical methods. 6-31G(d,p) basis set was used for calculations.	19
0.0034 ÷ 0.0073	Same as above except the invariant core of hydrogen atom was used for comparison.	19
0.002	Defined as a contour enclosing 98% of ED and from comparison with van der Waals radii from ref 22. The comparison was done on the basis of 13 samples; ED-defined radii in 10 samples were higher than van der Waals radii by at least one SD meaning the possibility that EDC > 0.002 e · Bohr^{-3} for those samples. Other 3 samples had the values of radii equal within one SD. 3-21G basis set was used for calculations.	20

try to find the ED value at which equation $V_{ED} = V_{exper}$ holds most accurately. Values of V_{exper} are available from several different methods.[24] Volumes defined by two methods were used here. First, these values are available from the parameters derived from the equation of state (ES). Particularly, here we used effective molecular volumes obtained from the perturbed hard-sphere equation of state (in the form of the Carnahan−Starling−van der Waals equation).[23] Second, van der Waals atomic radii obtained from crystallographic data are commonly used to calculate molecular volumes. Although both methods involve a certain load of theoretical assumptions and calculations, the basic source of the data is still of an empirical nature. Connolly solvent-excluded volumes (V_{vdW})[14] were calculated with a probe radius set to 0.0 Å.

The linear relationship between V_{ED} and V_{exper} was observed in practically all diapason of EDs used. A simple linear regression for the model

$$V_{exper} = a_0 + a_1 \cdot V_{ED} \tag{13}$$

was performed for each value of ED. The values of a_1 and a_0, their upper and lower confidence limits (LCL, UCL, with $\alpha = 0.05$), root-mean-squared error (RMSE), average relative error of prediction (RE, in percents), and r^2 were calculated. Ideally, intercept a_0 is equal to zero, and coefficient a_1 is equal to unit. Additionally, errors of the prediction (e.g. root-mean-squared error) should be equal to zero and r^2 is equal to unit. Particular conditions used to deduce the ED cutoffs

Table 4. Estimation of ED Cutoff Values from Linear Regression of Calculated and Empirical Molecular Volumes and Areas of Projection[a]

condition	V_{ES}	V_{vdW}[b]	A_{ES}[c]
LCL(a_1) < 1 and UCL(a_1) > 1	0.0024 ÷ 0.0026	0.0049 ÷ 0.0052	0.010 ÷ 0.015
LCL(a_0) < 0 and UCL(a_0) > 0	*d*	0.0040 ÷ 0.25	0.0013 ÷ 0.0031
r^2 = max(r^2)	0.0025	0.001	0.002
RMSE = min(RMSE)	0.0025	0.001	0.002
RE = min(RE)	0.006	0.0005	0.0024

[a] Empirical molecular parameters name the columns. [b] Three data points (inert gases Ar, Ne, Kr) were removed as outliers. [c] Ten data points with highest *A* were removed as outliers. [d] Inconclusive, $a_0 < 0$ at all range of ED.

values as well as the results of this deduction are presented in Table 4.

The relationships and parameters produced from the equations of state are based on the models of molecular interactions in which the basic physical event is a collision of molecules with each other. This suggests that the intrinsic geometrical parameter of molecular size embedded in the models and equations is the molecular area of projection not the molecular volume. Based on this assumption, the linear regressions of calculated areas of projection (A_{ED}) and the areas calculated from an equation of state ($A_{vdW} = \pi\sigma^2/4$) were performed (Table 4).

Only a rough estimation of EDC may be obtained from the results in Table 4. Still, it confirms the previously used values at least to the order of magnitude and this is quite satisfactory considering the exponential character of the dependence of electronic densities on the distance from the centers of atoms. It also extends the applicability of these values on the multiatomic molecules. There is a certain tendency to increase the reliable range of EDC from earlier used 0.001 ÷ 0.002 e·Bohr^{-3} to up to ~0.005 e·Bohr^{-3}. The discrepancies between the three methods are quite explainable. Molecular volumes derived from equations of state are bigger then those derived from van der Waals atomic radii due to the fact that when molecules interact with each other the volume of internal cavities and wrinkles on the surface of the molecules becomes excluded which increases the values of V_{vdW}. Another possible factor is the influence of the molecular rotations which may increase the effective volume of the molecules if derived from equations of state. As an overall result of these factors, the value of EDC produced from the V_{ES} regression is smaller than those produced from V_{vdW}. Regression of areas of projection calculated from equation of state-derived parameters, on the other hand, produces bigger values of EDC than the other methods because ab initio calculation of areas presented here does not take into account any molecular conformation other than staggered straight-chain conformation. The existence of multiple other conformations may cause the areas of projection only to decrease. The result of not taking into account the existence of conformations is also seen on the graph of A_{ED} vs A_{ES}. For the molecules of the higher size, experimentally determined areas are becoming lower than a linear dependence would suggest. The most pronounced deviation of this character is observed for normal long chain hydrocarbons (C_9, C_{12}, C_{16}; the lower points in each of the three pairs of points in the upper right part of the graph in Figure 5).

Trying to find an exact EDC value, one should keep in mind obvious limitations associated with this. As it was pointed out in one of the early studies, "there is no physical reason for associating size with a particular contour since various physical characteristics might recommend different contours".[18] In addition to the specific sources of deviations listed above, there are intrinsic differences between the size parameters obtained using different means. The fundamental

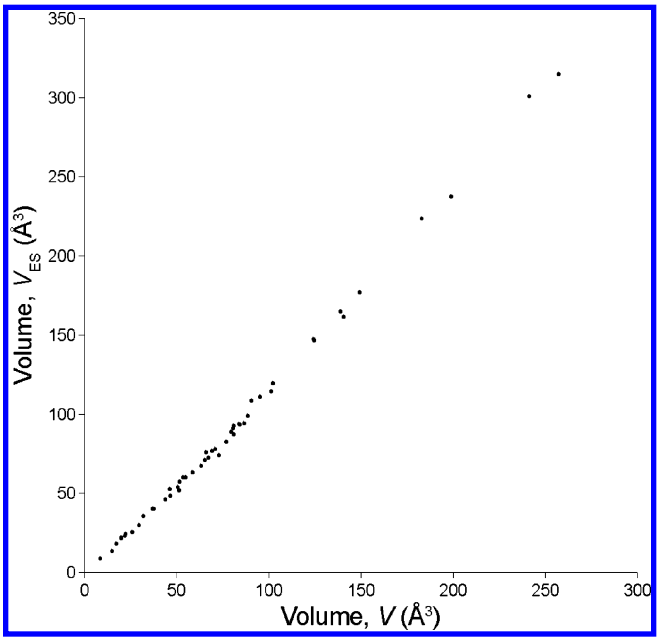

Figure 3. Plot of ED molecular volumes *V* versus molecular volumes obtained from the equation of state V_{ES}.

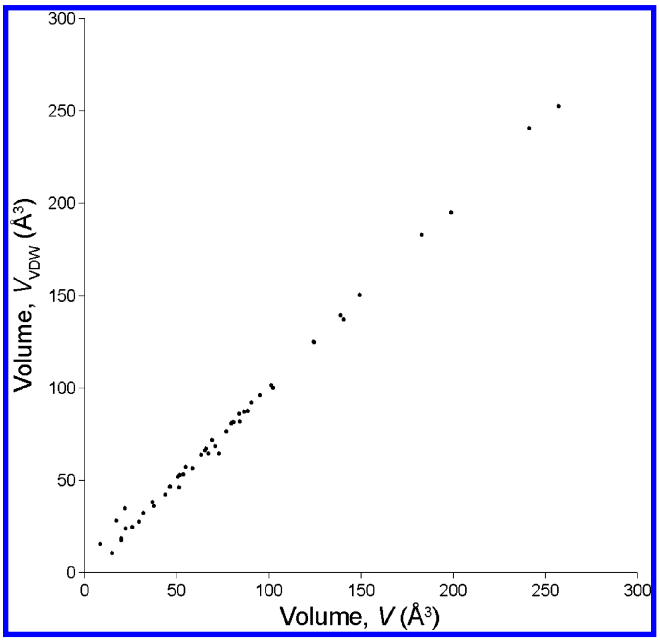

Figure 4. Plot of ED molecular volumes *V* versus molecular volumes V_{VDW} calculated from van der Waals atomic radii using Connolly algorithm.

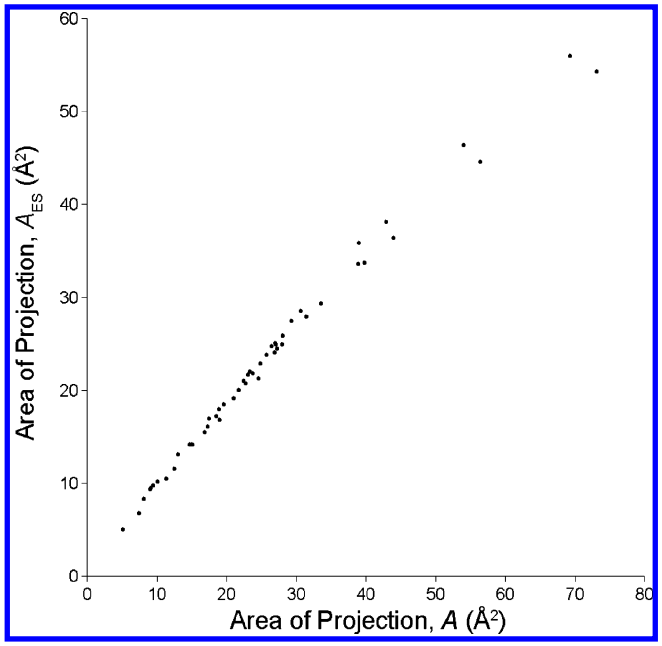

Figure 5. Plot of ED molecular areas of projection A versus molecular areas of projection obtained from the equation of state A_{ES}.

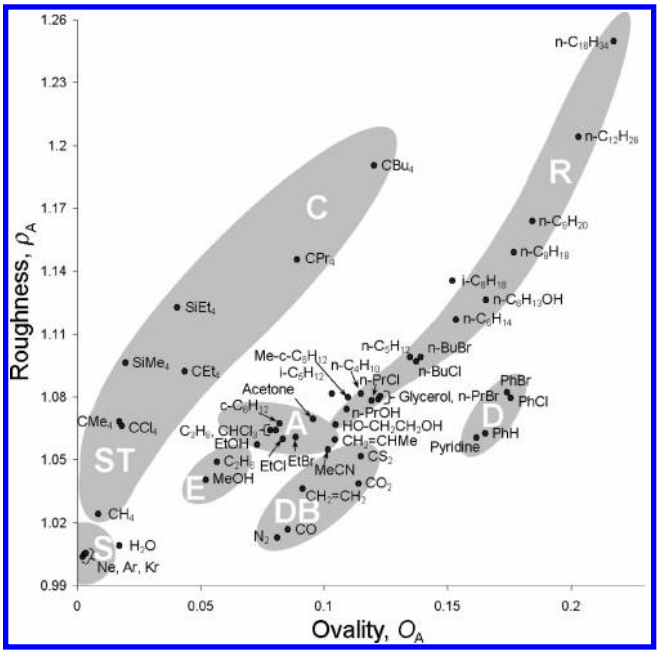

Figure 6. Plot of ovality O_A versus roughness ρ_A. Shapes of the molecules assigned subjectively (A − anglelike, C − crosslike, D − disklike, DB − dumbbell-like, E − egglike, R − rodlike, S − spherical, ST − starlike). A strong correlation of these descriptors between each other as well as with the size of the molecules exist here. Still, certain trends in the distribution of the points depending on the shape of the compounds are noticeable.

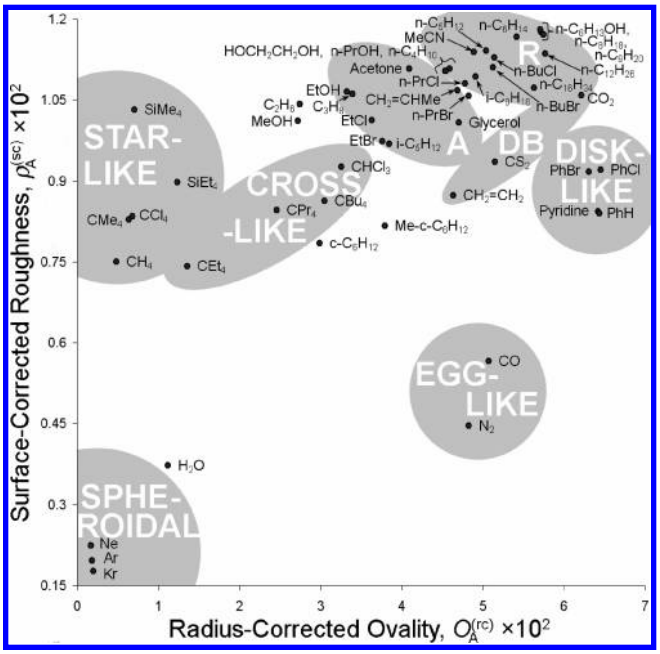

Figure 7. Plot of size-corrected ovality $O_A^{(sc)}$ and roughness $\rho_A^{(rc)}$. Shapes of the molecules assigned subjectively (A − anglelike, DB − dumbbell-like, R − rodlike).

difference is of a purely geometrical nature: ED-derived geometrical shape defining V_{ED} is not the same as a shape built from overlapping van der Waals atomic spheres; EOS effective molecular volumes are obtained in a way in which all size and shape characteristics as well as other structural and electronic characteristics of a molecule are integrated into one size parameter and, because of this, the equality $V_{EOS} = V_{ED}$ may be considered only as an approximation. Next, molecular volumes are always a function of the molecular interactions. How elastic molecules are or how sticky they are during those interactions depends on the whole electronic structure not only on the EDC value of a contour chosen. In this aspect, the number of empirical methods providing the data on the molecular size may be not of a great importance, although it is advantageous that two completely independent methods were used to determine EDC values.

6.2. Classification of Shapes. Electronic density cutoff values may vary in a substantially broad range without affecting the shape properties of a molecule.[5] This implies that the "exact" value of EDC is not required for the shape characterization, and this observation is especially valuable considering the discussion presented in the previous section. An EDC value of 0.005 e·Bohr^{-3} was chosen for the calculation of the shape descriptors O_A and ρ_A, corresponding corrected descriptors, and higher order momenta-derived parameters γ_{1A} and γ_{2A}.

Having calculated the molecular descriptors, it became interesting to perform some studies on how well the variations in molecular shape relate to the variations in the above descriptors. This article is limited to only exploratory analysis of these relations.

Just a visual observation of the 2D-plots of the data on a 50-molecule data set reveals certain tendencies in the distribution of the data points (Figures 6−9). Generally, there is certain continuity between the locations of the data points and the structure of the molecules they represent. In other words, the small changes in the structure of the molecule cause the small change in the location of the proper point on the graph. As a result, in general, a separation of molecules with particular structural features occurs. For example, atoms of inert gases are collocated, so are rodlike (at least in the conformation used for the computations) molecules of normal hydrocarbons and, separately, disklike aromatic compounds. Star-shaped structures (AB_4, B is of comparatively small size) are also separated and located someplace near spherical molecules. As the length of B increases, because of the conformation used for the calcula-

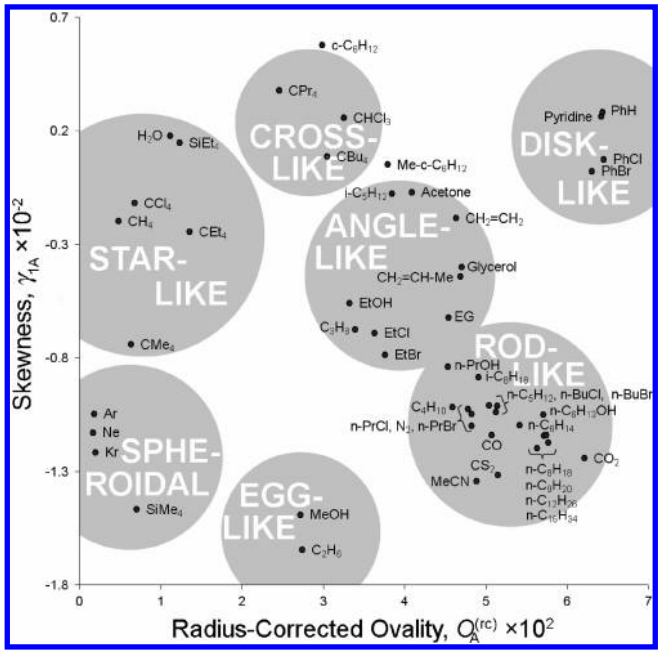

Figure 8. Plot of size-corrected ovality $O_A^{(sc)}$ versus skewness ρ_{1A}. Shapes of the molecules assigned subjectively.

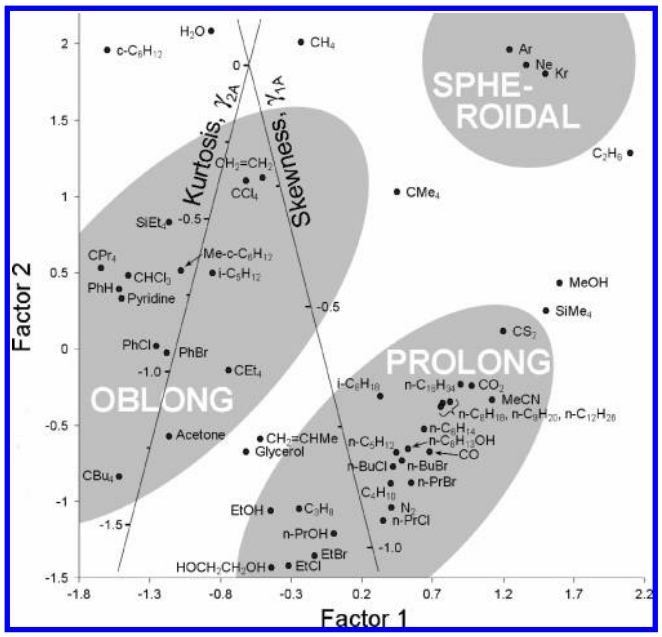

Figure 9. Graph of scores of two factors derived from principal component analysis (PCA) of two descriptors, skewness γ_{1A} and kurtosis γ_{2A}. PCA procedure in this case is equivalent to just a linear transformation of the system of coordinates to produce a better illustration of the locations of the points relative to each other. The original coordinate axes of γ_{1A} and γ_{2A} are also presented. (Coordinates of a point in the original system of coordinates could be reconstructed by drawing lines from that point to each axis and parallel to the other axis). Shapes of the molecules assigned subjectively. The trends in the localizations of the points are different from those seen on the plots of other descriptors.

tions of these molecules, the shapes of AB_4 become similar to a cross. This often shifts their positions toward the area of disklike molecules. The particular locations of molecules of different structures on the graphs are dependent on the descriptors used.

Skewness and kurtosis, γ_{1A} and γ_{2A}, show comparatively high negative correlation (−0.858 for standardized data). The relation of the values of the descriptors and the structural features of the compounds studied is not obvious from the first glance when these two descriptors are plotted against each other. The picture becomes clearer if a simple linear transformation is used to modify the original system of coordinate. Upon the application of principal component analysis the graph of the same data in the new system of coordinate reveals a certain discrimination of the shape features (Figure 9).

The closeness of the points and the clusters of points to each other may be analyzed more quantitatively and implementing several variables if hierarchical cluster analysis (HCA) is applied. This method invokes various definitions of distance and methods of clustering. Here, Ward's clustering method was employed, as it became quite a standard in the literature. To have a greater control over the variances introduced by correlated variables and to have a possibility to remove the correlations with the parameters of the molecular size, factor analysis was performed before HCA. Principal component analysis was used for finding the factors, then all or only the major factors were retained for further processing. Extracted orthogonal factors were used either unrotated, or rotated orthogonally (quartimax or varimax rotation), or nonorthogonally (quartimax or varimax prerotation followed by promax oblique rotation). Factor scores from each factor analysis were used for HCA either altogether, or only the scores of the factors with the smallest correlation with size-characterizing variables ($r_{(A)}$, $r_{(V)}$, A, V) were selected for HCA.

The graphical output of the hierarchical cluster analysis is a dendrogram representing similarities and closeness of the molecules in regard to the properties described by the variables used. Typical HCA trees are presented in Figures 10−12. Generally, the character of clustering has similar trends as those found from the observation of 2D plots of descriptors. The dendrogram in Figure 11 is obtained from the descriptors with low correlation with descriptors of the molecular size. As it appears from the tree, molecules CO_2 and *n*-hexadecane are recognized as similar objects and become joined into a cluster at early stage of clustering procedure.

It is interesting to examine the similar dendrograms built using the descriptors correlated with size. In general, addition or removal of size-dependent variables causes expected and explainable changes in clustering. Those changes indirectly confirm the relevancy of the variables to the molecular shape. When size-dependent descriptors O_A and ρ_A were used (together with parameters γ_{1A} and γ_{2A}, Figures 10 and 12), factor analysis allowed for the separation of the factors correlated and uncorrelated with molecular size. If only the scores on size-independent factors were used for building a dendrogram, diatomic and linear triatomic molecules O_2, N_2, CO, CO_2, and CS_2 clustered together with linear hydrocarbons; $SiMe_4$ showed high similarity with spherical atoms of inert gases (Figure 12). Then all factors were used for cluster analysis, small linear molecules were far from linear hydrocarbons, and tetramethylsilane was much less similar to the molecules of inert gases (Figure 10).

No attempts were made to employ the descriptors in any kind of discriminant analysis and cross-validation of it primarily because the initial assignment of molecules from the training set into a particular shape class seemed to be artificial. At the same time, using abstract geometrical solids

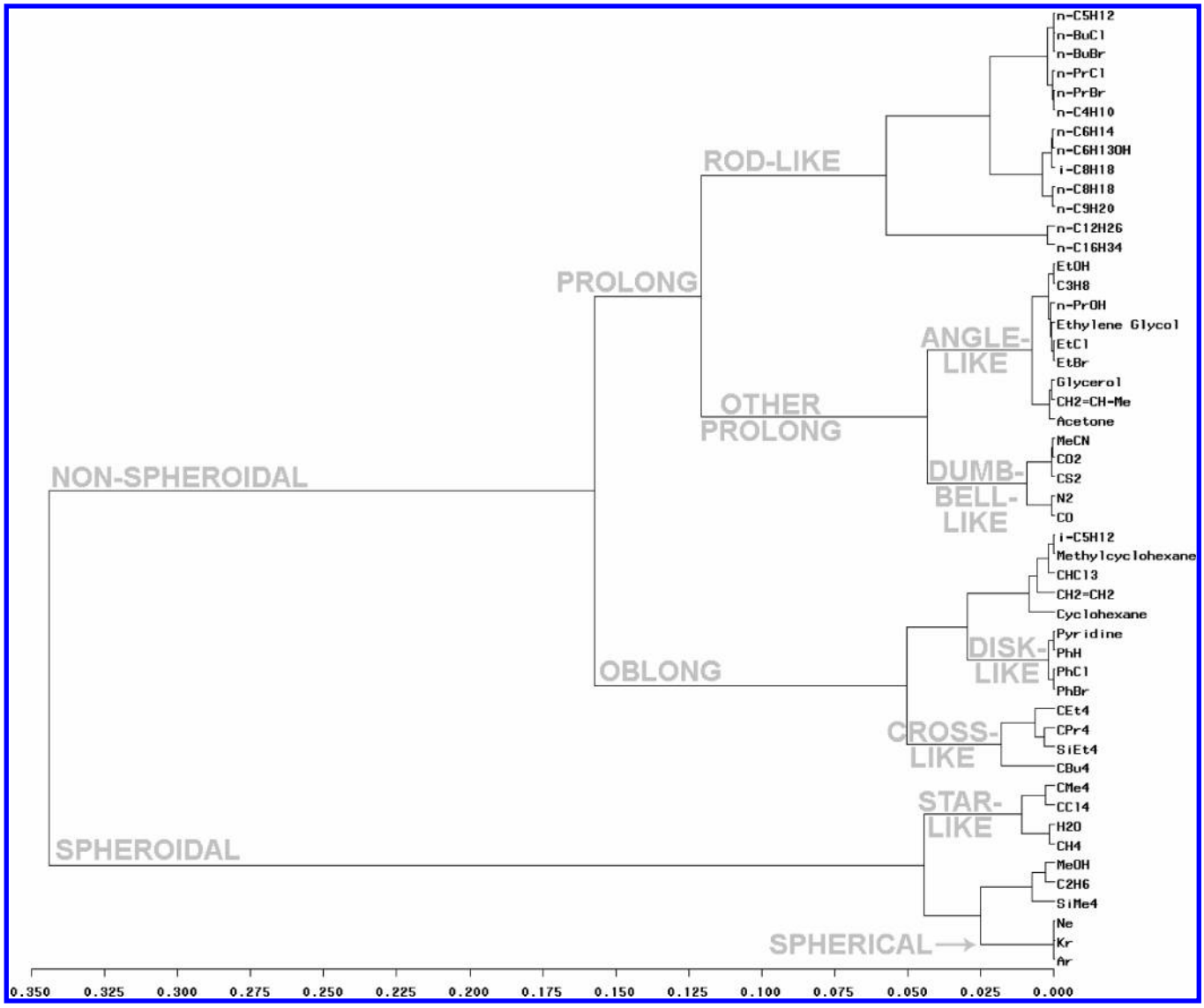

Figure 10. Shape classification dendrogram built using hierarchical cluster analysis. Four descriptors were used for the analysis (O_A, r_A, γ_{1A}, γ_{2A}), factor analysis with varimax prerotation and promax rotation of all 4 factors extracted was performed, the HCA dendrogram was generated from factor scores employing Ward's minimum-variance method. Shapes of the molecules assigned subjectively.

of various shapes and the calculated shape descriptors of those to investigate the classification power of the descriptors would be interesting and may become in the future a topic for a separate study. Apart from the above, the amount of the data (50 molecules) would not allow any rigorous estimation of the shape discriminating power of the descriptors (such as an estimation using leave-one-out procedure or "jackknifing"). On the other hand, statistics did allow us to confirm the fact that the values of the shape parameters and the assignment of molecules to abstract geometrical shape types are dependent. In other words, at least for some shape classes, if a molecule is assigned to a shape class "*X*", its expected value of any of calculated here shape descriptors will be different from the expected value of the molecules assigned to shape class "NOT *X*" (at test probability levels $p = 0.01$). The same statement for the differences in variances in populations "*X*" and "NOT *X*" was proven to be true for the majority of descriptors tried yet at lower levels of significance in some cases (test probability $p = 0.05$). Nonparametric tests for median, variance, and distribution (Wilcoxon, Kruskal-Wallis, Siegel-Tukey, Kolmogorov-Smirnov) were employed for the study (see Supporting Information for details and results).

7. DISCUSSION

From a very general point of view, it may be not completely clear how a calculation of only two geometrical properties, volume and area of projection, can produce at least 4 or 6 different descriptors. In fact, these two properties comprise 127 variables (1 value of volume and 126 values of areas of projection on different directions, as in the version presented in this work). These variables, in turn, are a representation of an infinite number of points, a combination of which makes up a body of a molecule. As soon as at least some information about the shape is conveyed by the several final descriptors, the compression of the initial data becomes quite impressive (yet, of course, this compression is not lossless).

For convenience, all shape descriptors listed in Table 2 will be abbreviated further as ROKS descriptors (Roughness, Ovality, Kurtosis, and Skewness descriptors).

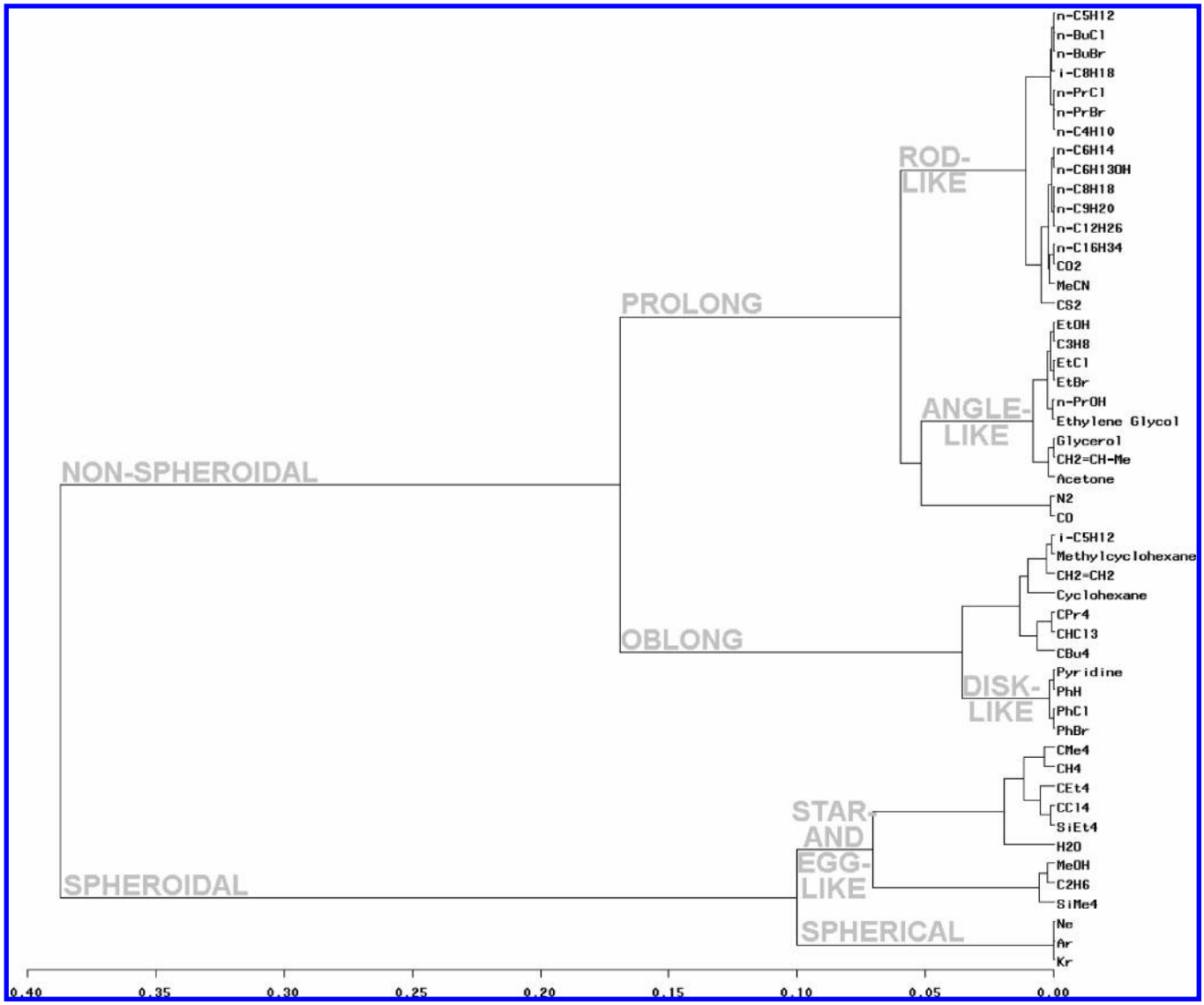

Figure 11. Shape classification dendrogram built using hierarchical cluster analysis. Four descriptors with low correlation with molecular size were used for the analysis ($O_A^{(rc)}$, $\rho_A^{(sc)}$, γ_{1A}, γ_{2A}), factor analysis with varimax prerotation and promax rotation of all 4 factors extracted was performed, and the HCA dendrogram was generated from factor scores employing Ward's minimum-variance method. Shapes of the molecules assigned subjectively.

Each of the methods mentioned above has distinct advantages when comparing them with each other. Often, ROKS descriptors do not have as much descriptive power as many other types of descriptors and methods, but the descriptors used in this paper may eliminate certain problems related to those other methods. First, it may be considered as an advantage, that the ROKS descriptors can deliver a concise and concentrated description of a molecular shape in a form of a limited number of numerical values. In this regard, SGM[5] employs higher-dimension[42] and more voluminous data to describe shapes. Some levels of the data appear not to be relevant to the actual process of molecular interaction or at least not to be revealed in molecular interactions of interest, e.g. in the biological systems. For example, generally not all density intervals are important to describe the molecular shape because high-electronic density regions of a molecule (e.g. a region deeper than the threshold for "single density domain" with ED of less than $\sim 0.2 \div 0.4$ e·Bohr^{-3}) are regularly not penetrated or touched by other interacting molecules. These high-density regions are, of course, important for the description of the structure of the molecule because they show the positions of nuclei and define many molecular properties, but geometrically these high-density regions do not define the molecular shape.

Next, index-based methods[6−9] often denominate or describe the shape by a discrete number, when it would be more natural to expect the use of continuous scales for the descriptors of shape. An issue that is presently unclear is the relationship of the shape descriptors listed here and well-known topological molecular connectivity indices. Initially introduced by Wiener and now existing in many modifications, molecular connectivity indices, such as Hosoya's parameter *W(G)*, a half of a sum of the distance matrix of a molecular graph, are able to characterize molecular complexity and reveal certain molecular features (e.g. sphericity, rings, or long chains). Some of the topological indices may seem to be too formal in design, yet they are often proved to have predictive and descriptive power, and it would not be unexpected to find a significant correlation of some topological indices of this kind and ROKS descriptors.

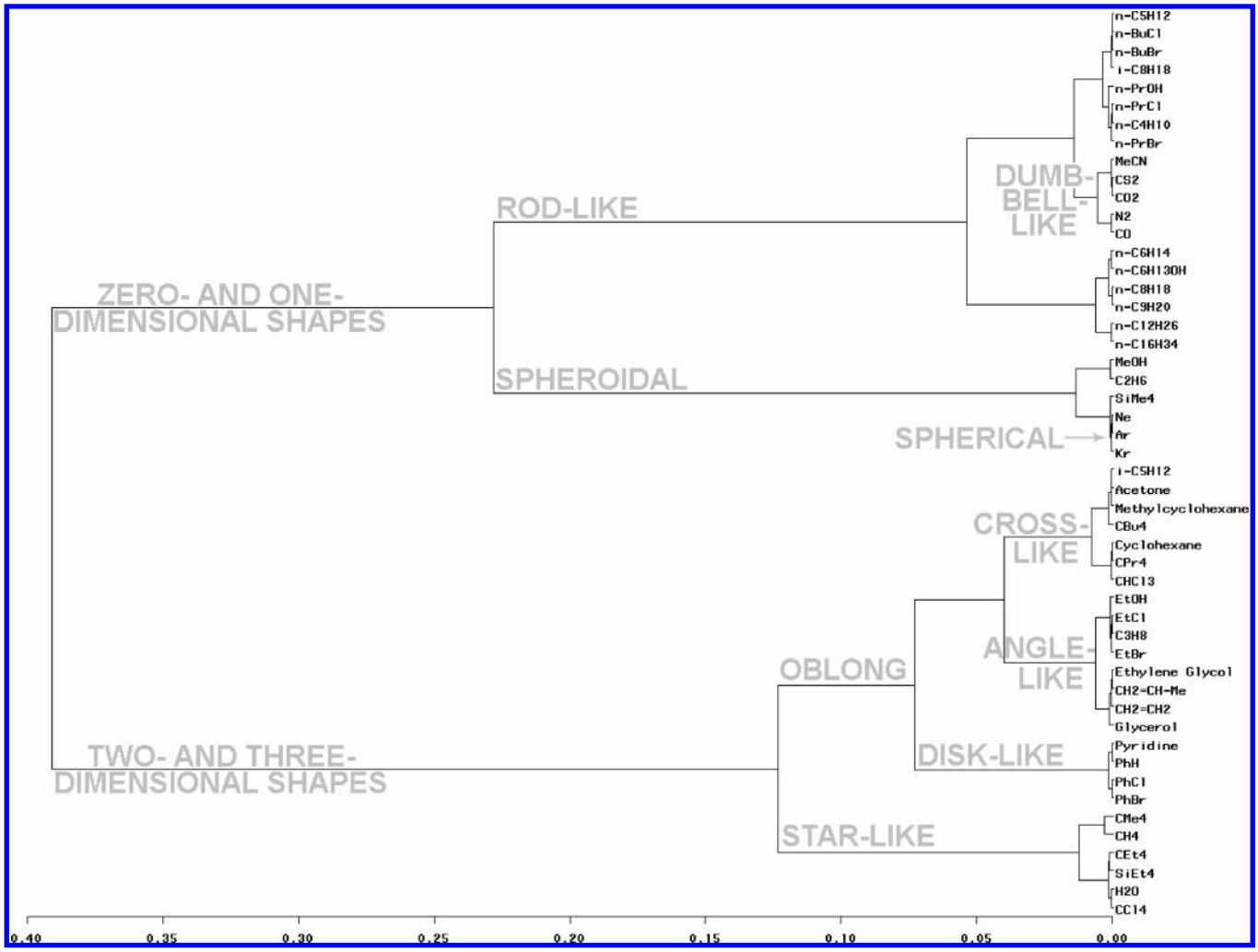

Figure 12. Shape classification dendrogram built using hierarchical cluster analysis. Four descriptors were used for the analysis (O_A, ρ_A, γ_{1A}, γ_{2A}), factor analysis with varimax prerotation and promax rotation of all 4 factors extracted was performed, and the HCA dendrogram was generated from factor scores of 2 factors employing Ward's minimum-variance method. Shapes of the molecules assigned subjectively (notation "0-, 1-, 2-, and 3-dimensional shapes" is from the analogy of the shapes with the shapes of point (zero-dimensional entity), line (one-dimensional entity), etc.). Character of clustering changes depending on factors submitted for HCA.

The fundamental difference of "index" descriptors and ROKS descriptors is the following. Generally speaking, the approach presented here is not to measure some absolute shape parameter or index characteristic to molecules but to measure the deviation of those molecules from a certain standard. For example, roughness is derived from the comparison of surface area or area of projection of a molecule to the same parameter of the ideal sphere with the same volume. Ovality O_A is derived from the deviation of the areas of projection from their mean or, in other words, from the area of projection of a body with an invariant area of projection i.e. a sphere. Next, the distributions of areas of projection are evaluated by the comparison with a normal distribution. A particular standard solid which would have a normally distributed areas of projection is, of course, not defined, but the existence of well defined and generally accepted statistical parameters to measure the deviation from the normal distribution makes it quite logical to use the higher momenta of distribution for the characterization of that distribution.

An advantage of the approach used here comparing with other statistical and distribution-based methods such as, for example, the WHIM method by R. Todeschini[12] is that the need for the additional weighting coefficient information is eliminated. At some point, the shape information is completely abstracted from the atomic composition of a molecule. Additionally, the shape information is not divided into directional components (eigenvectors or principal components). Shape information, if presented by ROKS descriptors, is also detached from any electronic, electrostatic, and other molecular properties.

ROKS descriptors generally do not allow distinguishing solids on the basis of their symmetry.[10,11] For example, solids of the same group of symmetry (i.e. "disks" and "rods") have distant values of the shape parameters, and, vice versa, solids of the different symmetry may have close values of those parameters (i.e. "crosslike" $CHCl_3$ and CBu_4, "disklike" benzene and pyridine) (Figures 3−9).

Finally, as a cautionary note, it is important to remember that different methods which may be close in their goals and approaches are usually suited and limited to their own areas of applicability. For example, a general distinction between finding an initial lead compound and optimizing of its activity exists in the process of drug discovery. First, a structure containing pharmacophore specific in binding to the protein pocket of interest is searched. Further development produces

a more complex inhibitor. Shape properties of the molecular candidates are important at both stages, but distinct methods or modifications of those methods should be employed at these two stages. Particularly in this example, the shape of the pharmacophore must be preserved at the stage of lead optimization.[43]

The question remains, what is the actual distribution of the areas of projection. Although the distributions are compared with the normal distribution, the geometrical sense of projection areas suggests that they are described by some other, logically more suitable for the case distribution. If the positions of atoms in various molecules may be assumed to be distributed normally, the same is not true about the distribution of the values of the areas of projection. In case of areas of projection of a molecule, for each given molecule there is always a nonzero minimum of the area. At the same time, an area of projection is equivalent to an area of the union of all the sections of the molecule perpendicular to the direction of the projection. Since it is a union, it is always either a maximum or a value larger than the maximum of those sections. Considering this, one would suggest to use the Weibull distribution or, maybe more correct, the log-Weibull distribution to describe the distribution of the calculated areas. The former (Weibull[30]) is a random distribution of values with a certain minimal value, such as the breaking strength of materials. The latter (log-Weibull or Extreme Values distribution[30]) is used to describe the distribution of the maximal values of the values distributed as the Weibull distribution. The exact mathematical type of the distributions of the molecular areas of projection was not established in this study since this question is of a more theoretical nature and is beyond the goals of this article.

8. CONCLUSIONS AND PERSPECTIVES

As a result of this study, volumes and areas of projection of 50 molecules were calculated ab initio for contours enclosed by varying values of electronic density. A reasonable interval of ED cutoffs was deduced from the investigation of the relations between empirical and theoretically calculated molecular parameters (A and V). A novel set of distribution-based molecular shape descriptors was proposed. The fundamental basis for building these descriptors is both generalized molecular radius and areas of projection. Descriptors derived from the areas of projection were calculated in this work. Examination of the values of the calculated variables in relation to the geometrical shapes of the molecules used in the study reveals apparent correlation of the descriptors and certain generic shapes represented by the molecules. Further studies will include analysis of descriptors derived from molecular radii r_F, r_V, and r_S as defined above. From the general considerations, the basic properties and the descriptive power of both radii- and area-based parameters are expected to be similar to each other.

As it was pointed out above the descriptors used in this publication do not provide a complete description of the shape and do not allow an exact reconstruction of the molecular shape from the values given. This stems from the fact that only several central moments of distribution are to be taken into account. There are other obvious limitations such as the inability of the approach to make a distinction between enantiomers (yet, diastereomers, in general, could be distinguished by the shape parameters). The calculations described in this article do not provide any account for the existence of different molecular conformations. Still, the possibility for the conformational analysis is built-in into the method and may well be realized in the future. Finally, there is a general limitation arising from the novelty of the method. More data and more examples of practical applications of the approach are needed to establish its validity. Although the statistical analysis cannot provide an estimation of discriminating power and errors of discrimination at this point, the fact that those shape parameters are derived from the intrinsic shape descriptors provides certain "content-validity" of the descriptors. Regarding the advantages of using ROKS descriptors, they are already evident: the descriptors provide tremendous reduction of the molecular structure data, they are invariant to orientation and translation of a molecule, and, as it has been noted, they are fundamentally related to the shape properties of a molecule. Additionally, ROKS descriptors could be calculated from the structural data obtained either ab initio, semiempirically, or from the models of the molecules built of van der Waals atomic spheres. The possibility to utilize the output of quantum-mechanical calculations fundamentally increases the value of the descriptors and counterpoints them and the descriptors which are intrinsically non-ab initio (e.g. Wiener index).

The descriptors of molecular shape proposed in this article have their own area of practical applicability which has yet to be defined. The most straightforward use of those could be as a filter for a primary selection of the structures for further biological or other screening. In case if some information about the receptor and known inhibitors is available, the initial selection on the basis of similarity to those inhibitors may substantially reduce the pool of compounds to screen and, therefore, reduce the costs associated with the synthesis and screening. This is also true if the screening is done in silico. Calculation of shape parameters, on the other hand, takes certain machine time but once done and stored in the database, the values could be used many times for various biological targets. If the information about the active site of the receptor is not available, the diversity of the candidate pool is something desirable. For example, commercially available libraries are in most cases advertised and sold as a collection of diverse (by the chemical and physical properties listed) compounds. This is another way of cutting costs for pharmaceutical companies: a quantity of multiple copies of "diverse" library or a set of libraries is designed, produced, and purchased once and could be used for varying targets at the initial steps of drug discovery.[1,2] A small set of parameters allowing exclusive characterization of molecular shape would be indispensable in the design of those libraries.[44]

Supporting Information Available: Formulas for weighted standard deviation, skewness, and kurtosis; histograms for distributions of molecular areas of projection for all compounds studied; SAS programs for calculations of molecular volumes and areas of projection; detailed algorithms and instructions for those programs; calculated values of V_{vdW}, V_{ES}, V, A, A_{ES}, ρ_A, $\rho_A^{(rc)}$, γ_{1A}, $\gamma_{1A}^{(sc)}$, $\gamma_{1A}^{(sc)}$, and $\gamma_{1A}^{(sc)}$; and results of nonparametric statistical tests on identification of the molecular shapes. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Borman, S. *Chem Eng. News* **2004**, *82*, 40, 32−40.

(2) (a) Gillet, V. J. In *Computational Medicinal Chemistry for Drug Discovery*; Bultinck, P., De Winter, H., Langenaeker, W., Tollenaere, J. P., Eds.; Marcel Dekker: New York, 2004; pp 617−639. (b) Sauer, W. H. B.; Schwarz, M. K. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987−1003. (c) For the example of commercially available libraries designed to satisfy certain molecular descriptors' profile see the products by ChemBridge (San Diego, CA). Similarly, for in-house libraries, there is a trend to reduce their size without a loss of diversity. Wyeth (Madison, NJ) recently reduced its collection by 3 times, see: Mullin, R. *Chem Eng. News* **2004**, *82*, 30, 23−32.

(3) Arteca, G. A. *Rev. Comput. Chem*. **1996**, *9*, 191−253.

(4) Raevsky, O. A. *Russian Chem. Rev.* **1999**, *68*, 505−524.

(5) (a) Mezey, P. G. *Shape in Chemistry: An Introduction to Molecular shape and Topology*; VCH Publishers Inc.: New York, 1993; 224 p. (b) Mezey, P. G. In *Encyclopedia of Computational Chemistry*; Schleyer, P. V. S., Ed.; Wiley: Chichester, 1998; Vol. 4, pp 2582−2589.

(6) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(7) Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.

(8) Bertz, S. H. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601.

(9) (a) Randic, M. *J. Am. Chem. Soc*. **1975**, *97*, 6609−6615. (b) Hall, L. H.; Kier, L. B. *Rev. Comput. Chem*. **1991**, *2*, 367−422. See also other works of these authors on the molecular indices.

(10) (a) Zabrodsky, H.; Peleg, S.; Avnir, D. *J. Am. Chem. Soc*. **1992**, *114*, 7843−7851. (b) Zabrodsky, H.; Peleg, S.; Avnir, D. *J. Am. Chem. Soc.* **1993**, *115*, 8278−8289.

(11) (a) Zabrodsky, H.; Avnir, D. *J. Am. Chem. Soc.* **1995**, *117*, 462−473. (b) Lipkowitz, K. B.; Gao, D.; Katzenelson, O. *J. Am. Chem. Soc*. **1999**, *121*, 5559−5564.

(12) (a) Todeschini, R.; Lasagni, M.; Marengo, E. *J. Chemometrics* **1994**, *8*, 263−272. (b) Todeschini, R.; Gramatica, P.; Provenzani, R.; Marengo, E. *Chemom. Intell. Lab. Systems* **1995**, *27*, 221−229.

(13) Gavezzotti, A. *J. Am. Chem. Soc*. **1983**, *105*, 5220−5225.

(14) Connolly, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 1118−1124.

(15) (a) Richards, F. M. *J. Mol. Biol.* **1974**, *82*, 1−14. (b) Gerstein, M.; Tsai, J.; Levitt, M. *J. Mol. Biol*. **1995**, *249*, 955−966. (c) Tsai, J.; Taylor, R.; Chothia, C.; Gerstein, M. *J. Mol. Biol.* **1999**, *290*, 253−266.

(16) (a) Bader, R. F. W.; Carroll, M. T.; Cheeseman, J. R.; Chang, C. *J. Am. Chem. Soc*. **1987**, *109*, 7968−7979. (b) Bader, R. W. F.; Preston, H. J. T. *Theor. Chim. Acta* **1970**, *17*, 384.

(17) Wong, M. W.; Wiberg, K. B.; Frisch, M. J. *J. Comput. Chem.* **1995**, *16*, 385−394.

(18) Bader, R. F. W.; Henneker, W. H.; Cade, P. E. *J. Chem. Phys.* **1967**, *46*, 3341.

(19) Arteca, G. A.; Grant, N. D.; Mezey, P. G. *J. Comput. Chem.* **1991**, *12*, 1198−1210.

(20) Francl, M. M.; Hout, R. F., Jr.; Hehre, W. J. *J. Am. Chem. Soc.* **1984**, *106*, 563−570.

(21) A standard quantum-mechanical software, Gaussian 03, uses a preset value of EDC = 0.001 e•Bohr^{-3} when a molecular volume is calculated using keyword *Volume*.

(22) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441−451.

(23) Ben-Amotz, D.; Herschbach, D. R. *J. Phys. Chem.* **1990**, *94*, 1038−1047.

(24) Hirschfelder, J. O.; Curtiss, C. F.; Bird, R. B. *Molecular Theory Of Gases And Liquids*; John Wiley & Sons: New York, 1954; 1219 p.

(25) (a) Rohrbaugh, R. H.; Jurs, P. C. *Anal. Chim. Acta* **1987**, *199*, 99−109. (b) Webster, C. E.; Drago, R. S.; Zerner, M. C. Molecular Dimensions for Adsorptives. *J. Am. Chem. Soc.* **1998**, *120*, 5509−5516.

(26) Meyer, A. Y.; Farin, D.; Avnir, D. *J. Am. Chem. Soc.* **1986**, *108*, 7897−7905.

(27) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan B. T. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 669−677.

(28) Oberg, K.; Edlund, U.; Eliasson, B.; Shchukarev, A.; Seshadri, K.; Allara, D. *J. Phys. Chem. B* **2000**, *104*, 10627−10634.

(29) Software packages Cerius^2 by Accelrys, Inc. (San Diego, CA) and CODESSA by Semichem, Inc. (Kansas City, MO).

(30) Weisstein, E. W. *MathWorld*, http://mathworld.wolfram.com/, Wolfram Research, Inc. (Champaign, IL), 2004.

(31) Matsumoto, T.; Inoue, H.; Chiba J. *J. Appl. Phys*. **1992**, *71*, 1020−1025.

(32) (a) Bodor, N.; Gabanyi, Z.; Wong, C.-K. *J. Am. Chem. Soc*. **1989**, *111*, 3783−3786. (b) CS ChemOffice v. 7.0.0 by CambridgeSoft (Cambridge, MA) 2001.

(33) Vollet, D. R.; Donatti, D. A.; Ibanez Ruiz, A. *Phys. Rev*. **2004**, *69*, 064202-1-6.

(34) A terminological note: molecular roughness (in the sense of having nonminimal surface or projection area at given volume) is not opposite to smoothness (in mathematical meaning), rough molecules are usually smooth (i.e. have smooth surfaces).

(35) Arteca, G. A.; Mezey, P. G. *Theor. Chim. Acta* **1991**, *81*, 79−93.

(36) Nayyar, M. L. Piping Handbook, 7th ed.; McGraw-Hill: New York..., 2000; A.269.

(37) (a) Theodorou, D. N.; Suter, U. W. *Macromolecules* **1985**, *18*, 1206−1214. (b) Rudnick, J.; Gaspari, G. *Science* **1987**, *237*, 384−389. (c) Baumgartner, A. *J. Chem. Phys*. **1993**, 98, 7496−7501.

(38) (a) Lipkowitz, K. B.; Peterson, M. A. *J. Comput. Chem.* **1993**, *14*, 121−125. (b) Lipkowitz, K. B.; Peterson, M. A. *J. Comput. Chem.* **1995**, *16*, 285−295.

(39) Effects of a level of calculations and a basis set used on the values of the obtained parameters was tracked before[17,19−20] and were not investigated here.

(40) Hydrocarbons: methane, ethane, ethylene (CH_2═CH_2), propane, propene (CH_2═CH-Me), butane, *n*-pentane (*n*-C_5H_{12}), *i*-pentane (*i*-C_5H_{12}), neopentane (CMe_4), *n*-hexane (*n*-C_6H_{14}), cyclohexane (c−C_6H_{12}), methylcyclohexane (Me-c−C_6H_{12}), *n*-octane (*n*-C_8H_{18}), *i*-octane (*i*-C_8H_{18}), *n*-nonane (*n*-C_9H_{20}), 3,3-diethylpentane (CEt_4), *n*-dodecane (*n*-$C_{12}H_{26}$), 4,4-dipropylheptane (CPr_4), *n*-hexadecane (*n*-$C_{16}H_{34}$), 5,5-dibutylnonane (CBu_4); alcohols: methanol (MeOH), ethanol (EtOH), 1-propanol (n-PrOH), 1-hexanol (*n*-$C_6H_{13}OH$), ethylene glycol (EG), glycerol; alkyl halides: chloroform, carbon tetrachloride, chloroethane (EtCl), bromoethane (EtBr), 1-chloropropane (n-PrCl), 1-bromopropane (n-PrBr), 1-chlorobutane (n-BuCl), 1-bromobutane (n-BuBr); aromatic hydrocarbons: pyridine, benzene (PhH), chlorobenzene (PhCl), bromobenzene (PhBr), other organic molecules: acetonitrile (MeCN), acetone; alkyl silanes: tetramethylsilane ($SiMe_4$), tetraethylsilane ($SiEt_4$); inorganic molecules: neon, argon, krypton, nitrogen, water, carbon monooxide, carbon dioxide, carbon disulfide. Abbreviations used in figures are given here in parentheses unless the same name or molecular formula was used for notation.

(41) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. *J. Comput. Chem.* **1995**, *16*, 273−284.

(42) A certain degree of confusion exists regarding of the term *dimension* used here. On one hand each variable describing shape or size has its own geometrical dimensionality in the sense of how many coordinates are involved in the calculation of the variable. For example, as it was used above, a measure of surface or area is two dimensional ($\sim$*length*^{2}), a measure of volume is three-dimensional ($\sim$*length*^{3}), and a variable defined as the ratio of two variables of the same dimensionality is of zero dimension. On the other hand, a descriptor may be a composition of several variables. For example, molecular volume, if assumed as one value per molecule, is a one-dimensional descriptor;[3] the distance matrix is a two-dimensional molecular descriptor; shape matrices in SGM could be of variable dimensionality, usually at least two-dimensional. Dimensionality in CoMFA methods should, in this sense, be assumed to be equal to the number of points in the grid, and all the descriptors listed in Table 1 are one-dimensional.

(43) For illustration, see the details of the development of a specific thrombin inhibitor in Hann, M. M.; Leach, A. R.; Harper, G. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856−864.

(44) Pearlman, R. S.; Smith, K. M. *Perspect. Drug Discovery Des.* **1998**, 339−353.

CI050005L