Roberto Todeschini graduated in Chemistry in 1972 and his main research activities concern chemometrics in all his aspects, the study of quantitative structure–activity relationships (QSAR), molecular descriptors, multicriteria decision making, and software development. President of the Italian Chemometric Society and member of the editorial advisory boards of relevant scientific reviews, he is full Professor of Chemometrics at the Department of Environmental Sciences of the University of Milano-Bicocca (Milano, Italy). He is author of more than 130 publications and international reviews and of the books: The Data Analysis Handbook, by I. E. Frank and R. Todeschini, Elsevier, 1994, and Handbook of Molecular Descriptors, by R. Todeschini and V. Consonni, Wiley–VCH, 2000.

Viviana Consonni studied environmental sciences (B.Sc. in 1997) and then chemistry (Ph.D in 2000) at the University of Milano-Bicocca (Italy). She is now research fellow at the Milano Chemometrics and QSAR Research Group. She has published a book, Handbook of Molecular Descriptors, by R. Todeschini and V. Consonni, Wiley–VCH, 2000, and 15 research papers. She is also co-author of the DRAGON software for the calculation of molecular descriptors.

2
Descriptors from Molecular Geometry

Roberto Todeschini and Viviana Consonni

2.1
Introduction

Molecular descriptors, which play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, health research and quality control, are obtained when molecules, thought of as real objects, are transformed into a molecular representation enabling mathematical treatment. Many molecular descriptors have been proposed; they are derived from different theories and approaches with the aim of predicting biological and physicochemical properties of molecules [1].

The information content of a molecular descriptor depends on the kind of molecular representation that is used and on the defined algorithm for its calculation. There are simple molecular descriptors, derived by counting some atom-types or structural fragments in the molecule, others derived from algorithms applied to a topological representation (molecular graph) and usually called topological or 2D-descriptors, and there are molecular descriptors derived from a geometrical representation that are called geometrical or 3D-descriptors.

Because a geometrical representation involves knowledge of the relative positions of the atoms in 3D space, i.e. the (x, y, z) atomic coordinates of the molecule atoms, geometrical descriptors usually provide more information and discrimination power than topological descriptors for similar molecular structures and molecule conformations. Despite their high information content, geometrical descriptors usually also have drawbacks. They require geometry optimization and, therefore, the overhead to calculate them. For flexible molecules, moreover, several molecule conformations can be possible – on one hand, new information is available and can be exploited, but, on the other hand, the problem complexity can increase significantly. Finally, most geometrical descriptors such as grid-based descriptors need alignment rules to achieve molecule comparability.

For these reasons, topological descriptors, fragment counts and other simple descriptors are usually preferred for screening large databases of molecules. On the other hand, searching for relationships between molecular structures and complex properties, such as biological activity, can often efficiently be performed by use of geometrical descriptors, by exploiting their large information content.

As shown in the Figure 2-1, within the set of geometrical descriptors, several classes of descriptors can be distinguished, e.g. quantum-chemical, grid-based, volume and surface descriptors, etc.

Here, only some of the geometrical descriptors have been considered; most are derived directly from the (x, y, z) coordinates and do not require prior alignment of molecules for analysis.

In the first section, attention will be paid to several simple molecular descriptors mainly proposed by different authors to describe molecular size or shape. In the other sections sets of geometrical descriptors calculated according to homogeneous theoretical schemes will be presented; these descriptors are the WHIM, GETAWAY, 3D autocorrelation, 3D-MoRSE, RDF, EVA and EEVA descriptors.

2.2 Geometrical Descriptors for Molecular Size and Shape

Most geometrical descriptors are calculated directly from the (x, y, z) coordinates of the molecule atoms and other quantities derived from the coordinates, e.g. interatomic distances or distances from a specified origin (e.g. the molecule barycenter). Many of these are derived from the molecular geometry matrix **G** defined

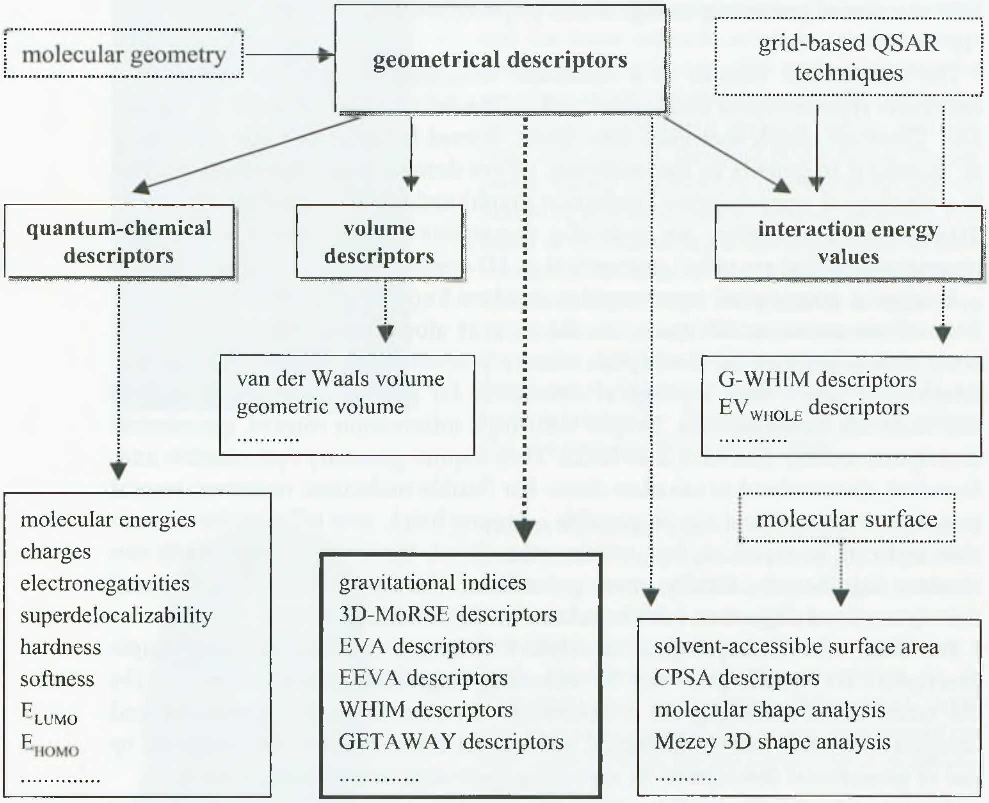

Fig. 2-1 Outline of the different sets of geometrical descriptors

by all the geometrical distances r_{ij} between atom pairs. The geometry matrix is a simple molecular representation in which atoms are viewed as single points in the molecule space. To account for more chemical information, the atoms in the molecule can be represented by their atomic masses and molecular descriptors can be derived from the molecule inertia matrix, **I**, from atom distances relative to the center of mass, and by weighting interatomic distances with functions of atomic masses.

The *geometry matrix* **G** (or *geometric distance matrix*) of a molecule, obtained from atom coordinates, is a square symmetric matrix whose entry r_{st} is the geometric distance calculated as the Euclidean distance between atoms s and t (Eq. (1)):

$$\mathrm{G} \equiv \begin{vmatrix} 0 & r_{12} & \cdots & r_{1A} \\ r_{21} & 0 & \cdots & r_{2A} \\ \cdots & \cdots & \cdots & \cdots \\ r_{A1} & r_{A2} & \cdots & 0 \end{vmatrix} \tag{1}$$

where A is the number of atoms in the molecule. Diagonal entries are always zero.

Although the geometry matrix contains information about molecular configurations and conformations, it does not contain information about atom connectivity. Thus, for several applications, it is accompanied by a *connectivity table* in which, for each atom, the identification numbers of the atoms bonded to it are listed. The geometry matrix can also be calculated on geometry-based standardized bond lengths and bond angles and derived by embedding a graph on a regular two-dimensional or three-dimensional grid; in these cases, the geometry matrix is often referred to as the *topographic matrix* [2].

By using the geometry matrix instead of the topological distance matrix to represent a molecular graph, a number of atom descriptors and related molecular descriptors can be derived. These are called *topographic indices* or topological indices of a new generation, because they combine three-dimensional information given by geometric interatomic distances with topological information given by the molecular graph [3–5].

The *i*th row sum of the geometry matrix is called the *geometric distance degree*, ${}^{G}\sigma_i$; it is an atom descriptor used, for example, in the development of the 3D-connectivity indices $\chi\chi$ [3]. In general, the row sum of this matrix is a measure of the centrality of an atom; atoms that are close to the center of the molecule have smaller atomic sums whereas those far from the center have large atomic sums. Because the smallest and the largest row sums give the extreme values of the first eigenvalue of the **G** matrix, when all the atoms are equivalent, i.e. the distance degrees are all the same, the geometric distance degree yields exactly the *first eigenvalue*. Together with the first eigenvalue, another simple size descriptor is defined as the *sum of the absolute eigenvalues* of the geometry matrix.

The average sum of all geometric distance degrees is a molecular descriptor called *average geometric distance degree*, i.e. (Eq. (2)):

$$ {}^{G}\bar{\sigma} = \frac{1}{A} \cdot \sum_{i=1}^{A} {}^{G}\sigma_i = \frac{1}{A} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} r_{ij} \tag{2} $$

whereas the half sum of all geometric distance degrees is another molecular descriptor called 3D-Wiener index [6, 7] by analogy with the Wiener index it, which is calculated from the topological distance matrix.

The maximum value entry in the *i*th row of the geometry matrix is a local descriptor called the *geometric eccentricity* ${}^{G}\eta_i$ representing the longest geometric distance from the *i*th atom to any other atom in the molecule (Eq. (3)):

$$ {}^{G}\eta_i = \max_j r_{ij} \tag{3} $$

From the eccentricity definition, the *geometric radius* ${}^{G}R$ and *geometric diameter* ${}^{G}D$ can immediately characterize a molecule. The radius of a molecule is defined as the minimum geometric eccentricity, and the diameter is defined as the maximum geometric eccentricity in the molecule, according to Eq. (4):

$$ {}^{G}R = \min_i {}^{G}\eta_i \quad \text{and} \quad {}^{G}D = \max_i {}^{G}\eta_i \tag{4} $$

These terms are size descriptors also depending on the molecular shape, such as their topological counterpart. In fact, the *geometrical shape coefficient*, I_3 [8], has been defined as a function of the geometric radius and diameter as (Eq. (5)):

$$I_3 = \frac{{}^{G}D - {}^{G}R}{{}^{G}R} \tag{5}$$

Triangular descriptors (or *triplet descriptors*) can easily be calculated from the geometry matrix. They describe the relative positions of three atoms or group centroids in the molecule. Each possible triplet of non-hydrogen atoms is taken as a triangle, and different triangle measures have been proposed such as individual triangle side lengths (i.e. geometric interatomic distances), triangular perimeter and area; these measures are integerized and transformed into single bit integers of defined length by different procedures, and their distribution is used to describe the molecule. They are used both to characterize molecular shape and for 3D pharmacophore database searching [9–11]. Similar to the triangular descriptors are potential pharmacophore point pairs (*PPP pairs*) and potential pharmacophore point triangles (*PPP triangles*), which are 3D fingerprints encoding, respectively, the distance information between all possible combinations of two and three potential pharmacophore points [12]. Potential pharmacophore points usually considered are hydrogen bond donors and acceptors, sites of potential negative and positive charge, and hydrophobic atoms. Moreover, a set of molecular descriptors can be obtained by summing the geometric distances between all possible combinations of predefined heteroatom-type pairs, such as N...N, N...O, O...O, N...S, N...P, N...Cl, O...P, etc.

From the atomic coordinates, the distances between each atom and the center of mass can be also calculated. The simplest descriptor based on this approach is the *span*, R [13], a size descriptor defined as the radius of the smallest sphere, centered on the center of mass, completely enclosing all atoms of a molecule (Eq. (6)):

$$R = \max_i r_i \tag{6}$$

where r_i is the distance of the ith atom from the center of mass.

The *average span*, calculated as the average value of conformational changes and denoted by $\bar{R}$, is used to describe long-chain molecules, such as macromolecules, polymers and proteins.

A geometrical shape descriptor based on the radii of three spheres centered at the barycenter of the molecule is the *Meyer visual descriptor of globularity*, R_M [14]. The first sphere of radius R_1 has a volume equal to the van der Waals volume; the second sphere of radius R_2 has a volume equal to the molecular volume, and the third sphere of radius R_3 is defined as the sphere embedding the whole molecule (i.e. the span R). The shape descriptor is then defined as (Eq. (7)):

$$R_M = \frac{R_3 - R_2}{R_2 - R_1} \tag{7}$$

A spherical top corresponds to small values of the R_M term.

Related to the previous descriptors is also the *end-to-end distance*, r_{ee} [15], which is defined for long chain molecules as (Eq. (8)):

$$r_{ee} = \|\mathbf{r}_1 - \mathbf{r}_A\| \tag{8}$$

where $\mathbf{r}$ is the vector of the atom coordinates with respect to the center of mass.

When the information carried by the atom masses is added to the interatomic distances, several other molecular descriptors can be defined.

Among these, the *gravitational indices* [16] are geometrical descriptors reflecting the mass distribution in a molecule, defined as (Eq. (9)):

$$G_1 = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \frac{m_i \cdot m_j}{r_{ij}^2} \quad \text{and} \quad G_2 = \sum_{b=1}^{B} \left(\frac{m_i \cdot m_j}{r_{ij}^2} \right)_b \tag{9}$$

where m_i and m_j are the atomic masses of the considered atoms, r_{ij} the corresponding interatomic distance, A and B the number of atoms and bonds of the molecule, respectively. The G_1 index takes into account all atom pairs in the molecule whereas the G_2 index is restricted to pairs of bonded atoms. These indices are related to the bulk cohesiveness of the molecules accounting, simultaneously, for both atomic masses (volumes) and their distribution within the molecular space. For modeling purposes the square root and cubic root of the gravitational indices were also proposed [17]. Both indices can be extended to any other atomic property different from atomic mass, such as atomic polarizability, van der Waals volume and electronegativity.

Exploiting the distance between each atom and the molecule center of mass, the *radius of gyration*, R_G [13, 18], is a size descriptor based on the distribution of atomic masses in a molecule, defined as (Eq. (10)):

$$R_G = \sqrt{\frac{\sum_{i=1}^{A} m_i \cdot r_i^2}{MW}} \tag{10}$$

where r_i is the distance of the ith atom from the center of mass of the molecule, m_i is the corresponding atomic mass, A the number of atoms and MW the molecular weight.

The radius of gyration can also be calculated from the principal moments of inertia, I; for planar molecules ($I_C = 0$) it is defined as (Eq. (11)):

$$R_G = \sqrt{\frac{(I_A \cdot I_B)^{1/2}}{MW}} \tag{11}$$

for non-planar molecules as (Eq. (12)):

$$R_G = \sqrt{\frac{2\pi \cdot (\mathrm{I_A} \cdot \mathrm{I_B} \cdot \mathrm{I_C})^{1/3}}{MW}} \tag{12}$$

The radius of gyration is a measure of molecular compactness for long-chain molecules and, specifically, small values are obtained when most of the atoms are close to the center of mass.

Several other geometrical descriptors are derived from the principal moments of inertia, I, of a molecule. Principal moments of inertia I are physical quantities related to the rotational dynamics of a molecule. The moment of inertia about any axis is defined as (Eq. (13)):

$$\mathrm{I} = \sum_{i=1}^{A} m_i \cdot r_i^2 \tag{13}$$

where A is the number of atoms, and m_i and r_i are the atomic mass and the perpendicular distance from the chosen axis of the ith atom of the molecule, respectively.

Principal moments of inertia are the moments of inertia corresponding to that particular and unique orientation of the axes for which one of the three moments has a maximum value, another a minimum value, and the third is either equal to one of the other or is intermediate in value between the other two. The corresponding axes are called principal axes of a molecule (or principal inertia axes). Conventionally, principal moments of inertia are labeled as $I_\mathrm{A} \leq I_\mathrm{B} \leq I_\mathrm{C}$; they can simply be calculated by diagonalization of the molecule inertia matrix, being the corresponding eigenvalues.

In general, the three principal moments of inertia have different values but, depending on the molecular symmetry, they show characteristic equalities; for this, a number of shape descriptors is defined in terms of principal moments of inertia. Moreover, principal inertia axes are used to provide a unique reference framework for calculation of several geometrical descriptors, e.g. the WHIM descriptors and the shadow indices reported below.

The *inertial shape factor*, S_I [19], is a shape measure based on the principal moments of inertia, I, and defined as (Eq. (14)):

$$\mathrm{S_I} = \frac{\mathrm{I_B}}{\mathrm{I_A} \cdot \mathrm{I_C}} \tag{14}$$

This descriptor cannot be calculated for planar molecules.

Molecular eccentricity, ε [20], is another shape descriptor defined as (Eq. (15)):

$$\varepsilon = \frac{(\mathrm{I_A^2} - \mathrm{I_C^2})^{1/2}}{\mathrm{I_A}} \qquad 0 \leq \varepsilon \leq 1 \tag{15}$$

where $\varepsilon = 0$ corresponds to spherical top molecules and $\varepsilon = 1$ to linear molecules.

Moreover, the *asphericity*, Ω_A [20], which is an anisometry descriptor for the deviation from the spherical shape, has been defined as (Eq. (16)):

$$\Omega_A = \frac{1}{2} \cdot \frac{(I_A - I_B)^2 + (I_A - I_C)^2 + (I_B - I_C)^2}{I_A^2 + I_B^2 + I_C^2} \qquad 0 \leq \Omega_A \leq 1 \tag{16}$$

where $\Omega_A = 0$ corresponds to spherical top molecules and $\Omega_A = 1$ to linear molecules. For prolate (cigar-shaped) molecules, $I_A \approx I_B > I_C$ and $\Omega_A \approx 0.25$, whereas for oblate (disk-shaped) molecules, $I_A > I_B \approx I_C$ and $\Omega_A \approx 1$.

The *spherosity index*, Ω_S, is an anisometry descriptor defined as a function of the eigenvalues of the covariance matrix of the atomic coordinates (Eq. (17)):

$$\Omega_S = \frac{3 \cdot \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} \qquad 0 \leq \Omega_S \leq 1 \tag{17}$$

The spherosity index varies from zero for flat molecules, such as benzene, to unity for totally spherical molecules [21].

Other approaches to description of molecular structure are briefly reported below. They involve knowledge of molecular geometry together with molecular properties such as van der Waals volume or radius and surface area.

The *Kaliszan shape parameter*, η [22], has been defined as the ratio of the longest to the shortest side of a rectangle having the minimum area that can envelope a molecular structure drawn assuming van der Waals radii for atoms and standard bond lengths.

A slightly different shape parameter is the *length-to-breadth ratio*, L/B, which is defined as the ratio of the longest to the shortest side of the rectangle that envelopes a molecular structure and at the same time maximizes the L/B ratio [23, 24]. In general, the length-to-breadth ratio is the ratio of the longest L to the shortest B side of a rectangle containing some molecular projection, having unequivocally defined a specific molecular orientation. The length-to-breadth ratio was calculated from the dimensions of rectangles that envelope the molecule oriented along the principal inertia axes [25].

Focusing attention on substituents instead of the whole molecule, the *Sterimol parameters* were proposed by Verloop [26] to describe the size and shape of substituents in a congeneric series. These were evaluated by measuring the dimensions of substituents in a restricted number of directions by a computer program (STERIMOL) which simulates 3D model building of substituent groups, using the Corey–Pauling–Koltun volume (CPK atomic models). For flexible substituents minimum energy conformations are considered.

Based on the same approach of embedding a molecule in specified rectangles, six *shadow indices* have been defined as a set of geometrical descriptors calculated by projecting the molecular surface on to three mutually perpendicular planes XY, XZ and YZ, assuming van der Waals radii for atoms [27]. Basically, a molecule is flattened into a plane by disregarding the third dimension; the area of the molecule which is projected on to the considered two dimensions defines the shadow

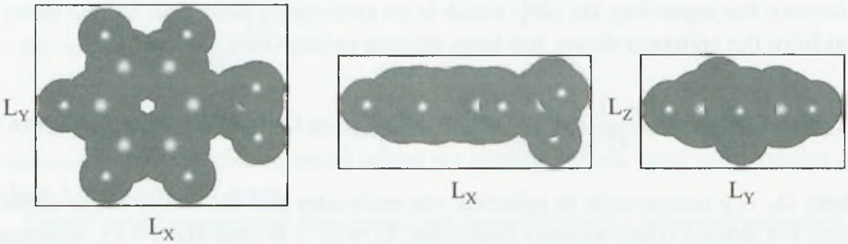

Fig. 2-2 Projections and embedding rectangles of the toluene molecule in the three principal planes

area of interest (Fig. 2-2). To obtain invariance to rotation of the calculated projections, the X, Y, Z molecule axes are previously aligned along the three principal inertia axes.

The *ovality index*, *O*, is an anisometry descriptor based on the property that, for a fixed volume, the spherical shape presents the minimum surface [28]. It is calculated from the ratio between the actual molecular surface area SA and the minimum surface area SA_0 corresponding to the molecule of the van der Waals volume V_{VDW} (Eq. (18)):

$$O = \frac{SA}{SA_0} = \frac{SA}{4\pi R^2} = \frac{SA}{4\pi \cdot \left(\frac{3 \cdot V_{VDW}}{4\pi}\right)^{2/3}} \qquad O \geq 1 \tag{18}$$

where *R* is the molecule radius. The ovality index is equal to 1 for spherical top molecules and increases with increasing linearity of the molecule.

The inverse of the ovality index [29] is the *globularity factor*, G ($0 < G \leq 1$), which is between zero and one. For molecules with the same volume, the most spherical species have G values approximating unity, and for molecules of non-equal volume, *G* reflects the relative compactness. When both the effective surface area and the volume of the molecule are available, the *surface–volume ratio* $G' = SA/V$ can be used as a descriptor of molecular congestion. More specifically, it was interpreted as a measure of the capability of a compound to adapt its shape to the requirements of an approaching reagent.

2.3 WHIM Descriptors

WHIM descriptors (*Weighted Holistic Invariant Molecular descriptors*) are geometrical descriptors based on statistical indices calculated on the projections of the atoms along principal axes [30, 31].

WHIM descriptors are built in such a way as to capture relevant molecular 3D information regarding molecular size, shape, symmetry and atom distribution

Tab. 2-1 WHIM descriptors

Formula	*Eq. no.*	*Name*	*Molecular feature*
$\lambda_m \quad m = 1, 2, 3$	(19)	d-WSIZ indices	Axial dimension
$\mathrm{T} = \lambda_1 + \lambda_2 + \lambda_3$	(20)	WSIZ index	Global dimension
$\mathrm{A} = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3$	(21)	WSIZ index	Global dimension
$\mathrm{V} = \prod_{m=1}^{3}(1 + \lambda_m) - 1 = \mathrm{T} + \mathrm{A} + \lambda_1\lambda_2\lambda_3$	(22)	WSIZ index	Global dimension
$\vartheta_m = \frac{\lambda_m}{\sum_m \lambda_m} \quad m = 1, 2, 3$	(23)	d-WSHA indices	Axial shape
$K = \frac{3}{4} \cdot \sum_{m=1}^{3} \left\| \frac{\lambda_m}{\sum_m \lambda_m} - \frac{1}{3} \right\|$	(24)	WSHA index	Global Shape
$\eta_m = \frac{\lambda^2 \cdot A}{\sum_i t_i^4} \quad m = 1, 2, 3$	(25)	d-WDEN indices	Axial density
$\mathrm{D} = \eta_1 + \eta_2 + \eta_3$	(26)	WDEN index	Global density
$\gamma_m = \left\{1 - \left[\frac{n_s}{A} \cdot \log_2 \frac{n_s}{A} + n_a \cdot \left(\frac{1}{A} \cdot \log_2 \frac{1}{A}\right)\right]\right\}^{-1} \quad m = 1, 2, 3$	(27)	d-WSYM indices	Axial symmetry
$\mathrm{G} = (\gamma_1 \cdot \gamma_2 \cdot \gamma_3)^{1/3}$	(28)	WSYM index	Symmetry

λ refers to eigenvalues of the weighted covariance matrix; t refers to atomic coordinates with respect to the principal axes; A is the number of molecule atoms; n_s is the number of symmetric atoms along a principal axis and n_a the number of asymmetric atoms.

with respect to invariant reference frames. They are divided into two main classes: *directional WHIM descriptors* and *global WHIM descriptors*. A summary of WHIMs is shown in Table 2-1.

Directional WHIM descriptors are calculated as univariate statistical indices on the projections of the atoms along each individual principal axis whereas global WHIMs are directly calculated as a combination of the former, thus simultaneously accounting for variation of molecular properties along the three principal directions in the molecule. In this case, any information individually related to each principal axis disappears and the description is related only to a global view of the molecule.

Within the WHIM approach, a molecule is seen as a configuration of points (the atoms) in the three-dimensional space defined by the Cartesian axes (x, y, z). To obtain a unique reference frame, principal axes of the molecule are calculated. Projections of the atoms along each of the principal axes are then performed and their dispersion and distribution around the geometric center are evaluated.

Indeed, the algorithm consists in calculating the eigenvalues and eigenvectors of a weighted covariance matrix of the centered Cartesian coordinates of a molecule, obtained from different *weighting schemes* for the atoms (Eq. (29)):

$$s_{qq'} = \frac{\sum_{i=1}^{A} w_i(q_i - \bar{q})(q'_i - \bar{q}')}{\sum_{i=1}^{A} w_i} \tag{29}$$

where $s_{qq'}$ is the weighted covariance between the atomic coordinates q and q' ($q, q' = x, y, z$), A is the number of atoms, w_i the weight of the ith atom, q_i and q'_i represent the coordinates of the ith atom, and $\bar{q}$ the corresponding average value.

Six different weighting schemes were proposed:

1. the unweighted case u ($w_i = 1$ $i = 1, A$, where A is the number of atoms for each compound);
2. atomic mass, m;
3. the van der Waals volume, v;
4. the Sanderson atomic electronegativity, e;
5. the atomic polarizability, p;
6. the electrotopological state indices S of Kier and Hall.

All the weights are scaled with respect to the carbon atom.

Depending on the kind of weighting scheme, different covariance matrixes and, therefore, different principal axes are obtained. For example, using atomic masses as the weighting scheme, the directions of the three principal axes are the directions of the principal inertia axes. Thus, the WHIM approach can be viewed as a generalization of searching for the principal axes with respect to a defined atomic property (the weighting scheme). Based on the same principles of the WHIMs, *COMMA2 descriptors* have recently been proposed [32]. They consist of 11 descriptors given by moment expansions for which the zero-order moment of a property field is non-vanishing.

WHIM descriptors are invariant to translation, because of the centering of the atomic coordinates, and invariant to rotation, because of the uniqueness of the principal axes; they are, therefore, not affected by prior alignment of molecules.

To make the WHIM approach clearer, let us look at a simple example. Consider chlorobenzene to be the molecule for analysis; it can be thought of as the configuration of points shown in Figure 2-3, the atomic Cartesian coordinates being those shown in Table 2-2.

This reference frame is obviously not unique, because it depends on how the molecule has been drawn and how its conformation has been optimized. Thus, to calculate unique molecular descriptors independent of the reference frame the principal axes have to be found. Figure 2-4 shows the principal axes of chloro-

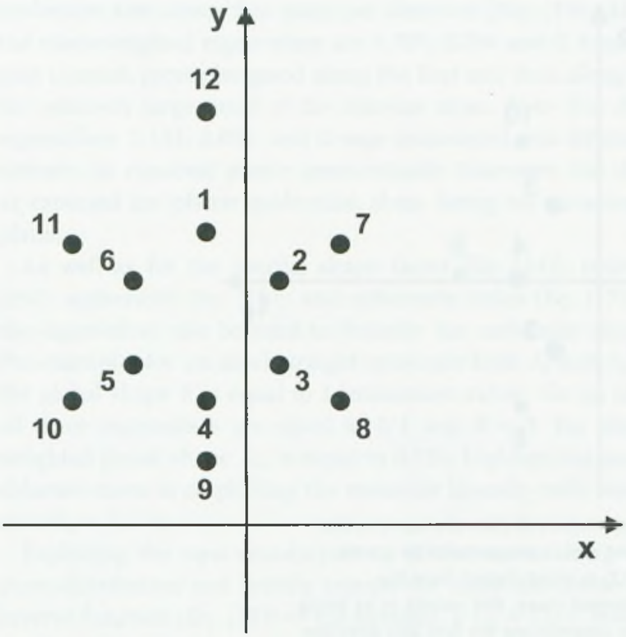

Fig. 2-3 Geometrical representation of chlorobenzene based on the Cartesian coordinates. The chlorine atom (12) shows the highest distance from the aromatic ring

benzene computed by considering each point weighted by the corresponding atomic mass. Note that the first principal axis is along the direction of the heteroatom and the origin of the reference frame coincides with the geometrical center of the molecule. The atom coordinates with respect to the new reference frame

Tab. 2-2 Cartesian atomic coordinates of an optimized geometry of chlorobenzene

ID	*Atom*	*x*	*y*	*z*
1	C	−0.662	4.186	0
2	C	0.549	3.489	0
3	C	0.547	2.093	0
4	C	−0.662	1.395	0
5	C	−1.871	2.093	0
6	C	−1.873	3.489	0
7	H	1.511	4.030	0
8	H	1.502	1.540	0
9	H	−0.662	0.291	0
10	H	−2.826	1.540	0
11	H	−2.835	4.030	0
12	Cl	−0.662	5.911	0

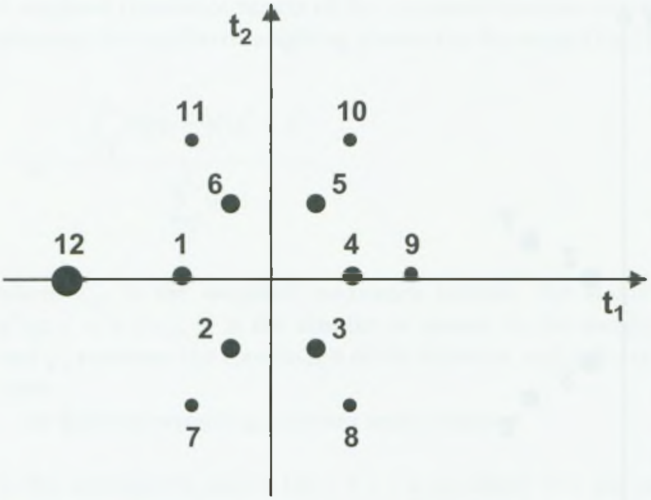

Fig. 2-4 Geometrical representation of chlorobenzene in the space of principal axes. Point size is proportional to atomic mass. The chlorine atom (12) is most distant from the aromatic ring and has the largest mass; this results in its being the most important atom in determining the first axis direction

together with scaled atomic masses are shown in Table 2-3. In general, the effect of weighting atoms with atomic properties consists of redirecting the principal axes towards molecule regions with large property values.

The eigenvalues λ_1, λ_2 and λ_3 of the weighted covariance matrix of the molecule atomic coordinates play a fundamental role in the WHIM descriptor calculation. Each eigenvalue represents a dispersion measure (i.e. the weighted variance) of the projected atoms along the considered principal axis, thus accounting for the

Tab. 2-3 Scaled atomic masses and coordinates relative to the principal axes of chlorobenzene

ID	*Atom*	*m*	t_1	t_2	t_3
1	C	1	−1.346	0	0
2	C	1	−0.649	−1.211	0
3	C	1	0.748	−1.209	0
4	C	1	1.446	0	0
5	C	1	0.748	1.209	0
6	C	1	−0.649	1.211	0
7	H	0.084	−1.189	−2.173	0
8	H	0.084	1.300	−2.164	0
9	H	0.084	2.549	0	0
10	H	0.084	1.300	2.164	0
11	H	0.084	−1.189	2.173	0
12	Cl	2.952	−3.071	0	0

molecular size along that principal direction (Eqs (19)–(22)). For chlorobenzene, the mass-weighted eigenvalues are 3.709, 0.794 and 0, highlighting that the molecule is much more elongated along the first axis than along the second, because of the relatively large mass of the chlorine atom. Note that if the three unweighted eigenvalues 2.333, 2.055, and 0 were considered this difference would be less significant, as expected purely geometrically. Moreover, the third eigenvalue is zero as expected for planar molecules, there being no variance out of the molecular plane.

As well as for the inertial shape factor (Eq. (14)), molecular eccentricity (Eq. (15)), asphericity (Eq. (16)), and spherosity index (Eq. (17)), relationships among the eigenvalues can be used to describe the molecular shape (Eqs (23) and (24)). For example, for an ideal straight molecule both λ_2 and λ_3 are equal to zero and the global shape K is equal to 1 (maximum value); for an ideal spherical molecule all three eigenvalues are equal to 1/3 and $K = 0$. For chlorobenzene the mass-weighted global shape K_m is equal to 0.736, highlighting once again the role of the chlorine mass in amplifying the molecule linearity with respect to the unweighted case $K_u = 0.500$.

Exploiting the new coordinates t_m of the atoms along the principal axes, the atom distribution and density around the molecule center can be evaluated by an inverse function (Eq. (25)) of the kurtosis, κ ($\eta = 1/\kappa$). Low values of the kurtosis are obtained when the data points (i.e. the atom projections) assume opposite values relative to the center. When an increasing number of data values are within the extreme values along a principal axis, the kurtosis value increases (i.e. kurtosis ranges from 1.8 for a uniform distribution of points to 3.0 for a normal distribution). When the kurtosis value tends to infinity the corresponding η value tends to zero.

In an analogous way, from the analysis of the new coordinates t_m of the atoms, molecular symmetry is evaluated on the basis of the number n_s of symmetric atoms with respect to the molecule center, i.e. atoms with opposite coordinates along the considered axis, and the number n_a of asymmetric atoms (Eq. (27)).

In conclusion, for each weighting scheme, w, 11 molecular directional WHIM descriptors (ϑ_3 is excluded) were proposed, thus resulting in a total of 66 directional WHIM descriptors.

For planar compounds, λ_3, γ_3, and η_3 are always equal to zero. The global WHIMs are five for each of the six proposed weighting schemes, w, plus the symmetry indices Gu, Gm and Gs, giving a total number of 33 descriptors.

WHIM descriptors have been used to model toxicological indices [31, 33, 34], several physicochemical properties of PCBs [35, 36] and PAHs [37], hydroxyl radical reaction rate constants [38], and soil sorption partition coefficients [39].

2.4 GETAWAY Descriptors

The GETAWAY (*GEometry, Topology, and Atom-Weights AssemblY*) descriptors [40] have recently been proposed as chemical structure descriptors derived from a new representation of molecular structure, the *Molecular Influence Matrix* (MIM), denoted by **H** and defined as follows (Eq. (30)):

$$H = M \cdot (M^{\mathrm{T}} \cdot M)^{-1} \cdot M^{\mathrm{T}} \tag{30}$$

where M is the molecular matrix consisting of the centered Cartesian coordinates x, y, z of the molecule atoms (hydrogens included) in a chosen conformation, and the superscript T refers to the transposed matrix. Atomic coordinates are assumed to be calculated relative to the geometrical center of the molecule to obtain translational invariance. The molecular information matrix is a symmetric $A \times A$ matrix, where A represents the number of atoms, and shows rotational invariance with respect to the molecule coordinates, and is thus independent of molecule alignment.

The diagonal elements h_{ii} of the molecular influence matrix, called *leverages*, range from 0 to 1 and encode atomic information related to the "influence" of each molecule atom in determining the whole shape of the molecule; in fact mantle atoms always have higher h_{ii} values than atoms near the molecule center. Moreover, the magnitude of the maximum leverage in a molecule depends on the size and shape of the molecule. As derived from the geometry of the molecule, leverage values are effectively sensitive to significant conformational changes and to the bond lengths that account for atom types and bond multiplicity.

Each off-diagonal element h_{ij} represents the degree of accessibility of the jth atom to interactions with the ith atom, or, in other words, the tendency of the two considered atoms to interact with each other. A negative sign for the off-diagonal elements means that the two atoms occupy opposite molecular regions relative to the center, hence the extent of their mutual accessibility should be low.

Table 2-4 shows the molecular influence matrix of chlorobenzene, the three-dimensional structure of which has been optimized by minimizing the conformational energy. Atom numbering of chlorobenzene is shown in Figure 2-5.

Tab. 2-4 Molecular influence matrix of chlorobenzene; atom-numbering refers to Figure 2-5

ID	*C_1*	*C_2*	*C_3*	*C_4*	*C_5*	*C_6*	*H_7*	*H_8*	*H_9*	*H_{10}*	*H_{11}*	*Cl_{12}*
C_1	**0.065**	0.031	−0.036	−0.070	−0.036	0.031	0.057	−0.063	−0.123	−0.063	0.057	0.148
C_2	0.031	**0.075**	0.042	−0.034	−0.077	−0.044	0.134	0.076	−0.059	−0.136	−0.079	0.071
C_3	−0.036	0.042	**0.079**	0.039	−0.039	−0.077	0.075	0.141	0.068	−0.071	−0.138	−0.082
C_4	−0.070	−0.034	0.039	**0.075**	0.039	−0.034	−0.061	0.067	0.132	0.067	−0.061	−0.159
C_5	−0.036	−0.077	−0.039	0.039	**0.079**	0.042	−0.138	−0.071	0.068	0.141	0.075	−0.082
C_6	0.031	−0.044	−0.077	−0.034	0.042	**0.075**	−0.079	−0.136	−0.059	0.076	0.134	0.071
H_7	0.057	0.134	0.075	−0.061	−0.138	−0.079	**0.242**	0.135	−0.108	−0.246	−0.141	0.130
H_8	−0.063	0.076	0.141	0.067	−0.071	−0.136	0.135	**0.250**	0.118	−0.129	−0.246	−0.143
H_9	−0.123	−0.059	0.068	0.132	0.068	−0.059	−0.108	0.118	**0.232**	0.118	−0.108	−0.280
H_{10}	−0.063	−0.136	−0.071	0.067	0.141	0.076	−0.246	−0.129	0.118	**0.250**	0.135	−0.143
H_{11}	0.057	−0.079	−0.138	−0.061	0.075	0.134	−0.141	−0.246	−0.108	0.135	**0.242**	0.130
Cl_{12}	0.148	0.071	−0.082	−0.159	−0.082	0.071	0.130	−0.143	−0.280	−0.143	0.130	**0.337**

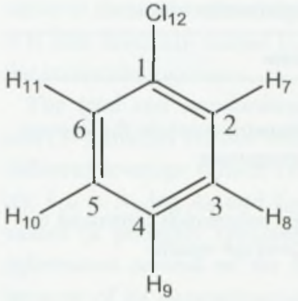

Fig. 2-5 Atom numbering of chlorobenzene

Leverage values of the atoms of chlorobenzene, bromobenzene, and iodobenzene are shown in Figure 2-6. It can be noted that the outer atoms have larger leverage values than the carbon atoms of the aromatic ring. Then, among the outer atoms, the halogens have the largest value and this value increases from chlorobenzene to bromobenzene and to iodobenzene, because it is sensitive to the increase in bond length. Note also that equivalent atoms have equal leverage values.

By combining the elements of the MIM matrix **H** with those of the geometry matrix **G**, another symmetric $A \times A$ molecular matrix, called *influence/distance matrix*, **R**, has been derived as follows (Eq. (31)):

$$[\mathrm{R}]_{ij} = \left[\frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{ij}}\right]_{ij} \quad i \neq j \tag{31}$$

where h_{ii} and h_{jj} are the leverages of the two considered atoms, and r_{ij} is their geometric distance. The diagonal elements of the matrix **R** are zero, while each off-diagonal element *i-j*, resembling the single terms in the sums of the gravitational indices (Eq. (9)), is calculated from the ratio of the geometric mean of the corre-

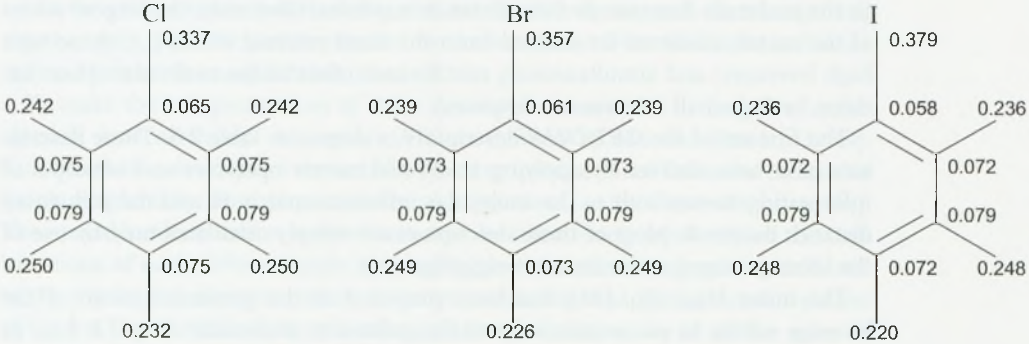

Fig. 2-6 Leverage values of the atoms of chlorobenzene, bromobenzene, and iodobenzene

Tab. 2-5 GETAWAY descriptors based on matrix operators and information indices

Formula	*Eq. no.*	*Name*
$H_{GM} = 100 \cdot \left(\prod_{i=1}^{A} h_{ii} \right)^{1/A}$	(32)	Geometric mean on the leverage magnitude
$\mathrm{I_{TH}} = \mathrm{A_0} \cdot \log_2 \mathrm{A_0} - \sum_{g=1}^{G} \mathrm{N_g} \cdot \log_2 \mathrm{N_g}$	(33)	Total information content on the leverage equality
$\mathrm{I_{SH}} = \frac{\mathrm{I_{TH}}}{\mathrm{A_0} \cdot \log_2 \mathrm{A_0}} = 1 - \frac{\sum_{g=1}^{G} \mathrm{N_g} \cdot \log_2 \mathrm{N_g}}{\mathrm{A_0} \cdot \log_2 \mathrm{A_0}}$	(34)	Standardized information content on the leverage equality
$HIC = \bar{I}_{\mathrm{H}} = -\sum_{i=1}^{A} \frac{h_{ii}}{D} \cdot \log_2 \frac{h_{ii}}{D}$	(35)	Mean information content on the leverage magnitude
$RARS = \frac{1}{A} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} \frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{ij}} = \frac{1}{A} \cdot \sum_{i=1}^{A} RS_i$	(36)	Average row sum of the influence/ distance matrix
$RCON = \sum_{b=1}^{B} (RS_i \cdot RS_j)_b^{1/2}$	(37)	R-connectivity index
$REIG = \lambda_1$	(38)	First eigenvalue of the **R** matrix

A is the number of molecule atoms (hydrogen included); A_0 is the number of non-hydrogen atoms; $\mathrm{N_g}$ is the number of atoms with the same leverage value and G the number of equivalence classes; $D = 1$, 2 or 3 (1 for linear, 2 for planar and 3 for non-planar molecules); B is the number of molecule bonds.

sponding ith and jth diagonal elements of the matrix **H** to the interatomic distance r_{ij} provided by the geometry matrix **G**.

The squared root product of the leverages of two atoms is divided by their interatomic distance to make the contributions from pairs of atoms which are far apart less significant, according to the basic idea that interactions between atoms in the molecule decrease as their distance increases. Obviously, the largest values of the matrix elements are derived from the most external atoms (i.e. those with high leverages) and simultaneously next to each other in the molecular space (i.e. those having small interatomic distances).

The first set of the GETAWAY descriptors is shown in Table 2-5. These descriptors have been derived by applying traditional matrix operators and concepts of information theory both to the molecular influence matrix **H** and the influence/ distance matrix **R**. Most of these descriptors are simply calculated only by use of the leverages used as the atomic weightings.

The index H_{GM} (Eq. (32)) has been proposed as the geometric mean of the leverage values to encompass information related to molecular shape. It has, in fact, been found that in an isomeric series of hydrocarbons, the H_{GM} index is sensitive to the molecular shape increasing from linear to more branched molecules; it is also inversely related to molecular size, decreasing as the number of atoms in the molecule increases.

The *total* and *standardized information content on the leverage equality* (Eqs (33) and (34)) mainly encode information on molecular symmetry; if all the atoms have different leverage values, i.e., the molecule does not have any element of symmetry, $I_{TH} = A_0 \log A_0$ and $I_{SH} = 1$; otherwise, if all the atoms have equal leverage values (a perfectly symmetric theoretical case), $I_{TH} = 0$ and $I_{SH} = 0$. The *total information content on the leverage equality* I_{TH} is more discriminating than I_{SH}, because of its dependence on molecular size, and thus it could be thought of as a measure of molecular complexity. These indices have been demonstrated to be useful in modeling physicochemical properties related to entropy and symmetry [41].

The *HIC* descriptor (Eq. (35)) seems to encompass more information related to molecular complexity than the total and standardized information content on the leverage equality. Differently from I_{TH} and I_{SH}, *HIC* can, for example, recognize the different substituents in a series of monosubstituted benzenes. It is also sensitive to the presence of multiple bonds.

Both *RARS* (Eq. (36)) and *RCON* (Eq. (37)) are based on the row sums of the influence/distance matrix, because these encode useful information that could be related to the presence of significant substituents or fragments in the molecule. It has, in fact, been observed that larger row sums correspond to terminal atoms that are located very next to other terminal atoms such as those in substituents on a parent structure. Moreover, the *RCON* index is very sensitive to the molecular size and to conformational changes and cyclicity.

The *REIG* descriptor (Eq. (38)) has been defined by analogy with the Lovasz–Pelikan index [42], which is an index of molecular branching calculated as the first eigenvalue of the adjacency matrix.

RARS and *REIG* indices are closely related; their values decrease as molecular size increases and seem to be a little more sensitive to molecular branching than to cyclicity and conformational changes.

GETAWAY descriptors in the other set, shown in Table 2-6, are based on spatial autocorrelation formulas, weighting the molecule atoms in such a way as to account for atomic mass, polarizability, van der Waals volume, and electronegativity together with 3D information encoded by the elements of the molecular influence matrix **H** and influence/distance matrix **R**.

To make the comprehension of these descriptors clearer, let us look first at the classical autocorrelation descriptors ATS as defined by Moreau–Broto [43, 44]. Autocorrelation descriptors are usually 2D-descriptors derived from the molecular graph weighted by atom physicochemical properties (i.e. the atom weightings w_i). The spatial autocorrelation is then evaluated by considering separately all the contributions of each different path length (*lag*) in the molecular graph, as collected in the topological distance matrix. In other words, the total spatial autocorrelation at *lag* k ATS_k is obtained by summing all the products $w_i \cdot w_j$ of all the pairs of atoms i and j, for which the topological distance equals the *lag* as (Eq. (47)):

Tab. 2-6 GETAWAY descriptors based on autocorrelation functions

Formula	*Eq. no.*	*Name*
$HATS_k(w) = \sum_{i=1}^{A-1} \sum_{j>i} (w_i \cdot h_{ii}) \cdot (w_j \cdot h_{jj}) \cdot \delta(k; d_{ij}) \quad k = 0, 1, 2, \ldots, d$	(39)	HATS indices
$HATS(w) = HATS_0(w) + 2 \cdot \sum_{k=1}^{d} HATS_k(w)$	(40)	HATS total index
$H_k(w) = \sum_{i=1}^{A-1} \sum_{j>i} h_{ij} \cdot w_i \cdot w_j \cdot \delta(k; d_{ij}; h_{ij}) \quad k = 0, 1, 2, \ldots, d$	(41)	H indices
$HT(w) = H_0(w) + 2 \cdot \sum_{k=1}^{d} H_k(w)$	(42)	H total index
$R_k(w) = \sum_{i=1}^{A-1} \sum_{j>i} \frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{ij}} \cdot w_i \cdot w_j \cdot \delta(k; d_{ij}) \quad k = 1, 2, \ldots, d$	(43)	R indices
$RT(w) = 2 \cdot \sum_{k=1}^{d} R_k(w)$	(44)	R total index
$R_k^+(w) = \max_{ij} \left(\frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{ij}} \cdot w_i \cdot w_j \cdot \delta(k; d_{ij}) \right) \quad i \neq j \text{ and } k = 1, 2, \ldots, d$	(45)	Maximal R indices
$RT^+(w) = \max_k(R_k^+(w))$	(46)	Maximal R total index

A is the number of molecule atoms (hydrogen included); d is the topological diameter; d_{ij} is the topological distance between atoms i and j; and w_i is a physicochemical weight for the ith atom.

$$ATS_k = \sum_{i=1}^{A-1} \sum_{j>i} w_i \cdot w_j \cdot \delta(k; d_{ij}) \quad k = 0, 1, 2, \ldots, d \tag{47}$$

where d_{ij} is the topological distance between atoms i and j, d is the topological diameter, i.e. the maximum topological distance in the molecule, and $\delta(k; d_{ij})$ is a Dirac-delta function defined as:

$$\delta(k; d_{ij}) = \begin{cases} 1 & \text{if } d_{ij} = k \\ 0 & \text{if } d_{ij} \neq k \end{cases}$$

An example of ATS calculation for chlorobenzene is shown here. The topological distances of chlorobenzene atoms have been collected in Table 2-7. Let us calculate the ATS descriptor with *lag* $k = 4$ and the weight equal to the scaled atomic mass m (Table 2-3). We have to consider all the atom pairs with a topological distance equal to 4; these atom pairs have been highlighted in Table 2-7.

Tab. 2-7 Topological distance matrix of chlorobenzene. Toplogical distances equal to 4 are in bold face

ID	C_1	C_2	C_3	C_4	C_5	C_6	H_7	H_8	H_9	H_{10}	H_{11}	Cl_{12}
C_1	0	1	2	3	2	1	2	3	4	3	2	1
C_2		0	1	2	3	2	1	2	3	4	3	2
C_3			0	1	2	3	2	1	2	3	4	3
C_4				0	1	2	3	2	1	2	3	4
C_5					0	1	4	3	2	1	2	3
C_6						0	3	4	3	2	1	2
H_7							0	3	4	5	4	3
H_8								0	3	4	5	4
H_9									0	3	4	5
H_{10}										0	3	4
H_{11}											0	3
Cl_{12}												0

The value of $\mathrm{ATS}_4(m)$ is calculated as:

$$\begin{aligned}\mathrm{ATS}_4(m) &= m_1m_9 + m_2m_{10} + m_3m_{11} + m_4m_{12} + m_5m_7 + m_6m_8 + m_7m_9 \\ &\quad + m_8m_{10} + m_9m_{11} + m_{10}m_{12} + m_{11}m_7 + m_{12}m_8 \\ &= 1 \times 0.084 + 1 \times 0.084 + 1 \times 0.084 + 1 \times 2.952 + 1 \times 0.084 \\ &\quad + 1 \times 0.084 + 0.084 \times 0.084 + 0.084 \times 0.084 + 0.084 \times 0.084 \\ &\quad + 0.084 \times 2.952 + 0.084 \times 0.084 + 2.952 \times 0.084 = 3.896\end{aligned}$$

By analogy with the Moreau–Broto autocorrelation descriptors, ATS, the GETAWAY descriptors (Table 2-6) have been defined, weighting each atom of the molecule by using physicochemical weights combined with the elements of **H** or **R** matrix, thus also accounting for the 3D features of the molecules.

The function $\delta(k; d_{ij}; h_{ij})$ used for the *H* indices (Eq. 41) is a Dirac-delta function defined as:

$$\delta(k; d_{ij}; h_{ij}) = \begin{cases} 1 & \text{if } d_{ij} = k \text{ and } h_{ij} > 0 \\ 0 & \text{if } d_{ij} \neq k \text{ or } h_{ij} \leq 0 \end{cases}$$

While the *HATS* indices (Eq. (39)) are calculated by use of the same formula as the Moreau–Broto ATS descriptors, the *H* indices exploit the off-diagonal elements of the **H** matrix, which can be either positive or negative. Therefore, the *H* indices have been defined by following the basic principles of spatial autocorrelation as above, however for a given *lag* (i.e. topological distance) the product of the atom weights is multiplied by the corresponding MIM value h_{ij} and only those contributions with a positive MIM value are considered. This means that, for a given atom i, only those atoms j at topological distance d_{ij} with a positive h_{ij} value have the chance to interact with the ith atom.

The weights used for the GETAWAY descriptors are those proposed for the calculation of the WHIM descriptors [31], i.e. atomic mass (m), atomic polarizability (p), atomic electronegativity (e), van der Waals atomic volume (v), plus the unit weight (u).

HATS, *H*, *R* and *maximal R* indices are molecular descriptors for structure–property correlations, but they can also be used as molecular profiles suitable for similarity/diversity analysis studies. These descriptors, as based on spatial autocorrelation, encode information on structural fragments and therefore seem to be particularly suitable for describing differences in congeneric series of molecules. Differently from the Moreau–Broto autocorrelations, GETAWAYs are geometrical descriptors encoding information on the effective position of substituents and fragments in the molecular space. They are, moreover, independent of molecule alignment and, to some extent, account also for information on molecular size and shape and for specific atomic properties.

Joint use of GETAWAY and WHIM descriptors is suggested, exploiting both the local information of the former set of descriptors and the holistic information of the latter. The GETAWAY descriptors have been evaluated on several data sets of pharmacological and environmental interest and their performance has been more than satisfactory [41].

2.5 3D Autocorrelation Descriptors

Spatial autocorrelation coefficients are frequently used in molecular modeling and QSAR to account for spatial distribution of molecular properties.

The simplest descriptor P for a molecular property is obtained by summing the (squared) atomic property values p_i. Mathematically (Eq. (48)):

$$P = \sum_{i=1}^{A} p_i^2 \tag{48}$$

where A is the number of atoms and P is the global property descriptor which depends on the kinds of atoms in the molecule and not on the molecular structure.

The spatial autocorrelation descriptors are an extension of this global property descriptor that combine chemical information given by property values in specified molecule regions and structural information. These are based on a conceptual dissection of the molecular structure and the application of an autocorrelation function to molecular properties measured in different molecular regions.

The first autocorrelation descriptors are those of Moreau and Broto [43] who applied an autocorrelation function to the molecular graph to measure the distribution of atomic properties on the molecule topology. They then extended this approach to 3D molecular geometry by replacing the topological distance by the interatomic distance r_{ij} [44].

Some ordered distance intervals are specified each defined by a lower and upper value of r_{ij}. All interatomic distances falling in the same interval are considered identical. For each distance interval the autocorrelation function AC_k is obtained by summing all the products of the property values of atoms i and j whose interatomic distance r_{ij} falls within the considered interval $[r_l, r_u]_k$ (Eq. (49)):

$$AC_k(r_l, r_u) = \sum_{i,j} p_i \cdot p_j \quad (r_l \leq r_{ij} \leq r_u) \tag{49}$$

Wagener et al. [45] have recently proposed 3D *autocorrelation descriptors of molecular surface properties*. These are calculated by applying autocorrelation functions to properties at distinct points on the molecular surface. The points are randomly distributed according to a preset density in order to model a continuous surface. Distance intervals are defined as above and the autocorrelation function AC_k is obtained by summing the products of property values at points i and j having a distance belonging to the considered distance interval (Eq. (50)):

$$AC_k(r_l, r_u) = \frac{1}{N_k} \sum_{i,j} p_i \cdot p_j \quad (r_l \leq r_{ij} \leq r_u) \tag{50}$$

where N_k is the number of distances belonging to the k interval.

As these autocorrelation values are influenced by how the molecule surface is defined, the van der Waals surface and all the van der Waals radii should be used; moreover, point densities equal to or greater than 5 points/Å^2 and distance intervals equal to or less than 1 Å have been suggested.

Autocorrelation values calculated for a number of distance intervals constitute a unique fingerprint of the molecule, and are thus suitable for similarity analysis of molecules. Figure 2-7 shows the autocorrelation vector of estradiol calculated by using MEP as the surface property.

3D autocorrelation descriptors have been shown to be useful in QSAR studies, because they are unique for a given geometry, are sensitive to changes in conformation and do not require any molecule alignment, being invariant to roto-translation.

A typical disadvantage of all the autocorrelation descriptors might be that the original information on the molecular structure or surface cannot be reconstructed.

2.6 3D-MoRSE and RDF Descriptors

3D-MoRSE (*3D-Molecule Representation of Structures based on Electron diffraction*) *descriptors* are based on the idea of obtaining information from the 3D atomic coordinates by use of the transform used in electron diffraction studies for prepar-

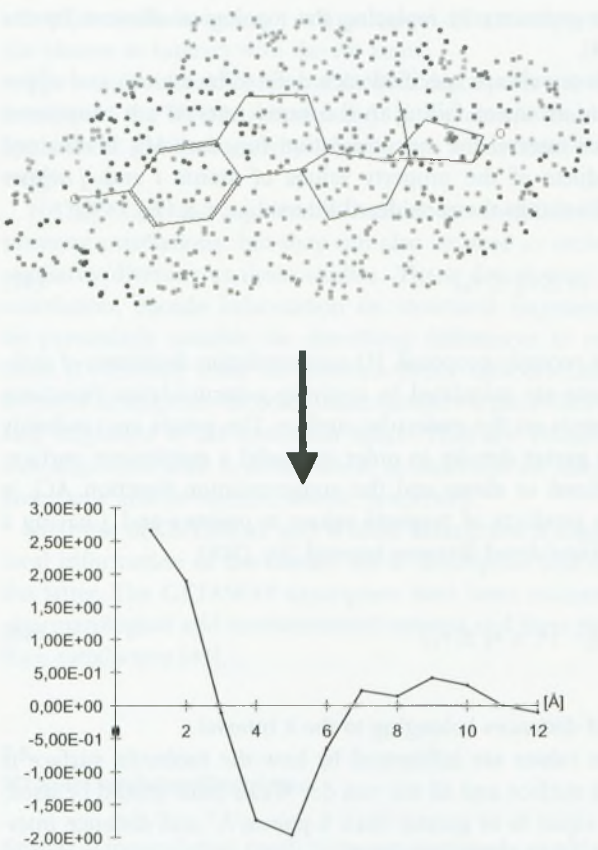

Fig. 2-7 Autocorrelation vector of estradiol calculated by using MEP as the surface property

ing theoretical scattering curves [46]. A generalized scattering function, called the molecular transform, can be used as the functional basis for deriving, from a known molecular structure, the specific analytical relationship of both X-ray and electron diffraction. The general molecular transform is (Eq. (51)):

$$G(s) = \sum_{i=1}^{A} f_i \cdot \exp(2\pi i \cdot \mathrm{r}_i \cdot s) \qquad (51)$$

where s represents the scattering in various directions by a collection of A atoms located at points r_i; f_i is a form factor taking into account the direction dependence of scattering from a spherical body of finite size. The scattering value, s, measures the scattering angle as (Eq. (52)):

$$s = 4\pi \cdot \sin(\vartheta/2)/\lambda \qquad (52)$$

where ϑ is the scattering angle and λ the wavelength of the electron beam.

This equation is usually used in the modified form suggested in 1931 by Wierl [47]. On substituting the form factor by the atomic property, w, considering the molecule to be rigid and setting the instrumental constant equal to unity, the following expression is obtained (Eq. (53)):

$$I(s) = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} w_i \cdot w_j \cdot \frac{\sin(s \cdot r_{ij})}{s \cdot r_{ij}} \tag{53}$$

where $I(s)$ is the scattered electron intensity, w is an atomic property, chosen as the atomic number, r_{ij} are the interatomic distances between the ith and jth atoms, and A is the number of atoms.

Soltzberg and Wilkins introduced a number of simplifications to obtain a binary code. Only the zero crossing of the $I(s)$ curve, i.e. the s values at which $I(s) = 0$ in the range 1–31 Å^{-1}, were considered. The s range is then divided into 100 equal intervals, each described by a binary variable equal to 1 if the interval contains a zero crossing, 0 otherwise. Thus, a code consisting of a 100-dimensional binary vector is obtained.

Gasteiger et al. [48] returned to the initial $I(s)$ curve and maintained the explicit form of the curve. For atomic weightings w, various physicochemical properties such as atomic mass, partial atomic charge, or atomic polarizability were considered. To obtain uniform length descriptors, the intensity distribution $I(s)$ is made discrete, its value being calculated at a sequence of evenly distributed values, e.g. 32 or 64 values in the range of 1–31 Å^{-1}. Clearly, the more values are chosen, the finer becomes the resolution in the representation of the molecule.

The MoRSE descriptors have been shown to have good modeling power for different biological and physicochemical properties and can be used even for the simulation of infrared spectra [49].

RDF (Radial Distribution Function) descriptors based on a radial distribution function have recently been proposed; they have some characteristics in common with the $I(s)$ function used to obtain the 3D-MoRSE descriptors. These descriptors are based on the geometrical interatomic distance and constitute a radial distribution function code.

Formally, the radial distribution function of an ensemble of A atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of radius R. The general form of the radial distribution function code (*RDF code*) is represented by (Eq. (54)):

$$g(R) = f \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} w_i \cdot w_j \cdot e^{-\beta \cdot (R - r_{ij})^2} \tag{54}$$

where f is a scaling factor, w are characteristic properties of the atoms i and j, r_{ij} is the geometrical distance between the ith and jth atoms, and A is the number of atoms [50]. The exponential term contains the interatomic distance r_{ij} and

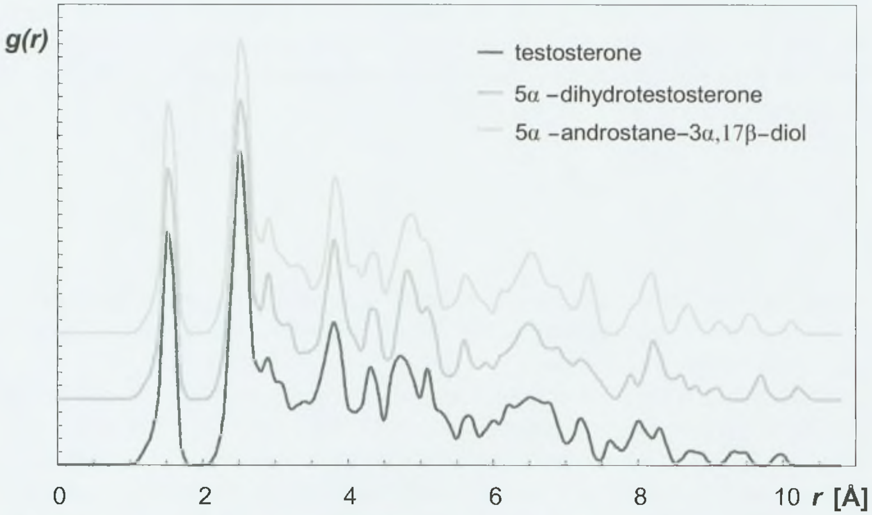

Fig. 2-8 RDF codes of testosterone, 5α-dihydrotestosterone, and 5α-androstane-3α,17β-diol

the smoothing term β, which defines the probability distribution of the individual interatomic distance; β can be interpreted as a temperature factor that defines the movement of atoms. $g(R)$ is generally calculated at a number of discrete points with defined intervals. An RDF code of 128 values was proposed; this was obtained by setting β in the range 100 to 200 Å^{-2} and a step size for R ca. 0.1–0.2 Å. RDF codes for testosterone, 5α-dihydrotestosterone, and 5α-androstane-3α,17β-diol are shown in Figure 2-8.

By including characteristic atomic properties w of atoms i and j, the RDF code can be used in different tasks to fit the requirements of the information to be represented. These atomic properties enable the discrimination of the atoms of a molecule for almost any property that can be attributed to an atom.

The radial distribution function in this form meets all the requirements for a 3D structure descriptor. It is independent of the number of atoms, i.e. the size of a molecule; it is unique regarding the three-dimensional arrangement of the atoms; and it is invariant against translation and rotation of the entire molecule. Additionally, the RDF code can be restricted to specific atom types or distance ranges to represent specific information in a certain three-dimensional structure space, e.g. to describe steric hindrance or the structure and/or activity properties of a molecule.

The RDF code is interpretable by using simple rule sets, and thus provides the possibility of converting the code back into the corresponding 3D structure. Besides information about interatomic distances in the entire molecule, the RDF code provides further valuable information, e.g. about bond distances, ring types, planar and non-planar systems and atom types. This is a most valuable consideration for computer-assisted code elucidation.

2.7 EVA and EEVA Descriptors

EVA (*EigenVAlue*) *descriptors* were recently proposed by Willett's group [51] as a new approach to extracting chemical structure information from mid- and near-infrared spectra. The approach is to use, as a multivariate descriptor, the vibrational frequencies of a molecule, a fundamental molecular property characterized reliably and easily from the potential energy function. The EigenVAlue (EVA) descriptors are a function of the eigenvalues obtained from the normal coordinate matrix; they correspond to the fundamental vibrational frequencies of the molecule, which can be calculated using standard quantum or molecular mechanical methods.

Because the number of vibrational normal modes varies with the number of atoms, A, in a molecule (actually $3A$–6 for a molecule without axial symmetry or $3A$–5 for a linear molecule), each set of eigenvalues will usually be of different dimensionality. Thus to obtain comparability among the molecules and uniform length descriptors, the frequency range typically chosen is from 0 to 4000 cm^{-1} to encompass the frequencies of all fundamental molecular vibrations, and the eigenvalues are projected on to a bounded frequency scale (BFS) where the vibrational frequencies are represented by points along this axis, affording a scale of fixed dimensionality. A Gaussian function of standard deviation σ is then centered at each eigenvalue projection over the BFS axis, resulting in a series of $3A$–6 (or $3A$–5) identical and overlapping Gaussians.

The value of the *EVA function* at any point x on the BFS axis is determined by summing the contributions from each and every one of the $3A$–6 (or $3A$–5) overlaid Gaussians at that point (Eq. (55)):

$$EVA_x = \sum_{i=1}^{3A\text{-}6} \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-(x-\lambda_i)^2/2\sigma^2} \tag{55}$$

where λ_i is the ith vibrational frequency (eigenvalue) for the molecular structure.

As the final step, the EVA function is sampled at fixed increments of L cm^{-1} along the BFS axis, this sampling results in the $4000/L$ values that define the EVA descriptor of uniform length.

The choice of σ defines the extent to which the fundamental vibrations overlap: σ values determine the number and extent to which vibrations of a particular frequency in one structure can statistically be related to those in the other structures (interstructural overlap); such values also govern the extent to which vibrations within the same structure may overlap at non-negligible values (intrastructural overlap).

In the original EVA method a fixed Gaussian standard deviation is used, meaning that each vibrational frequency is equally weighted before regression analysis. In the new EVA-GA method [52] the standard deviation can have localized values at different regions on the BFS. In some more detail the BFS is divided into a number of bins of equal size and a localized value of the standard deviation is

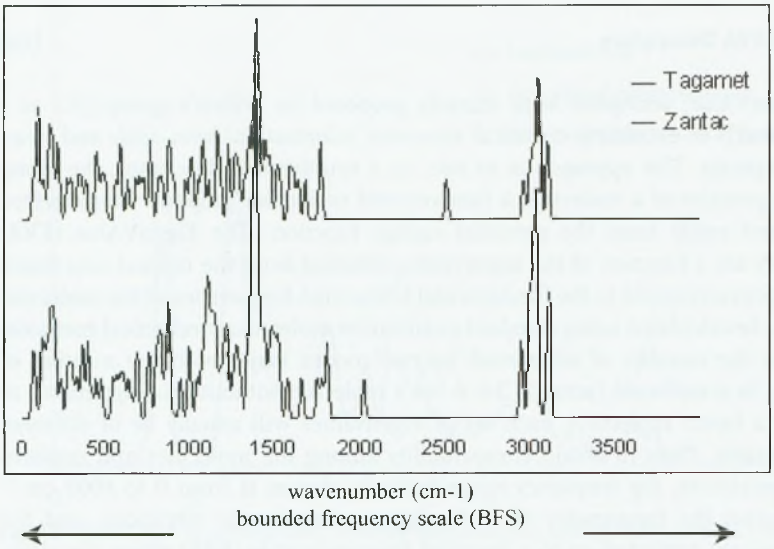

Fig. 2-9 Bounded frequency scale with superimposed EVA descriptors for two compounds.

associated with each bin. A frequency value falling within a bin is thus expanded using the associated local standard deviation. A genetic algorithm (GA) is used to search for optimum combinations of standard deviation values for each data set and modeled response. If a standard deviation value of zero is permitted, the GA also enables reduction of the descriptor number and thus simplified models.

With the frequency range fixed, the sampling term L determines the total number of EVA descriptor elements; L should be minimized, to reduce computational overhead, and maximized to catch all the useful information. The optimum value of L depends on the selected σ value; a general rule-of-thumb is to choose L so that it is $<2\sigma$.

A characteristic value of the Gaussian standard deviation σ is 10 cm^{-1} (range 1–25 cm^{-1}) and a characteristic value of the sampling increment L is 5 cm^{-1} (range 2–50 cm^{-1}); this results in 800 ($4000/L$) descriptor variables.

EVA descriptors are 3D-descriptors, independent of any molecular alignment, giving information about molecular size, shape and electronic properties. They are, moreover, only moderately dependent on conformation. They are mainly used as descriptors for QSAR and they have been shown to perform well with different data sets for prediction of biological responses [53, 54]. EVA descriptors have, moreover, been evaluated for use in similarity searching of structure databases [55]; however, the main drawback for this use is the overhead required to calculate the vibrational frequencies.

EEVA descriptors (or *Electronic EigenVAlue descriptors*) are a vector-descriptor proposed as a modification of EVA [56]. Semi-empirical molecular orbital energies, i.e. the eigenvalues of the Schrödinger equation, are used instead of the vibrational frequencies of the molecule.

Each molecular orbital energy of the molecule is first projected on to a bounded energy scale. EEVA descriptors are then defined by the same function (Eq. (55)) as the EVA descriptors, where MO energies E_i are used instead of vibrational frequencies λ_i. A fixed standard deviation σ equal to 0.1 eV and a sampling interval L set at $\sigma/2$ have been proposed.

This procedure provides a descriptor vector with dimensionality much lower than that of the EVA descriptor vector.

EEVA descriptors have been applied to reference QSAR data sets such as PCBs, PCDDs, PCDFs, and steroids [57]. Despite their good performance EEVA descriptors are not easily understandable, because the encoded electronic information cannot be taken straight back to molecular structure. The authors highlight that the current practical value of the EEVA descriptors relies almost solely on their predictive ability, which should always be carefully tested by use of validation techniques.

2.8 Conclusions

Several geometrical descriptors have been proposed in the last ten years, emphasizing the great interest of the scientific community in this approach to catching molecular information and the need for more sophisticated molecular descriptors useful for development of QSAR/QSPR models.

Molecular description based on geometry is mandatory when the conformational problem must be faced, although this increases the complexity and time of descriptor computation. For the same reason, the practicability of geometrical descriptors still remains an open problem for similarity/diversity analysis and screening of large databases.

References

1 R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley–VCH, Weinheim (GER), **2000**.

2 A. T. Balaban, *From Chemical Topology to Three-Dimensional Geometry*, A. T. Balaban (Ed.), Plenum Press, New York (NY), **1997**, 1–24.

3 M. Randic, *Stud. Phys. Theor. Chem.* **1988**, *54*, 101–108.

4 M. V. Diudea, D. Horvath, A. Graovac, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 129–135.

5 M. Randic, M. Razinger, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 140–147.

6 O. Mekenyan, D. Peitchev, D. Bonchev, N. Trinajstic, I. P. Bangov, *Arzneim. Forsch.* **1986**, *36*, 176–183.

7 B. Bogdanov, S. Nikolic, N. Trinajstic, *J. Math. Chem.*, **1989**, *3*, 299–309.

8 P. A. Bath, A. R. Poirrette, P. Willett, F. H. Allen, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 714–716.

9 G. W. Bemis, I. D. Kuntz, *J. Comput. Aid. Molec. Des.* **1992**, *6*, 607–628.

10 P. A. Bath, A. R. Poirrette, P. Willett, F. H. Allen, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141–147.

11 A. C. Good, I. D. Kuntz, *J. Comput. Aid. Molec. Des.* **1995**, *9*, 373–379.

12 R. D. Brown, Y. C. Martin, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.

13 M. V. Volkenstein, *Configurational Statistics of Polymeric Chains*, Wiley–Interscience, New York (NY), **1963**.

14 A. Y. Meyer, *J. Comput. Chem.* **1986**, *7*, 144–152.

15 P. J. Flory, *Statistical Mechanics of Chain Molecules*, Wiley–Interscience, New York (NY), **1969**.

16 A. R. Katritzky, L. Mu, V. S. Lobanov, M. Karelson, *J. Phys. Chem.* **1996**, *100*, 10400–10407.

17 M. D. Wessel, P. C. Jurs, J. W. Tolan, S. M. Muskal, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726–735.

18 C. Tanford, *Physical Chemistry of Macromolecules*, Wiley, New York (NY), **1961**.

19 D. G. Lister, J. N. Macdonald, N. L. Owen, *Internal Rotation and Inversion*, Academic Press, London (UK), **1978**.

20 G. A. Arteca, *Reviews in Computational Chemistry*, Vol. 9, K. B. Lipkowitz, D. Boyd (Eds), VCH, New York (NY), **1991**, 191–253.

21 D. D. Robinson, T. W. Barlow, W. G. Richards, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 939–942.

22 R. Kaliszan, *Quantitative Structure–Chromatographic Retention Relationships*, Wiley, New York (NY), **1987**.

23 G. M. Janini, K. Johnston, W. L. Zielinski Jr, *Anal. Chem.* **1975**, *47*, 670–674.

24 S. A. Wise, W. J. Bonnett, F. R. Guenther, W. E. May, *J. Chromatogr. Sci.* **1981**, *19*, 457–465.

25 E. R. Collantes, W. Tong, W. J. Welsh, *Anal. Chem.* **1996**, *68*, 2038–2043.

26 A. Verloop, *The STERIMOL Approach to Drug Design*, Marcel Dekker, New York (NY), **1987**.

27 P. C. Jurs, M. N. Hasan, P. J. Hansen, R. H. Rohrbaugh, *Physical Property Prediction in Organic Chemistry*, C. Jochum, M. G. Hicks, J. Sunkel (Eds.), Springer, Berlin (Germany), **1988**, 209–233.

28 N. Bodor, P. Buchwald, M.-J. Huang, *SAR & QSAR Environ. Res.* **1998**, *8*, 41–92.

29 A. Y. Meyer, *J. Comput. Chem.* **1988**, *9*, 18–24.

30 R. Todeschini, M. Lasagni, E. Marengo, *J. Chemom.* **1994**, *8*, 263–273.

31 R. Todeschini, P. Gramatica, *3D QSAR in Drug Design – Vol. 2*, H. Kubinyi, G. Folkers, Y. C. Martin (Eds.), Kluwer/ESCOM, Dordrecht (The Netherlands), **1998**, 355–380.

32 B. D. Silverman, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1470–1476.

33 J. C. Pinheiro, M. M. C. Ferreira, O. A. S. Romero, *Theochem* **2001**, *572*, 35–44.

34 T. Suzuki, K. Ide, M. Ishida, S. Shapiro, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 718–726.

35 P. Gramatica, N. Navas, R. Todeschini, *Chemom. Intell. Lab. Syst.* **1998**, *40*, 53–63.

36 Z. Daren, *Computers Chem.* **2001**, *25*, 197–204.

37 R. Todeschini, P. Gramatica, E. Marengo, R. Provenzani, *Chemom. Intell. Lab. Syst.* **1995**, *27*, 221–229.

38 P. Gramatica, V. Consonni, R. Todeschini, *Chemosphere* **1999**, *38*, 1371–1378.

39 P. Gramatica, M. Corradi, V. Consonni, *Chemosphere* **2000**, *41*, 763–777.

40 V. Consonni, R. Todeschini, M. Pavan, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682–692.

41 V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693–705.

42 L. Lovasz, J. Pelikan, *Period. Math. Hung.* **1973**, *3*, 175–182.

43 G. Moreau, P. Broto, *Nouv. J. Chim.* **1980**, *4*, 359–360.

44 P. Broto, G. Moreau, C. Vandycke, *Eur. J. Med. Chem.* **1984**, *19*, 66–70.

45 M. Wagener, J. Sadowski, J. Gasteiger, *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.

46 L. J. Soltzberg, C. L. Wilkins, *J. Am. Chem. Soc.* **1977**, *99*, 439–443.

47 K. Wierl, *Ann. Phys. (Leipzig)* **1931**, *8*, 521–564.

48 J. H. Schuur, P. Selzer, J. Gasteiger, *J. Am. Chem. Soc.*, **1996**, *36*, 334–344.

49 J. Schuur, J. Gasteiger, *Anal. Chem.* **1997**, *69*, 2398–2405.

50 M. C. Hemmer, V. Steinhauer, J. Gasteiger, *Vibrat. Spect.* **1999**, *19*, 151–164.

51 T. W. Heritage, A. M. Ferguson, D. B. Turner, P. Willett, *3D QSAR in Drug Design – Vol. 2*, H. Kubinyi, G. Folkers, Y. C. Martin (Eds.), Kluwer/ESCOM, Dordrecht (The Netherlands), **1998**, 381–398.

52 D. B. Turner, P. Willett, *J. Comput. Aid. Molec. Des.* **2000**, *14*, 1–21.

53 D. B. Turner, P. Willett, A. M. Ferguson, T. W. Heritage, *J. Comput. Aid. Molec. Des.* **1999**, *13*, 271–296.

54 D. B. Turner, P. Willett, *Eur. J. Med. Chem.* **2000**, *35*, 367–375.

55 C. M. Ginn, D. B. Turner, P. Willett, A. M. Ferguson, T. W. Heritage, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23–27.

56 K. Tuppurainen, *SAR & QSAR Environ. Res.* **1999**, *10*, 39–46.

57 K. Tuppurainen, M. Viisas, R. Laatikainen, M. Peräkylä, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 607–613.