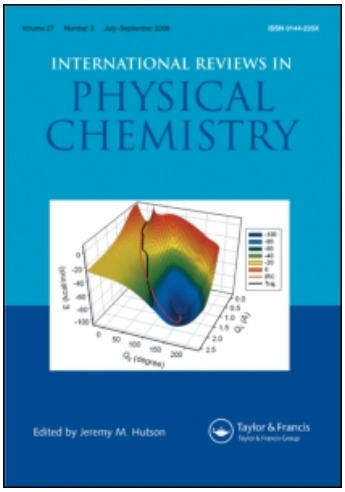

International Reviews in Physical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713724383

Quantum chemistry of macromolecular shape

Paul G. Mezey

Online publication date: 26 November 2010

To cite this Article Mezey, Paul G.(1997) 'Quantum chemistry of macromolecular shape', International Reviews in Physical Chemistry, 16: 3, 361 — 388
To link to this Article: DOI: 10.1080/014423597230226
URL: http://dx.doi.org/10.1080/014423597230226

INTERNATIONAL REVIEWS IN PHYSICAL CHEMISTRY, 1997, VOL. 16, NO. 3, 361–388

Quantum chemistry of macromolecular shape

by PAUL G. MEZEY

Mathematical Chemistry Research Unit, Department of Chemistry and Department of Mathematics and Statistics, University of Saskatchewan, 110 Science Place, Saskatoon, SK, Canada, S7N 5C9

Some of the new developments in the quantum-chemical study of macromolecular shapes are reviewed, with special emphasis on the additive fuzzy electron density fragmentation methods and on the algebraic-topological shape group analysis of global and local shape features of fuzzy three-dimensional bodies of electron densities of macromolecules. Earlier applications of these methods to actual macromolecules are reviewed, including studies on the anticancer drug taxol, the proteins bovine insulin and HIV protease, and other macromolecules. The results of test calculations establishing the accuracy of these methods are also reviewed. The spherically weighted affine transformation technique is described and proposed for the deformation of electron densities approximating the changes occurring in small conformational displacements of atomic nuclei in macromolecules.

1. Introduction

During the recent decades, quantum chemistry methods have provided unprecedented insight into the properties and behaviour of small molecules. In several areas of research the accuracy of quantum-chemical computational techniques compare well and even exceed the accuracy of experimental methods. Furthermore, modern quantum chemistry allows one to study many molecular problems where experimental information is not available. The development of some of the fundamental computational methods of quantum chemistry [1–10] and various experimental and theoretical studies on electron densities [11–24] have contributed to the evolution of the basic concepts of chemistry. The computational methodologies, primarily the Hartree–Fock–Roothaan–Hall molecular orbital technique [1–4, 10], various related methods developed for the treatment of electron correlation and, more recently, some of the computational techniques based on density functional methods [25–47], have become accessible to all chemists with a state-of-the-art desktop computer.

Whereas the early methodologies were rather limited to small molecules, some new developments in the representation of electronic densities and related properties of macromolecules allow an extension of many of the quantum chemistry approaches to large systems, including proteins. In this report a particular approach, the additive fuzzy density fragmentation (AFDF) method [48–52] is reviewed, where the emphasis is on electronic density obtained within the molecular orbital framework, but without the actual determination of a macromolecular wavefunction. This method can be applied for the construction of approximate macromolecular electron densities, density matrices, approximate energy relations, and approximate macromolecular forces.

The emphasis on electronic density $\rho(\mathbf{r})$, as opposed to a molecular wavefunction Ψ, is motivated by both fundamental quantum-chemical as well as practical computational considerations.

0144–235X/97 $12·00

Electronic density is an observable, whereas a molecular wavefunction is only a quantum-mechanical tool. *Electron density is reality*, *whereas a wavefunction is only a formal 'square root' of reality*. In a molecule there is nothing else but a nuclear distribution and an electron distribution. It appears sensible to focus on the actual physical entity, electron density, that describes fully the molecular shape and also reflects the actual nuclear arrangement. All information concerning a given molecular conformation is contained in the electron density.

From the practical computational perspective, molecular electron densities have 'nicer' properties than molecular wavefunctions. The exponential convergence of $\rho(\mathbf{r})$ to zero is very rapid with the distance from the nearest nucleus. Furthermore, using the AFDF approach, the electronic density $\rho(\mathbf{r})$ can be decomposed into fuzzy electron density fragments in an exactly additive manner, while preserving the same rapid convergence properties. By contrast, the molecular wavefunction Ψ has less uniform convergence properties and Ψ is not easily decomposable in a strictly additive manner into 'local wavefunctions' representing molecular fragments.

The macromolecular quantum chemistry methods based on the AFDF approach also provide the tools for the extension of earlier quantum-chemical shape analysis methods to large molecules, including proteins.

In section 2, the main concepts and methods of the application of the AFDF approach are described briefly, including the results of numerical tests establishing the accuracy of the approximations employed. These tests indicate that the AFDF methodology is suitable for providing *ab initio* quality results at a fraction of the computational cost of conventional calculations. Several earlier macromolecular applications are reviewed, including studies on various polypeptides, the proteins bovine insulin and HIV protease, and other macromolecules such as the anticancer drug taxol [53–57].

The AFDF methods described in this section include the molecular electron density loge assembler (MEDLA) method [53–59], the adjustable density matrix assembler (ADMA) techniques [50, 52, 60–63], the ADMA–FORCE approach to macromolecular forces [61, 62], as well as the related quantum-chemical representations of local molecular moieties, such as functional groups [64–67].

In section 3, an auxiliary technique is described, suitable for the calculation of small deformations of approximate electron densities of macromolecules. This technique, the spherically weighted affine transformation (SWAT) method [68], is suggested for the rapid estimation of macromolecular electron densities if an electron density is available for a macromolecular nuclear arrangement only slightly different from the actual conformation considered.

In section 4, the shape group method (SGM), an earlier quantum-chemical shape analysis technique [65, 69–73] is extended to large systems, with emphasis on the global features of macromolecular electron densities and on the detailed shapes of their local regions.

2. Additive fuzzy density fragmentation and density matrix assembler methods

We shall use the $\rho(\mathbf{r}, K)$ notation for the self-consistent field (SCF) linear-combination-of-atomic orbitals (LCAO) *ab initio* electronic density of a molecule taken at some fixed nuclear conformation K, where $\mathbf{r}$ is the three-dimensional position vector variable. Using the conventional Hartree–Fock–Roothaan–Hall formalism, we assume that this electron density is expressed in terms of a basis set $\varphi(K)$ of atomic orbitals $\varphi_i(\mathbf{r}, K)$ $(i = 1, 2, \ldots, n)$ used for the expansion of the molecular wavefunction,

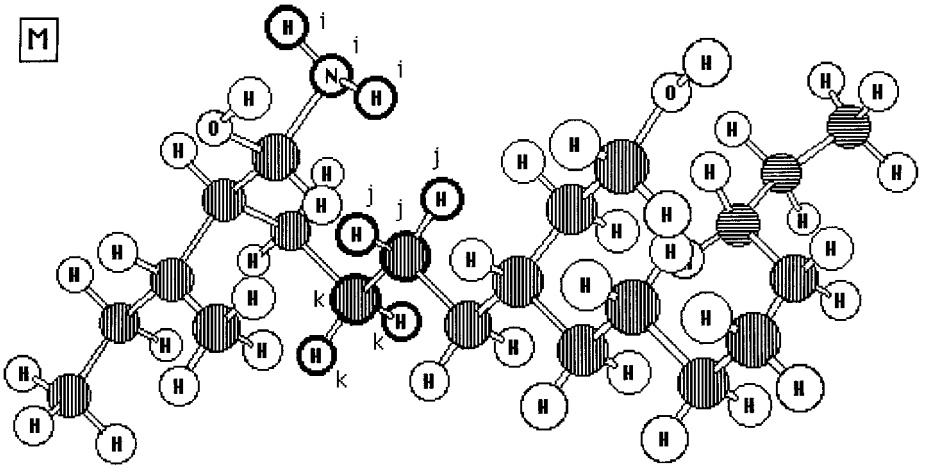

Figure 1. Example of a large molecule M and three nuclear families f_i, f_j and f_k, serving as 'anchor points' for fuzzy electron density fragments $F_i(K)$, $F_j(K)$ and $F_k(K)$ respectively. These local fragment densities correspond to that of an NH_2 group, and two CH_2 groups, where the latter two groups differ as a consequence of their different local surroundings.

where n is the number of orbitals. Note that the location of basis functions is dependent on the nuclear configuration K, and this dependence is explicitly indicated in the notation. The $n \times n$ dimensional density matrix determined for the given nuclear configuration K is noted $\mathbf{P}(\varphi(K))$. The corresponding electronic density $\rho(\mathbf{r}, K)$ is computed as

$$\rho(\mathbf{r}, K) = \sum_{i=1}^{n} \sum_{j=1}^{n} P_{ij}(\varphi(K))\, \varphi_i(\mathbf{r}, K)\, \varphi_j(\mathbf{r}, K). \tag{1}$$

This equation provides a natural introduction for the AFDF principle [48–52] that, in turn, provides the basis for generating 'assembled' electron densities [53–59] and 'assembled' density matrices [50, 52, 60–63].

2.1. *The additive fuzzy density fragmentation principle*

The simplest illustration as well as implementation of the AFDF principle is the Mulliken–Mezey scheme [48–52], that is the approach used in the MEDLA method of Walker and Mezey [53–57] as well as in the ADMA macromolecular density matrix method of Mezey [50, 52, 60–63]. This fuzzy density fragmentation approach has been motivated by Mulliken's [5, 6] population analysis and charge assignment scheme.

The basis of the general Mulliken–Mezey AFDF scheme is a subdivision of the set of nuclei of the molecule M into m mutually exclusive families denoted by $f_1, f_2, \ldots, f_k, \ldots, f_m$. These nuclear families serve as reference and as formal 'anchor' points for a set of m fragment density functions $\rho^1(\mathbf{r}, K), \rho^2(\mathbf{r}, K), \ldots, \rho^k(\mathbf{r}, K), \ldots, \rho^m(\mathbf{r}, K)$. These density functions correspond to the actual additive fuzzy density fragments $F_1(K)$, $F_2(K), \ldots, F_k(K), \ldots, F_m(K)$.

These ideas are illustrated by the example of a 'macromolecule' M shown in figure 1. Only three of the nuclear families are highlighted, families f_i, f_j and f_k, serving as 'anchor points' for fuzzy electron density fragments $F_i(K)$, $F_j(K)$ and $F_k(K)$ respectively. These fragment densities correspond to that of a NH_2 group, and two CH_2 groups, where the latter two have different local surroundings.

In order to generate local fuzzy electron densities assigned to nuclear families

within an additive density fragmentation scheme, it is convenient to define a formal membership function $m_k(i)$ which indicates, if a given atomic orbital (AO) basis function $\varphi_i(\mathbf{r}, K)$ belongs to the set of AOs centred on a nucleus of family f_k,

$$m_k(i) = \begin{cases} 1 \text{ if AO } \varphi_i(\mathbf{r}) \text{ is centred on one of the nuclei of set } f_k, \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

Within the general Mulliken–Mezey AFDF scheme [48–52] the elements $P^k_{ij}(\varphi(K))$ of the $n \times n$ fragment density matrix $\mathbf{P}^k(\varphi(K))$ of the kth fragment $F_k(K)$ are defined as

$$P^k_{ij}(\varphi(K)) = [m_k(i)\, w_{ij} + m_k(j)\, w_{ji}]\, P_{ij}(\varphi(K)), \tag{3}$$

where for the w_{ij} and w_{ji} weighting factors

$$w_{ij} + w_{ji} = 1, \quad w_{ij},\ w_{ji} > 0, \tag{4}$$

hold. In the simplest version of the Mulliken–Mezey AFDF scheme the choice $w_{ij} = w_{ji} = 0{\cdot}5$ is taken, that is

$$P^k_{ij}(\varphi(K)) = 0{\cdot}5[m_k(i) + m_k(j)]\, P_{ij}(\varphi(K)). \tag{5}$$

This density matrix 'fragmentation' formula represents an approach equivalent to Mulliken's [5, 6] population analysis without integration, providing a justification for the terminology. Based on the more general scheme of equations (3) and (4), alternative choices for the weighting schemes have also been proposed [48–52].

Note that, for each index pair (i,j) the elements $P^k_{ij}(\varphi(K))$ of the AFDF fragment density matrices $\mathbf{P}^k(\varphi(K))$ are strictly additive:

$$P_{ij}(\varphi(K)) = \sum_{k=1}^{m} P^k_{ij}(\varphi(K)). \tag{6}$$

Consequently, the AFDF fragment density matrices $\mathbf{P}^k(\varphi(K))$ are also strictly additive and their sum is the density matrix $\mathbf{P}(\varphi(K))$ of the molecule M:

$$\mathbf{P}(\varphi(K)) = \sum_{k=1}^{m} \mathbf{P}^k(\varphi(K)). \tag{7}$$

Using these fragment density matrices $\mathbf{P}^k(\varphi(K))$, the additive fuzzy density fragments $\rho^k(\mathbf{r}, K)$ are defined as

$$\rho^k(\mathbf{r}, K) = \sum_{i=1}^{n} \sum_{j=1}^{n} P^k_{ij}(\varphi(K))\, \varphi_i(\mathbf{r}, K)\, \varphi_j(\mathbf{r}, K),\ k = 1, 2, \ldots, m. \tag{8}$$

According to equation (1), the molecular electron density $\rho(\mathbf{r}, K)$ depends linearly on the matrix elements $P_{ij}(\varphi(K))$; consequently, the exact additivity properties (6) and (7) for the fragment density matrices $\mathbf{P}^k(K)$ expressed with reference to basis set $\varphi(K)$ imply that the $\rho^k(\mathbf{r}, K)$ fuzzy fragment densities are, indeed, additive, and their sum is equal to the density $\rho(\mathbf{r}, K)$ of the molecule M of nuclear configuration K:

$$\rho(\mathbf{r}, K) = \sum_{k=1}^{m} \rho^k(\mathbf{r}, K). \tag{9}$$

Equation (9) provides a valid realization of an AFDF scheme.

The individual fuzzy fragment densities $\rho^k(\mathbf{r}, K)$ can be used for a local shape analysis of molecular moieties and functional groups [64–67].

Another important application of this AFDF scheme is in the construction of approximate electron densities of large molecules. If the size of a large molecule, for example a protein, renders a direct application of the Hartree–Fock–Roothaan–Hall

method impractical, then the AFDF approach can be used to circumvent the difficulties.

The first step is the classification of the nuclei of the large molecule M, referred to as the 'target molecule', into m mutually exclusive nuclear families $f_1, f_2, \ldots, f_k, \ldots, f_m$. Our goal is to generate a family of fuzzy density fragments $F_1(K), F_2(K), \ldots, F_k(K), \ldots, F_m(K)$, with 'anchor points' the respective nuclear families, where a simple superposition of these fuzzy density fragments provides an approximation to the macromolecular electron density of M.

Such fuzzy fragment densities F_k of M can be computed indirectly, if the AFDF scheme is applied to standard Hartree–Fock or post-Hartree–Fock electron densities obtained for a set of m small 'parent' molecules $M_1, M_2, \ldots, M_k, \ldots, M_m$, where each parent molecule M_k contains the respective nuclear family f_k with the same local nuclear geometry and the same local surroundings as they are found in the large target molecule M.

These ideas are illustrated in figure 2, where only three of the 'parent' molecules $M_i = M(F_i)$, $M_j = M(F_j)$ and $M_k = M(F_k)$ of the 'target' molecule M of figure 1 are shown. Within these parent molecules, the local arrangements and local surroundings of nuclear families f_i, f_j and f_k are the same as found in the large 'target' molecule M shown in figure 1. The AFDF technique can be applied to the Hartree–Fock electron densities of the small parent molecules, using a fragmentation where within each parent molecule $M_k = M(F_k)$ the respective nuclear family f_k is identified as an actual set of 'anchor points'. If each fuzzy electron density fragment F_k is taken from the respective parent molecule $M_k = M(F_k)$, then, by superimposing all these fragments, an approximate macromolecular electron density of target molecule M is obtained.

Evidently, the accuracy of this approach depends on the reproducibility of the local surroundings of each macromolecular density fragment within the parent molecules, which can be improved to any desired accuracy by increasing the size of the parent molecules.

In practice, a high-quality *ab initio* calculation is carried out for a 'custom-made' model of a parent molecule M_k that contains the density fragment $F_k(K_k)$ within local surroundings that matches the local surroundings of fragment $F_k(K)$ of the target molecule M within a large enough 'coordination shell'. The actual nuclear configuration K of macromolecular fragment $F_k(K)$ as well as its coordination shell are exactly reproduced by the nuclear configuration K_k and the coordination shell of the density fragment $F_k(K_k)$ within the 'custom-made' parent molecule M_k. Detailed test calculations have confirmed [53, 54, 57] that, in practice, a coordination shell of 4–5 Å thickness in each parent molecule is sufficient to generate local fragment electron densities at a level of accuracy that reproduces the results of conventional 6-31G** *ab initio* calculations of a target molecule M better than conventional *ab initio* calculations using smaller Gaussian basis sets such as 6-31G [57].

Although the method is rather simple, its accuracy is not surprising; all the local interactions within the fuzzy density fragment F_k and between the density fragment and the surrounding coordination shell within the macromolecule M and within the small parent molecule M_k are identical. Consequently, the fuzzy fragment electron density $\rho^k(\mathbf{r}, K_k)$ obtained for fragment F_k by a high-quality *ab initio* calculation for the 'custom-made' small parent molecule M_k is a good approximation of the density $\rho^k(\mathbf{r}, K)$ of fragment F_k in the large molecule M. Both fragment electron densities $\rho^k(\mathbf{r}, K_k)$ and $\rho^k(\mathbf{r}, K)$ decrease exponentially with the distance from the nearest nucleus of the nuclear family f_k (in the actual calculations, this exponential decay is approximated by

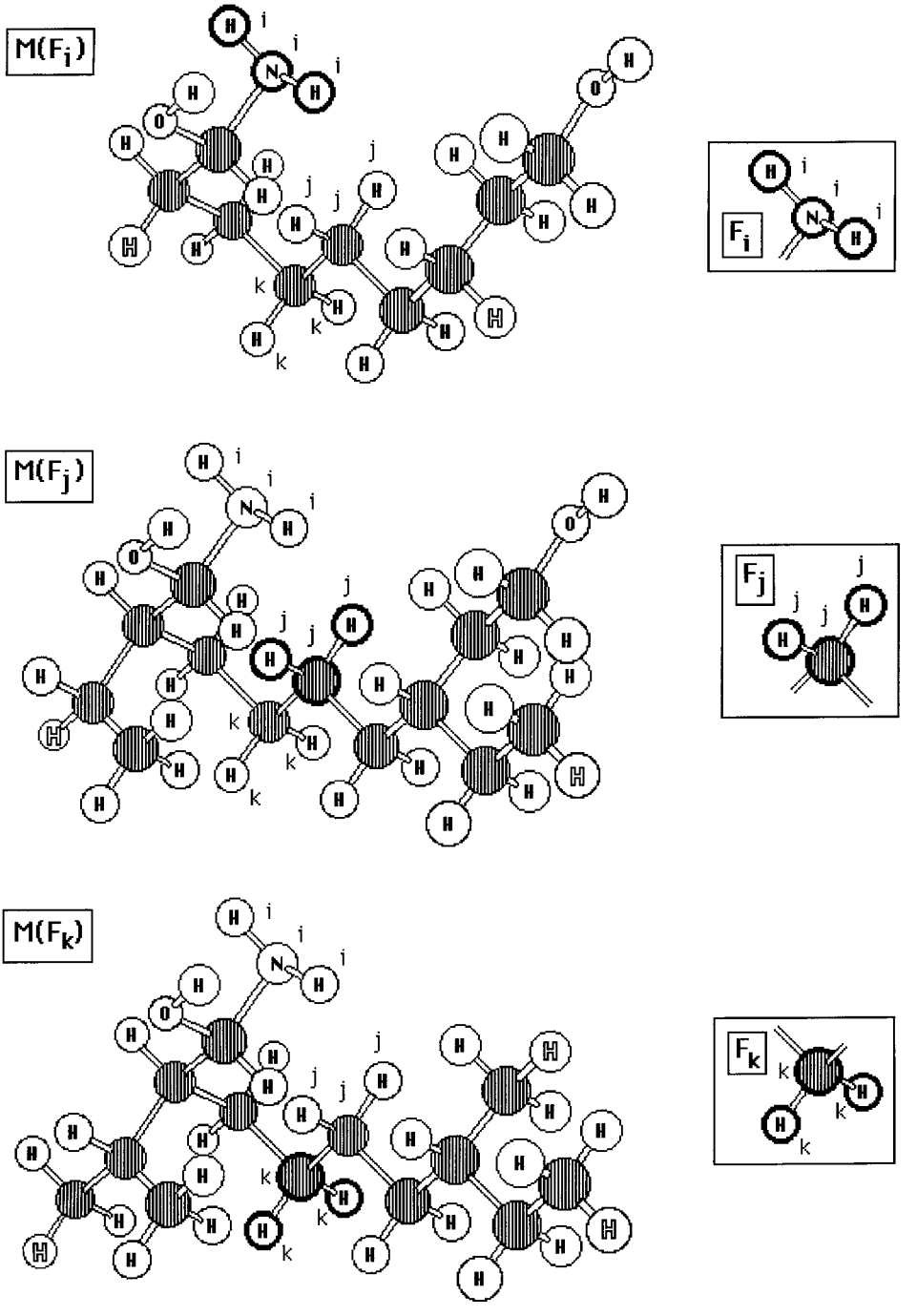

Figure 2. Three of the 'parent' molecules $M_i = M(F_i)$, $M_j = M(F_j)$, and $M_k = M(F_k)$ of the 'target' molecule M of figure 1 are shown. The local arrangements and local surroundings of nuclear families f_i, f_j and f_k, within parent molecules $M(F_i)$, $M(F_j)$ and $M(F_k)$ respectively are the same as in the target molecule M shown in figure 1. After computing the Hartree–Fock electron densities of these small parent molecules, within each parent molecule $M_k = M(F_k)$ the respective nuclear family f_k can be selected as an actual set of

multiple Gaussian functions). The method ensures a high degree of similarity between these fragment densities even if the matching surroundings are rather limited in size. A series of numerical tests of the simplest Mulliken–Mezey scheme has shown a high degree of similarity between electron densities computed using conventional *ab initio* techniques and methods based on the AFDF principle [53, 54, 57]. The actual deviation between $\rho^k(\mathbf{r}, K_k)$ and $\rho^k(\mathbf{r}, K)$ can be reduced to less than any small positive threshold; the accuracy of the method can be controlled by taking a large enough 'coordination shell' of matching surroundings within the parent molecule M_k.

Whereas for the MEDLA approach no restriction on basis set orientation is required, for a concise discussion of the MEDLA and ADMA methods, we assume that the electron densities of the large target and small parent molecules are represented by wavefunctions which are expressed using identical AO basis functions centred on corresponding nuclei. Also note that, for the actual target molecule M, no such wavefunction is calculated; however, the macromolecular density matrix that is constructed for M by the ADMA technique is compatible with such a wavefunction.

2.2. *The molecular electron density 'loge' assembler method*

The first practical implementation of the AFDF scheme for the construction of macromolecular electron densities was the MEDLA method (also referred to as molecular electron density 'lego' assembler) of Walker and Mezey [53–57], using a numerical electron density fragment database of pre-calculated custom-made electron density fragments. As numerous tests have demonstrated [53, 54, 57], the MEDLA method produces *ab initio* quality, in fact, nearly 6-31G** quality electron densities for large molecules. In particular, the test results justify the claim of '*ab initio* quality'.

In these tests [53, 54, 57], the electron densities of several molecules of moderate size were computed both by traditional *ab initio* methods and by the AFDF approach, using fragments from smaller parent molecules. The electron density results of the conventional *ab initio* Hartree–Fock–Roothaan–Hall SCF method using Gaussian basis sets ranging from STO 3G to 6-31G** were compared with one another and with the AFDF results obtained using the MEDLA method with fragments generated from 6-31G** calculations of the smaller parent molecules.

Specific tests included the following:

(*a*) detailed comparisons of electron densities obtained for the amino acid β-alanine [53];
(*b*) detailed comparisons of electron densities obtained for the model peptide system of glycyl-alanine [54];
(*c*) test of the reproducibility of the electron density of a hydrogen bond in a helical tetrapeptide [54];
(*d*) test of the reproducibility of a nonbonded interaction between a sulphur atom and a phenyl ring in a molecular fragment from the pentapeptide metenkephalin [54];
(*e*) test of reproducibility of aromatic rings and substituent effects in a series of aromatic molecules [57].

'anchor points'. By applying the AFDF technique and by taking each fuzzy electron density fragment F_k from the respective parent molecule $M_k = M(F_k)$, a simple superposition of all these fuzzy density fragments leads to an approximate macromolecular electron density of target molecule M.

The conventional *ab initio* Hartree–Fock–Roothaan–Hall SCF results at the 6-31G** basis set level were used as reference, since this was the basis set used for the generation of fuzzy density fragments. Clearly, an AFDF technique using density fragments obtained at the 6-31G** level cannot be expected to perform better than a direct application of the conventional *ab initio* method at the same basis set level.

In all these tests, the MEDLA method performed consistently better than conventional *ab initio* computations at any of the basis set levels tested except 6-31G**, the level used as reference. Comparisons included direct point-by-point comparisons throughout three-dimensional density grids, as well as integrated similarity measures such as the Carbó quantum similarity index [57]. According to these test results, the claim of '*ab initio* quality' is justified.

The AFDF methodology was applied to calculate *ab initio* quality electron densities for a series of macromolecules [54–56]. Justified by their exceptional biochemical importance, emphasis was placed on amino acids, peptides and proteins [54–56]. In particular, the computation of the electron densities of several proteins was completed, including crambin [54], bovine insulin [55], the gene-5 protein (g5p) of bacteriophage M13 [54], and the HIV-1 protease monomer, a protein of 1564 atoms in 99 amino acid residues [56]. These macromolecular electron densities have been calculated at the MEDLA 6-31G** level, that is using fragment densities obtained from custom-made parent molecules at the standard *ab initio* 6-31G** level. The resolution of the calculated electron densities exceeds the resolution of current experimental techniques, such as X-ray crystallography, by about two orders of magnitude. The MEDLA method serves as a 'computational microscope', providing detailed images of the fuzzy bodies of large molecules in any desired conformation.

2.3. *The adjustable density matrix assembler method*

The ADMA method is a more advanced application of the AFDF approach, which focuses on the fragment density matrices $\mathbf{P}^k(\varphi(K_k))$. The ADMA technique does not require a numerical fragment density database; only a more concise database of the fragment density matrices and basis set information are needed. If for the various fragment density matrices a mutual compatibility condition is satisfied, then these matrices can be assembled into an approximate macromolecular density matrix $\mathbf{P}(\varphi(K))$, which represents the same level of accuracy as a MEDLA numerical electron density generated on an ideal infinite-resolution grid. The macromolecular density matrix, combined with the basis set information, can be used for the computation of electron densities, and macromolecular density matrices are advantageous if our goal is the estimation of molecular properties other than electron density.

Within the general AFDF scheme, the mutual compatibility requirements for a family of additive fragment density matrices $\mathbf{P}^k(\varphi(K_k))$ obtained from small parent molecules M_k involves two additional conditions.

(*a*) Fragment AO basis set orientation condition. *All the fragment density matrices* $\mathbf{P}^k(\varphi(K_k))$ *should refer to local coordinate systems where the coordinate axes are parallel to and have the same orientation as the reference axes of a common macromolecular coordinate system.*

(*b*) Compatible target–parent fragmentation condition. *If the set of nuclei of the target molecule M are classified into m families,* $f_1, f_2, \ldots, f_k, \ldots, f_m$, *then each parent molecule* M_k *may contain only complete nuclear families* $f_{k'}$ *from the large target molecule M.*

Condition (*a*) can always be satisfied by a simple similarity transformation of a fragment density matrix $\mathbf{P}^k(\varphi(K_k))$ using a suitable orthogonal transformation matrix $\mathbf{T}^{(k)}$ of the AO sets.

Similarly, condition (*b*) can also be satisfied for any target macromolecule M, by an appropriate choice of nuclear families f_k for the various fragments and by a suitable choice of the 'coordination shells' of parent molecules M_k.

The general AFDF approach, combined with the two mutual compatibility conditions, is referred to as the MC-AFDF approach. These conditions imply that the AO basis functions centred at nuclei of a family f_k are the same in all parent molecules where family f_k occurs, independently of the role of f_k as the central family or a family in the 'coordination shell' within the parent molecule M_k.

In some of the parent molecules, the peripheral regions of the coordination shells may have some 'dangling bonds'; in these cases the parent molecule M_k may contain some additional peripheral hydrogen nuclei (or possibly other nuclei) linked to these formal bonds. For each parent molecule M_k, the AOs centred on these extra nuclei are listed at the end of the AO list of the local basis set. Since these extra nuclei are at large distances from the 'central' nuclear set f_k, their effect on the actual fragment density $\rho^k(\mathbf{r}, K_k)$ is negligible. Consequently, all contributions of the AOs of the additional, peripheral nuclei to the fragment density matrix $\mathbf{P}^k(\varphi(K_k))$ are ignored, and the corresponding rows and columns of the actual parent molecule density matrix $\mathbf{P}(K_k)$ are not included in the fragment density matrix $\mathbf{P}^k(\varphi(K_k))$. That is, the contribution of the fragment density matrix $\mathbf{P}^k(\varphi(K_k))$ to the generalized Mulliken–Mezey MC-AFDF scheme of the ADMA method involves only the orbital indices of the 'central' nuclear set f_k and the complete nuclear sets $f_{k'}$ which are part of the coordination shell used to reproduce the local macromolecular surroundings of set f_k within the parent molecule M_k.

The implementation of the MC-AFDF scheme of the ADMA method involves extensive index manipulation of various fragment density matrix contributions.

We denote the number of AOs in the nuclear families $f_1, f_2, \ldots, f_k, \ldots$, and f_m of the target macromolecule M by $n_1, n_2, \ldots, n_k, \ldots$, and n_m respectively. For each pair $(f_k, f_{k'})$ of nuclear families the quantity

$$c_{k'k} = \begin{cases} 1, \text{ if nuclear family } f_{k'} \text{ is present in parent molecule } M_k, \\ 0 \text{ otherwise,} \end{cases} \tag{10}$$

is defined. For each AO, $\varphi(\mathbf{r})$ three types of index notation are used. If the serial number of $\varphi(\mathbf{r})$ is b in the AO set

$$\{\varphi_{a,k'}(\mathbf{r})\}_{a=1}^{n_{k'}} \tag{11}$$

of nuclear family $f_{k'}$, then the notation $\varphi_{b,k'}(\mathbf{r})$ is used. With reference to the fragment density matrix $\mathbf{P}^k(\varphi(K_k))$, the jth AO of the basis set

$$\{\varphi_i^k(\mathbf{r})\}_{i=1}^{n_{P^k}} \tag{12}$$

of the kth fragment density matrix $\mathbf{P}^k(\varphi(K_k))$is denoted by $\varphi_j^k(\mathbf{r})$, where the total number of these AOs is denoted by n_{P^k}:

$$n_{P^k} = \sum_{k'=1}^{m} c_{k'k} n_{k'}. \tag{13}$$

With reference to the macromolecular density matrix $\mathbf{P}(K)$, the AO of serial index y in the AO set

$$\{\varphi_x(\mathbf{r})\}_{x=1}^{n} \tag{14}$$

of the density matrix $\mathbf{P}(K)$ of the target macromolecule M is denoted by $\varphi_y(\mathbf{r})$. For each AO,

$$\varphi_{a,\,k'}(\mathbf{r}) = \varphi_i^k(\mathbf{r}) = \varphi_x(\mathbf{r}), \tag{15}$$

the index x is determined by the index a in family k' as follows:

$$x = x(k', a, f) = a + \sum_{b=1}^{k'-1} n_b, \tag{16}$$

where the symbol f in $x(k', a, f)$ indicates that k' and a refer to a nuclear family.

From the element index i and serial index k of fragment density matrix $\mathbf{P}^k(\varphi(K_k))$, the index x can be determined as follows. Three quantities are defined for each index k and nuclear family $f_{k''}$ for which $c_{k''k} \neq 0$:

$$a'_k(k'', i) = i - \sum_{b=1}^{k''} n_b c_{bk}, \tag{17}$$

$$k' = k'(i, k) = \min\{k'' : a'_k(k'', i) \leqslant 0\} \tag{18}$$

and

$$a_k(i) = a'_k(k', i) + n_{k'}. \tag{19}$$

The AO index $x = x(k, i, P)$ in the density matrix $\mathbf{P}(K)$ of target molecule M is determined from indices i and k using index k' and the function $x(k', a, f)$:

$$x = x(k, i, P) = x(k', a_k(i), f). \tag{20}$$

Here the distinguishing symbol P in the index function $x(k, i, P)$ indicates that k and i refer to the fragment density matrix $\mathbf{P}^k(\varphi(K_k))$.

Using these index assignments, the density matrix $\mathbf{P}(K)$ of the target macromolecule M is assembled by an iterative procedure:

$$P_{x(k,\,i,\,P),\;y(k,\,j,\,P)}(K) = P_{x(k,\,i,\,P),\;y(k,\,j,\,P)}(K) + P_{ij}^k(K_k), \tag{21}$$

where only the non-zero elements of each fragment density matrix $\mathbf{P}^k(\varphi(K_k))$ are used. Since the fragment density matrices $\mathbf{P}^k(\varphi(K_k))$ are typically sparse, this approach offers some computational savings.

The size of parent molecules M_k is bounded; hence there is a bound on the computer time needed for their *ab initio* calculation as well as for the index reassignment for elements of each fragment density matrix. This bound is independent of the number of density fragments within the target macromolecule, consequently, the MC-AFDF ADMA method requires computer time that depends linearly on the number of fragments and on the size of the target macromolecule M.

The accuracy of the method, compared with a conventional *ab initio* calculation, depends on the size of the coordination shell in each parent molecule. By taking large enough parent molecules M_k, the difference between the MC-ADMA density matrix $\mathbf{P}(K)$ and the ideal conventional *ab initio* density matrix of the given basis set can be reduced below any positive threshold. Satisfactory accuracy can be obtained if a formal 'coordination shell' of approximately 4–5 Å thickness is used in each parent molecule M_k.

If the size of the coordination shell is insufficient, then the ADMA density matrix $\mathbf{P}(K)$ may deviate from idempotency and produce electron densities which do not exactly correspond to the right integral number of electrons. As the direct tests [53, 54, 57] on electron densities indicate, these errors may be negligible locally, but for macromolecules they may accumulate and for a protein of well over a thousand atoms

the error of the overall integrated electronic charge may be as much as half an electron. However, both idempotency and charge conservation can be restored using standard density matrix methods [74, 75] followed by a simple scaling [56]. Similar deviations from idempotency may occur if one attempts to compare and approximate electron densities at slightly displaced nuclear configurations. Some aspects of these two problems can be studied from a common perspective using methods [61, 62] based on the Löwdin [76–78] transforms. These methods are analogous to a technique employed in quantum crystallography for generating density matrices from experimental electron densities [79].

The idempotency condition for a density matrix $\mathbf{P}(K)$ expressed in terms of a non-orthonormal AO basis $\varphi(K)$ is given by the matrix product $\mathbf{P}(K)\mathbf{S}(K)\mathbf{P}(K)$ where $\mathbf{S}(K)$ is the overlap matrix for the AO basis. In this way, a formal product operation $*$ between density matrices can be defined using the overlap matrix, and the idempotency condition can be written as

$$\mathbf{P}(K)*\mathbf{P}(K) = \mathbf{P}(K). \tag{22}$$

If the AOs of a basis set $\varphi(K)$ are centred on the nuclei, then the main effect of a small conformation change $K \rightarrow K'$ of the nuclei is the displacement of the AO basis functions. If the ΔK displacement is small, then a simple approximation of the electronic density $\rho(\mathbf{r}, K')$ at the displaced nuclear arrangement K' can be obtained [52, 60–63] by taking the same density matrix $\mathbf{P}(K)$ and using it with the set of displaced AOs $\varphi_i(\mathbf{r}, K')$ at the new nuclear locations:

$$\rho_{\text{appr}}(\mathbf{r}, K') = \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}(\varphi(K))\,\varphi_i(\mathbf{r}, K')\,\varphi_j(\mathbf{r}, K'). \tag{23}$$

The error of this rather simplistic approach is surprisingly small.

If higher accuracy is needed, then an alternative method [52, 60–63] can be used, based on the Löwdin [76–78] transform and analogous to a method used in quantam crystallography for the construction of approximate 'experimental' density matrices from X-ray diffraction results [79]. The Löwdin transform of a density matrix $\mathbf{P}(K)$ involves pre- and post-multiplication by the matrix $\mathbf{S}(K)^{1/2}$:

$$\mathbf{S}(K)^{1/2}\mathbf{P}(K)\mathbf{S}(K)^{1/2}. \tag{24}$$

The resulting matrix is idempotent with respect to ordinary matrix multiplication.

If an overlap matrix $\mathbf{S}(K')$ expressed in terms of the basis functions $\varphi_i(\mathbf{r}, K')$ moved to the new nuclear configuration K', then the inverse Löwdin transform, when applied to $\mathbf{S}(K)^{1/2}\mathbf{P}(K)\mathbf{S}(K)^{1/2}$ gives a new approximation to the density matrix at the displaced nuclear geometry:

$$\mathbf{P}(K', [K]) = \mathbf{S}(K')^{-1/2}\mathbf{S}(K)^{1/2}\mathbf{P}(K)\mathbf{S}(K)^{1/2}\mathbf{S}(K')^{-1/2}. \tag{25}$$

This matrix $\mathbf{P}(K', [K])$ is an idempotent improved approximation of the density matrix $\mathbf{P}(K')$. Simple substitution shows that the matrix $\mathbf{P}(K', [K])$ is idempotent with respect to $*$ multiplication:

$$\mathbf{P}(K', [K])\mathbf{S}(K')\mathbf{P}(K', [K]) = \mathbf{P}(K', [K]). \tag{26}$$

One may regard the transformation of $\mathbf{P}(K)$ into the new density matrix $\mathbf{P}(K', [K])$ as a formal 'orthonormalization–deorthonormalization' carried out with respect to the original and displaced basis sets at the two nuclear geometries K and K' respectively.

The most time-consuming step in the forward-inverse Löwdin transforms of the technique is the calculation of the macromolecular overlap matrices $\mathbf{S}(K)$ and $\mathbf{S}(K')$. However, this is only a minor problem, since these overlap matrices are rather sparse for large molecules, and simple internuclear distance conditions can be used to identify most near-zero elements.

In some instances it is sufficient to estimate a macromolecular density matrix from a matrix already determined for a nuclear geometry that differs only slightly from the conformation of interest. In these cases, the ADMA density matrix $\mathbf{P}(K)$ can be readjusted for small nuclear geometry variations using the method described above.

It is worth emphasizing that the ADMA density matrix $\mathbf{P}(K)$ represents the same accuracy as an infinite-grid-resolution numerical MEDLA electron density; however, the ADMA method is more versatile. The ADMA macromolecular density matrix, with idempotency correction if needed, can also be used for the computation of a variety of molecular properties, including approximate macromolecular forces [61, 62].

For large molecules the quantum-chemical computation of forces acting on individual nuclei is a difficult problem. The ADMA method provides an approximation (ADMA–FORCE) for macromolecular force computation [61, 62]. If macromolecular electron densities are available, it is natural to consider the electrostatic Hellmann–Feynman [80, 81] theorem for force calculations [82]. The sensitivity of calculated Hellmann–Feynman forces to the quality of molecular wavefunction or to the quality of electron density is a limitation that reduces its applicability. However, the natural fragment size limitation of the AFDF approach reduces the effect of locally 'overcomplete' basis sets in ordinary Hartree–Fock–Roothaan–Hall techniques, which appears as one of the sources of errors in the application of the Hellmann–Feynman theorem. Work is in progress for establishing error estimates of the ADMA-FORCE method. Note that, for macromolecules, even a rough estimate of the forces, as provided by the ADMA–FORCE application of the electrostatic Hellmann–Feynman theorem, appears valuable.

Approximate ADMA electron densities of macromolecules can be computed relatively easily from assembled density matrices and basis set information $\varphi(K)$. If an approximate ADMA electron density $\rho(\mathbf{r}, K)$ of a macromolecule M of nuclear configuration K has been determined from the assembled density matrix $\mathbf{P}(K)$, then approximate forces can be computed using the electrostatic Hellmann–Feynman theorem [61, 62]. If in this configuration K the three-dimensional position vector of nucleus a of nuclear charge z_a is denoted by $\mathbf{R}_a = \mathbf{R}_a(K)$, and if $\mathbf{F}_a$ denotes the force operator representing the force acting on nucleus a, then, according to the electrostatic Hellmann–Feynman theorem [80–82], the expectation value of this force operator can be written as

$$\langle \mathbf{F}_a(K) \rangle = -z_a \int \rho(\mathbf{r}, K)\,(\mathbf{R}_a - \mathbf{r})\,|\mathbf{R}_a - \mathbf{r}|^{-3}\,\mathrm{d}\mathbf{r} + z_a \sum_{a \neq b}^{N} z_b (\mathbf{R}_a - \mathbf{R}_b)\,|\mathbf{R}_a - \mathbf{R}_b|^{-3}. \quad (27)$$

This force can be interpreted as a sum of nuclear repulsion and a classical contribution from the electronic charge density $\rho(\mathbf{r}, K)$. The integral in the first term of the expectation value can be computed efficiently if an ADMA density matrix $\mathbf{P}(K)$ and the associated basis set information $\varphi(K)$ are available [61, 62].

The fuzzy 'shares' of electron densities of local moieties and functional groups within macromolecules can be interpreted using the AFDF approach. Consider a nuclear family f_k that contains all the nuclei of the moiety or functional group of

interest and, by applying the AFDF method, a fuzzy electron density fragment of the moiety or functional group is obtained. Note that in this approach there are no sharp boundaries separating local electron density contributions, and the fuzzy electron density fragments have properties analogous to those of complete molecules.

2.4. *Density domains, functional groups and their representation using the additive fuzzy density fragmentation principle*

Following the terminology used in electron density shape analysis [65], a *molecular isodensity contour* (MIDCO) $G(K, a)$ is a collection of all points $\mathbf{r}$ of the three-dimensional space where the electronic density $\rho(K, \mathbf{r})$ is equal to a given threshold value a:

$$G(K, a) = \{\mathbf{r} : \rho(K, \mathbf{r}) = a\}. \tag{28}$$

The point set that includes all points of the MIDCO $G(K, a)$ and all the points within its interior is called a *density domain* and is denoted by $\mathrm{DD}(K, a)$:

$$\mathrm{DD}(K, a) = \{\mathbf{r} : \rho(K, \mathbf{r}) \geqslant a\}. \tag{29}$$

Whereas the density threshold a is a continuous parameter and each molecule has an infinite number of $\mathrm{DD}(K, a)$, nevertheless, for each nuclear configuration K there are only a finite number of topologically different bodies of DDs.

DDs form the basis for a quantum-chemical definition of chemical functional groups [64–66], and they also provide a natural representation of molecular bodies and chemical bonding in molecules [48, 51].

The motivation and justification of the approach followed here for the quantum-chemical representation of functional groups can be illustrated by the following example. Take two molecules, for example two methane molecules in an arrangement where the two carbon atoms are separated by a distance of 20 au. Even at this distance, the electron densities of these two molecules overlap slightly, that is at some low-density threshold a there exists an isodensity contour $G(K, a)$ that encloses the nuclei of both molecules. Nevertheless, the two molecules maintain their separate identities and, at some higher density threshold a', one finds two separate MIDCOs $G_1(K_1, a')$ and $G_2(K_2, a')$, each enclosing the nuclei of only one of these two molecules. Clearly, the separate identities of these two molecules are manifested by the existence of two separate MIDCOs $G_1(K_1, a')$ and $G_2(K_2, a')$.

How does this example relate to functional groups? In many instances, a given chemical functional group can be found in many different molecules; yet the functional group shows only limited variations. Functional groups appear to have some limited identity within molecules. This aspect can be captured by the same tool that indicated the separate identity of two molecules in the example of a pair of methane molecules: the existence of a MIDCO that separates the nuclei of the functional group from the rest of the nuclei of the molecule.

Be definition, a *quantum-chemical functional group* is a fuzzy electron density fragment (AFDF fragment) associated with a family of nuclei f_k, where for this set of nuclei exists some density threshold a such that the corresponding MIDCO $G(K, a)$ separates this family f_k from the rest of the nuclei of the molecule. Of course, the separating MIDCOs of functional groups correspond to somewhat higher-density thresholds a than those of the MIDCOs separating individual molecules; nevertheless, the limited autonomy of functional groups is a valid aspect that is reflected in the definition.

Consider an electron density threshold a. Take the family of functional groups

$$F_1, F_2, \ldots, F_m \tag{30}$$

of a macromolecule M of some nuclear configuration K, where the corresponding density domains

$$\mathrm{DD}_1(a, K), \mathrm{DD}_2(a, K), \ldots, \mathrm{DD}_m(a, K) \tag{31}$$

appear as separate entities at this threshold a.

The electron density contribution $\rho^i(\mathbf{r})$ of each functional group F_i can be calculated using the AFDF approach, based on the AFDF of the macromolecular density $\rho_M(\mathbf{r})$. The nuclear set chosen for each fuzzy fragment density is the nuclear set embedded in the corresponding density domain $\mathrm{DD}_i(a, K)$ representing functional group F_i.

The corresponding fuzzy fragment electron density contributions

$$\rho_{F_1}(\mathbf{r}), \rho_{F_2}(\mathbf{r}), \ldots, \rho_{F_i}(\mathbf{r}), \ldots, \rho_{F_m}(\mathbf{r}), \tag{32}$$

represent the 'share' of each functional group F_i within the total electron density $\rho_M(\mathbf{r})$ of the macromolecule M.

One may reconstruct the electron density $\rho_M(\mathbf{r})$ of macromolecule M by a simple superimposition of the fuzzy fragment densities of the family of $F_1, F_2, \ldots, F_m$ functional groups.

One should note that the selected density threshold value a identifies only some of the possible functional groups $F_1, F_2, \ldots, F_m$ of molecule M. If a different threshold value a' is chosen, then a different assignment of nuclei to individual density domains may result, providing a manifestation of the 'limited autonomy' of a different set of functional groups within the same macromolecule M. Note that the separate density domains for each functional group exists only within a limited range of density thresholds, nevertheless, the functional groups $F_1, F_2, \ldots, F_m$ and the corresponding fragment electron densities $\rho_{F_1}(\mathbf{r}), \rho_{F_2}(\mathbf{r}), \ldots, \rho_{F_m}(\mathbf{r})$ are not restricted to any specific threshold value a of the macromolecular electron density.

3. The spherically weighted affine transformation method for deformations of electron densities

Consider a macromolecular electron density corresponding to a nuclear configuration K, and a small distortion ΔK resulting in a nearly identical nuclear configuration K'. The n nuclei of the macromolecule are denoted by $A_1, A_2, \ldots, A_i, \ldots, A_n$, and the three-dimensional position vectors of nucleus A_i are denoted by $\mathbf{v}^{(i)}$ and by $\mathbf{t}^{(i)}$ in configurations K and K' respectively.

Any four non-coplanar nuclei of some indices p, q, r, s in conformation K define a tetrahedron (p, q, r, s, V), with vertices denoted by

$$\mathbf{v}^{(p)}, \mathbf{v}^{(q)}, \mathbf{v}^{(r)} \text{ and } \mathbf{v}^{(s)}. \tag{33}$$

The vertices of the corresponding tetrahedron (p, q, r, s, T) of the same four nuclei in the distorted 'target' conformation K' are denoted by

$$\mathbf{t}^{(p)}, \mathbf{t}^{(q)}, \mathbf{t}^{(r)} \text{ and } \mathbf{t}^{(s)}. \tag{34}$$

With reference to tetrahedron (p, q, r, s, V), any vector $\mathbf{v}$ of the three-dimensional space can be written as an affine combination

$$\mathbf{v} = c^{(p)}\mathbf{v}^{(p)} + c^{(q)}\mathbf{v}^{(q)} + c^{(r)}\mathbf{v}^{(r)} + c^{(s)}\mathbf{v}^{(s)}. \tag{35}$$

For the affine coordinates

$$c^{(p)}, c^{(q)}, c^{(r)} \text{ and } c^{(s)} \tag{36}$$

of vector **v** with respect to tetrahedron (p, q, r, s, V), the condition

$$c^{(p)} + c^{(q)} + c^{(r)} + c^{(s)} = 1 \tag{37}$$

holds.

A linear transformation that distorts the entire space so that the tetrahedron (p, q, r, s, V) becomes the tetrahedron (p, q, r, s, T) is defined by replacing each vector **v** with a vector **t** where the affine coordinates of **t** with respect to tetrahedron (p, q, r, s, T) are the same as the affine coordinates of vector **v** with respect to tetrahedron (p, q, r, s, V); that is, vector **v** of equation (35) is replaced by

$$\mathbf{t} = c^{(p)}\mathbf{t}^{(p)} + c^{(q)}\mathbf{t}^{(q)} + c^{(r)}\mathbf{t}^{(r)} + c^{(s)}\mathbf{t}^{(s)}. \tag{38}$$

For each pair (p, q, r, s, V) and (p, q, r, s, T) of (non-degenerate) tetrahedra, this transformation is a linear homotopy.

In [51] a simple extension of this transformation, the weighted affine transformation (WAT) method, was suggested where for a molecule of more than four nuclei all possible tetrahedra were considered. By combining all the corresponding linear transformations using nonlinear weight functions, a global transformation was proposed. This WAT method distorts the entire three-dimensional space so that each nucleus is moved precisely to its assigned new location and the surrounding electron density is distorted accordingly. Whereas for large deformations the transformed electron density may show excessive deviations from the actual deformation that occurs in an actual conformational change, if the nuclear displacements are small, then the WAT technique provides a useful approximation to the actual electron density change.

The number of individual transformations combined within the WAT method is $n_s = n(n-1)(n-2)(n-3)/4!$, where n is the number of nuclei in the molecule. The value of n_s grows with the fourth power of n, which renders the method practical only for small molecules or small molecular fragments. Clearly, for a macromolecule the number n_s becomes much too large and the WAT technique impractical. However, in local ranges of large molecules the effects of a displacement of some nuclei far away from the local range and the electron density surrounding such distant nuclei are of minor importance. This suggests a modification of the WAT method to involve only those local tetrahedral transformations which are relevant to each local range of the macromolecule. This reduces the number of tetrahedra involved in the overall transformation. By an additional nonlinear spherical weighting, which preserves the exact transformation for the nuclear locations, the modified technique, referred to as the spherically weighted affine transformation (SWAT) technique, becomes applicable for the generation of approximate electron density deformations associated with small conformational changes of macromolecules [68].

Equation (35) describing the affine representation of point **v** can be rearranged:

$$\mathbf{v} - \mathbf{v}^{(s)} = c^{(p)}(\mathbf{v}^{(p)} - \mathbf{v}^{(s)}) + c^{(q)}(\mathbf{v}^{(q)} - \mathbf{v}^{(s)}) + c^{(r)}(\mathbf{v}^{(r)} - \mathbf{v}^{(s)}). \tag{39}$$

The column vectors $\mathbf{v}^{(p)} - \mathbf{v}^{(s)}$, $\mathbf{v}^{(q)} - \mathbf{v}^{(s)}$ and $\mathbf{v}^{(r)} - \mathbf{v}^{(s)}$ define a matrix $\mathbf{S}^{(p, q, r, s, V)}$,

$$\mathbf{S}^{(p, q, r, s, V)} = \mathrm{mat}\left|(\mathbf{v}^{(p)} - \mathbf{v}^{(s)})\ (\mathbf{v}^{(q)} - \mathbf{v}^{(s)})\ (\mathbf{v}^{(r)} - \mathbf{v}^{(s)})\right|, \tag{40}$$

and the first three affine coordinates $c^{(p)}, c^{(q)}, c^{(r)}$ define a column vector $\mathbf{c}^{(p, q, r, s)}$,

$$\mathbf{c}^{(p, q, r, s)} = (c^{(p)}, c^{(q)}, c^{(r)})'. \tag{41}$$

Equation (39) can be rearranged to give

$$\mathbf{c}^{(p, q, r, s)} = (\mathbf{S}^{(p, q, r, s, V)})^{-1}(\mathbf{v} - \mathbf{v}^{(s)}). \tag{42}$$

For every non-degenerate tetrahedron (p, q, r, s, V), the inverse matrix $(\mathbf{S}^{(p,q,r,s,V)})^{-1}$ of $\mathbf{S}^{(p,q,r,s,V)}$ exists.

By analogous treatment of the vectors describing the displaced nuclear arrangement, one obtains

$$\mathbf{S}^{(p,q,r,s,T)} = \mathrm{mat}\left|(\mathbf{t}^{(p)} - \mathbf{t}^{(s)})\ (\mathbf{t}^{(q)} - \mathbf{t}^{(s)})\ (\mathbf{t}^{(r)} - \mathbf{t}^{(s)})\right| \tag{43}$$

and

$$\mathbf{t} - \mathbf{t}^{(s)} = \mathbf{S}^{(p,q,r,s,T)}\mathbf{c}^{(p,q,r,s)}. \tag{44}$$

Equations (42) and (44) give

$$\mathbf{t} - \mathbf{t}^{(s)} = \mathbf{S}^{(p,q,r,s,T)}(\mathbf{S}^{(p,q,r,s,V)})^{-1}(\mathbf{v} - \mathbf{v}^{(s)}). \tag{45}$$

Using the notation

$$\mathbf{D}^{(p,q,r,s)} = \mathbf{S}^{(p,q,r,s,T)}(\mathbf{S}^{(p,q,r,s,V)})^{-1} \tag{46}$$

and

$$\mathbf{u}^{(p,q,r,s)} = \mathbf{t}^{(s)} - \mathbf{D}^{(p,q,r,s)}\mathbf{v}^{(s)}, \tag{47}$$

the vector $\mathbf{t}$ can be written as

$$\mathbf{t} = \mathbf{D}^{(p,q,r,s)}\mathbf{v} + \mathbf{u}^{(p,q,r,s)}. \tag{48}$$

Within the SWAT technique, if one considers the overall transformation of a point $\mathbf{v}$, only those tetrahedra (p, q, r, s, V) and the associated linear transformations will be used from the complete family of n_s tetrahedra (p, q, r, s, V) and the corresponding n_s transformations for which all four vertices p, q, r and s of the tetrahedron fall within a suitably chosen radius R of point $\mathbf{v}$. This reduces the number of n_s of tetrahedra to be considered to a number $n_s(\mathbf{v}, R)$ where $n_s(\mathbf{v}, R) < n_s$. A suitable nonlinear weighting is determined only for the transformations associated with these tetrahedra.

For each point $\mathbf{v}$, one non-zero weight function is associated with each of the tetrahedra fulfilling the distance criterion. These $\mathbf{v}$-dependent and R-dependent weight functions $w^{(p,q,r,s)}(\mathbf{v}, R)$ are required to fulfil several conditions.

(i) If $\mathbf{v} = \mathbf{v}^{(i)}$, then the weighted average of the $n_s(\mathbf{v}, R)$ transformations assigns the point $\mathbf{v}^{(i)}$ of the ith nuclear position exactly to the point $\mathbf{t}^{(i)}$ of the new ith nuclear position.

(ii) As a function of $\mathbf{v}$, the weighted average of the $n_s(\mathbf{v}, R)$ selected transformations deforms the electronic density continuously.

(iii) The summation of the weight functions for all tetrahedra (p, q, r, s), which is equivalent to a summation of the weight functions for the selected $n_s(\mathbf{v}, R)$ tetrahedra fulfilling the distance criterion, must result in unity:

$$\sum_{(p,q,r,s)} w^{(p,q,r,s)}(\mathbf{v}, R) = 1. \tag{49}$$

In order to construct a suitable weight function $w^{(p,q,r,s)}(\mathbf{v}, R)$ for each tetrahedron, first the distance condition is tested, and one sets

$$h^{(p,q,r,s)}(\mathbf{v}, R) = 0 \tag{50}$$

if any of the vertices of a tetrahedron (p, q, r, s, V) falls on or outside the sphere of radius R and centre $\mathbf{v}$, that is if any of the following conditions holds:

$$\left|\mathbf{v}^{(p)} - \mathbf{v}\right| \geqslant R, \tag{51}$$

$$\left|\mathbf{v}^{(q)} - \mathbf{v}\right| \geqslant R, \tag{52}$$

$$\left|\mathbf{v}^{(r)} - \mathbf{v}\right| \geqslant R, \tag{53}$$

$$\left|\mathbf{v}^{(s)} - \mathbf{v}\right| \geqslant R. \tag{54}$$

For the actual point **v**, only those $\mathbf{v}^{(j)}$ vertices are used which fulfil the distance condition

$$\left|\mathbf{v}^{(j)} - \mathbf{v}\right| < R. \tag{55}$$

For each vertex $\mathbf{v}^{(i)}$ in the family of vertices falling within the sphere, a **v**-dependent function $f^{(i)}(\mathbf{v}, R)$ is defined as

$$f^{(i)}(\mathbf{v}, R) = \prod_{j(j \neq i)} d(\mathbf{v}, \mathbf{v}^{(j)}), \tag{56}$$

where $d(\mathbf{v}, \mathbf{v}^{(j)})$ is the distance between points **v** and $\mathbf{v}^{(j)}$, and where index j runs over all $\mathbf{v}^{(j)}$ vertices falling within the sphere. If point **v** coincides with any of these nuclear positions $\mathbf{v}^{(j)}$, then $f^{(i)}(\mathbf{v})$ becomes zero, except if $j = i$, that is if **v** coincides with the nuclear position $\mathbf{v}^{(i)}$.

In order to introduce an additional smooth, infinitely differentiable spherical weighting that is equal to unity at point **v** and is identically zero outside the sphere of radius R and centre **v**, the following functions are used:

$$f(y) = \begin{cases} \exp\left(\dfrac{-1}{y^2}\right), & \text{if } y > 0, \\ 0 & \text{if } y \leqslant 0, \end{cases} \tag{57}$$

and

$$F(r, R) = \frac{f(R+r)f(R-r)}{f^2(R)}. \tag{58}$$

The combined weighting function $g^{(i)}(\mathbf{v}, R)$ for each vertex $\mathbf{v}^{(i)}$ in the family of vertices falling within the sphere is defined as

$$g^{(i)}(\mathbf{v}, R) = F(d(\mathbf{v}, \mathbf{v}^{(i)}), R) f^{(i)}(\mathbf{v}, R), \tag{59}$$

that is as

$$g^{(i)}(\mathbf{v}, R) = F(d(\mathbf{v}, \mathbf{v}^{(i)}), R) \prod_{j(j \neq i)} d(\mathbf{v}, \mathbf{v}^{(j)}). \tag{60}$$

As a function of point **v**, the sphere defining the local surroundings of point **v** changes continuously, and the weights assigned to various vertices $\mathbf{v}^{(i)}$ as they are reached by the surface of the sphere also change continuously and smoothly from zero to a maximum of unity at the centre **v** of the sphere.

For each tetrahedron (p, q, r, s) that falls within the sphere about the given point **v**, a continuous function $h^{(p, q, r, s)}(\mathbf{v}, R)$ is defined as

$$h^{(p, q, r, s)}(\mathbf{v}, R) = g^{(p)}(\mathbf{v}, R) + g^{(q)}(\mathbf{v}, R) + g^{(r)}(\mathbf{v}, R) + g^{(s)}(\mathbf{v}, R). \tag{61}$$

If the point **v** coincides with any of the nuclear positions $\mathbf{v}^{(j)}$, then $g^{(p, q, r, s)}(\mathbf{v}, R)$ becomes zero, except if j is one of the indices p, q, r or s, that is if point **v** coincides with one of the nuclear positions $\mathbf{v}^{(p)}$, $\mathbf{v}^{(q)}$, $\mathbf{v}^{(r)}$ or $\mathbf{v}^{(s)}$.

The function $h_{\text{sum}}(\mathbf{v}, R)$ is the sum of all these $h^{(p, q, r, s)}(\mathbf{v}, R)$ functions:

$$h_{\text{sum}}(\mathbf{v}, R) = \sum_{(p, q, r, s)} h^{(p, q, r, s)}(\mathbf{v}, R). \tag{62}$$

The **v**-dependent and R-dependent weight functions $w^{(p, q, r, s)}(\mathbf{v}, R)$ are defined as

$$w^{(p, q, r, s)}(\mathbf{v}, R) = \frac{h^{(p, q, r, s)}(\mathbf{v}, R)}{h_{\text{sum}}(\mathbf{v}, R)}. \tag{63}$$

The overall transformation is defined as

$$\mathbf{t} = \sum_{(p,q,r,s)} w^{(p,q,r,s)}(\mathbf{v}, R)\,(\mathbf{D}^{(p,q,r,s)}\mathbf{v} + \mathbf{u}^{(p,q,r,s)}), \tag{64}$$

where for each point $\mathbf{v}$ the summation can be restricted to those tetrahedra which fall within the sphere about $\mathbf{v}$.

This weighting scheme ensures that each nuclear position $\mathbf{v}^{(j)}$ is transformed exactly to its counterpart nuclear position $\mathbf{t}^{(j)}$, while the entire electron density is deformed continuously. The SWAT method uses only those reference positions (nuclear positions) as vertices for affine transformations which are within the vicinity of each point $\mathbf{v}$ being transformed. This latter feature is advantageous for macromolecular electron densities, since only a subset of nuclear locations is involved in the transformation of any given point. The method has no origin or coordinate dependence. This SWAT algorithm has been implemented as a computer program [68].

If the SWAT-generated approximate macromolecular electron density at the new nuclear geometry K' is compared with the electron density calculated directly by the ADMA method for the same nuclear configuration K', then an interesting analogy with the le Chatelier principle applies. The change in nuclear geometry from K to K', and the associated change in the electron density generated by the SWAT method, relying on the density at the original configuration K, may be regarded analogous to a formal 'stress' applied to a thermodynamic system in equilibrium. The replacement of the SWAT electron density with an actual ADMA electron density at the new configuration K' may be regarded as a formal 'relaxation' of this 'stress'. This situation is analogous to the electron density shape changes due to electronic excitations where an apparent trend, called the quantum-chemical le Chatelier principle for molecular shapes (QCLCP-MS) seems to apply [65]; the shape change due to relaxation from the shape of the electronic density obtained in a vertical electronic excitation (before nuclear rearrangement) tends to reduce the initial shape change of the vertical excitation. The analogy takes a different form in the case of the SWAT transformation followed by a replacement of the density with a direct ADMA density at the new nuclear configuration. According to the QCLCP-MS, the replacement of the SWAT density with the ADMA density is expected to reduce the initial shape change generated by the SWAT transformation.

4. The extension of the shape group methods to the global and local analysis of macromolecular electron densities

The topological shape analysis techniques provide a numerical representation of shape information, and the numbers so determined form a shape code. Numerical shape codes can be compared with a computer, and numerical measures of molecular shape similarity can be computed. One apparent advantage is the elimination of the subjective element of visual shape comparisons, a potentially important aspect if large sequences of molecules are compared. Local and global similarity analyses based on molecular shape codes are of some theoretical interest in the study of the role of electron density in static molecular properties and in reactivity and may also have some practical interest in drug design [83–90] and toxicological risk assessment [58].

4.1. *The elements of the shape group methods*

The entire electron density of a molecule M can be described by an infinite family of MIDCOs $G(K, a)$ nested within one another. As follows directly from their

definition, equation (28), for two MIDCOs $G(K, a)$ and $G(K, a')$ of the same nuclear configuration K, the relation

$$G(K, a) \text{ encloses } G(K, a') \tag{65}$$

holds if $a \leqslant a'$.

The three-dimensional shape groups, describing directly the pattern of interrelations between the various three-dimensional curvature domains of the electronic density function, where the density value is regarded as the coordinate along the fourth dimension, have been discussed earlier [83, 84] and have been found to possess some advantages [65, 73]. Nevertheless, the interpretation of results and practical applications of shape analysis are somewhat simpler if the shape group methodology is applied in its two-dimensional realization of the family of MIDCOs. Note that the MIDCOs are regarded as two-dimensional surfaces embedded in the ordinary three-dimensional space.

The shape groups are algebraic-topological structures, defined as the homology groups of truncated objects, where the truncation is determined by local shape properties, for example by local curvature properties [69–72]. Although the shape groups are algebraic groups describing shape properties, they are not related to point symmetry groups. The presence of symmetry, however, may influence the shape groups. In the usual applications of shape groups, the local shape properties are specified in terms of shape domains, for example in terms of the local convex, concave or saddle-type regions of MIDCOs [65], where the local curvatures of individual MIDCOs are compared with tangent planes.

A more detailed shape description of MIDCOs is obtained if the tangent plane is replaced by some other, possibly curved objects, for example, if the MIDCO is compared with a series of tangent spheres of various radii r or with a series of oriented tangent ellipsoids T. The latter choice is advantageous if a shape characterization involving some reference directions is needed; these ellipsoids can be translated but not rotated as they are brought into tangential contact with various points of the MIDCO surface $G(K, a)$.

A general tangent object T may fall locally on the outside or on the inside, or it may cut into the given MIDCO surface $G(K, a)$ within any small neighbourhood of the surface point $\mathbf{r}$ of the tangential contact. Accordingly, with reference to T, all points of the MIDCO are classified into one of three classes. By carrying out this characterization for all points $\mathbf{r}$, the MIDCO $G(K, a)$ is formally decomposed into several local shape domains of types D_2, D_0 and D_1, that is into locally convex, locally concave and locally saddle-type shape domains respectively, where the terms concave, saddle type and convex are interpreted relative to the tangent object T. For example, if a tangent sphere of some curvature b is used for reference object T, then a 'locally convex D_2 domain of $G(K, a)$ relative to T' corresponds to a surface domain of $G(K, a)$ where at each point within the domain both the minimum local curvature and the maximum local curvature of the surface are less than the value b. Note that a typical MIDCO is an orientable surface, and a tangent sphere T may osculate to $G(K, a)$ either from the inside or from the outside of $G(K, a)$. These two cases correspond to a negative or a positive reference curvature value b respectively. More details of the actual determination of these D_2, D_0 and D_1 curvature domains of MIDCOs can be found in [65].

If T is chosen as a sphere, then orientation cannot play any role, and for a sphere of radius r the curvature is $b = 1/r$. Note that $b = 0$ corresponds to the case of the infinitely large sphere, the tangent plane.

The characterization depends on the curvature parameter b, and in a detailed shape analysis a continuum of b values is considered. For each specific reference curvature b, the local shape domains D_2, D_0 and D_1 generate a complete partitioning of the MIDCO surface $G(K, a)$. Select all D_μ domains of a specified type μ; for example take all the locally convex domains D_2 relative to reference curvature b. If these domains are excised from the MIDCO surface $G(K, a)$, then a truncated contour surface $G(K, a, \mu)$ is obtained that inherits some essential shape information from the original MIDCO surface $G(K, a)$. This shape information can now be detected and identified by simple topological means. This procedure is repeated for a whole range of reference curvature values b, that is for a whole series of truncated surfaces, which gives a detailed shape analysis of the original non-truncated MIDCO surface $G(K, a)$. Important simplification is possible as a consequence of the simple fact that within the entire range of possible reference curvature values b there are only a finite number of topologically different truncated MIDCOs $G(K, a, \mu)$.

The truncated surfaces are characterized by their topological invariants, and these invariants provide a numerical shape characterization.

The homology groups of the truncated surfaces are groups of algebraic topology which are themselves topological invariants. The ranks of the homology groups are the Betti numbers, which are perhaps the most important topological invariants.

The notations for the shape groups of the original MIDCO surface $G(K, a)$ follow from their definition as the homology groups $H^p_\mu(a, b)$ of the truncated surfaces $G(K, a, \mu)$, where the formal dimensions p of these three shape groups are zero, one and two. Numerical shape codes of the molecular electron density distributions are generated by the lists of the $b^p_\mu(a, b)$ Betti numbers of their $H^p_\mu(a, b)$ shape groups. For each shape domain and truncation pattern μ and for each reference curvature b of a given MIDCO $G(K, a)$ of density threshold a, three shape groups $H^0_\mu(a, b)$, $H^1_\mu(a, b)$, and $H^2_\mu(a, b)$ are determined, collectively expressing the essential shape information of the MIDCO $G(K, a)$. For each (a, b) pair of density threshold a and curvature parameter b, and for each shape domain truncation type μ, there are three Betti numbers, denoted $b^0_\mu(a, b)$, $b^1_\mu(a, b)$, and $b^2_\mu(a, b)$.

The essential steps in the application of the SGM can be summarized as follows.

Step 1. Choose a range of electron density thresholds a and a range of reference curvatures b. For each pair of values a and b within these ranges, each MIDCOs $G(K, a)$ is partitioned into local curvature domains relative to each value b. Note that in practice only a finite number of (a, b) pairs need to be considered in order to identify all the topologically different patterns of curvature domains. The local curvature of a MIDCO surface ($G(K, a)$ at some point $\mathbf{r}$ is characterized by a local curvature matrix called the local Hessian matrix. The points of $G(K, a)$ are classified into curvature domains of types $D_0(b)$, $D_1(b)$ or $D_2(b)$, by comparing the local canonical curvatures (the eigenvalues of the local Hessian matrices) at each surface point to the reference curvature b. A point $\mathbf{r}$ of $G(K, a)$ is assigned to a $D_0(b)$, $D_1(b)$ or $D_2(b)$ curvature domain, if none, one or two (respectively) of the eigenvalues of the local Hessian matrix of the surface at point $\mathbf{r}$ are smaller than b.

Step 2. For each (a, b) pair of values, all curvature domains $D_\mu(b)$ of a specified type μ are formally removed from the corresponding MIDCO $G(K, a)$, resulting in a truncated surface $G(K, a, \mu)$. For each molecule and for the

whole range of parameter values a and b, there are only a finite number of classes of topologically different truncated surfaces.

Step 3. The shape groups of the entire molecular electron density distribution are determined by calculating the algebraic homology groups for each topological equivalence class of the truncated surfaces. The Betti numbers are the ranks of these homology groups. The Betti numbers serve as numerical shape descriptors, describing the mutual relations of the various local shape domains for the entire range of MIDCOs $G(K, a)$.

4.2. *Numerical shape codes using* (a, b) *parameter maps*

For any fixed nuclear configuration K, the shape groups of a molecule depend on two parameters: the electronic density threshold a and the reference curvature b. By definition, a positive b value indicates a tangent sphere placed on the exterior side of the MIDCO surface, and a negative b value indicates a sphere placed on the interior side of the MIDCO. The ranges of a and b define an (a, b) map, a formal two-dimensional map, where the distribution of the Betti numbers of various shape groups along this map gives a detailed numerical shape characterization of the electronic density of the molecule M.

In practical computations, separate (a, b) maps are generated for each of the three types of Betti numbers $b^0_\mu(a, b)$, $b^1_\mu(a, b)$, and $b^2_\mu(a, b)$. The Betti numbers obtained for a given pair of values of parameters a and b are assigned to the given location of the (a, b) parameter map. The chemically most relevant shape information is described by the Betti numbers of type $b^1_\mu(a, b)$, that is by the ranks of the shape groups of dimension 1.

Usually, a grid of a and b values is considered within some interval $[a_{\min}, a_{\max}]$ of density thresholds a and some interval $[b_{\min}, b_{\max}]$ of reference curvature values b. The range of these parameters often covers several orders of magnitude, and it is advantageous to use logarithmic scales. The $\log |b|$ values are taken for negative values of the curvature parameter b.

In some applications, the range [0·001, 0·1 au.] is taken for the density threshold values a, the range of $[-1{\cdot}0, 1{\cdot}0]$ is taken for the curvature b of the test spheres, and a 41×21 grid is used on a logarithmic scale. The values of the Betti numbers at the grid points (a, b) form a numerical shape code matrix $\mathbb{M}^{(a, b)}$ representing the shape of the fuzzy electron density of the molecule.

4.3. *Numerical shape similarity measures from shape codes*

The numerical shape codes, in the form of matrices $\mathbb{M}^{(a, b)}$ of the (a, b) maps of Betti numbers $b^p_\mu(a, b)$ are suitable for the evaluation of numerical shape similarity measures between molecules. If n_a and n_b are the number of grid divisions for parameters a and b respectively, then the total number of elements in the shape code matrix is

$$t = n_a n_b. \tag{66}$$

The shape codes $\mathbb{M}^{(a, b), \mathrm{A}}$ and $\mathbb{M}^{(a, b), \mathrm{B}}$ of two molecules A and B respectively can be compared and a numerical shape similarity measure can be defined as

$$s(\mathrm{A}, \mathrm{B}) = \frac{m[\mathbb{M}^{(a, b), \mathrm{A}}, \mathbb{M}^{(a, b), \mathrm{B}}]}{t}, \tag{67}$$

where $m[\mathbb{M}^{(a, b), \mathrm{A}}, \mathbb{M}^{(a, b), \mathrm{B}}]$ is the number of matches between corresponding elements in the two matrices.

If the 41×21 grid is used, then the elements of matrix $\mathbb{M}^{(a, b)}$ can be stored in an

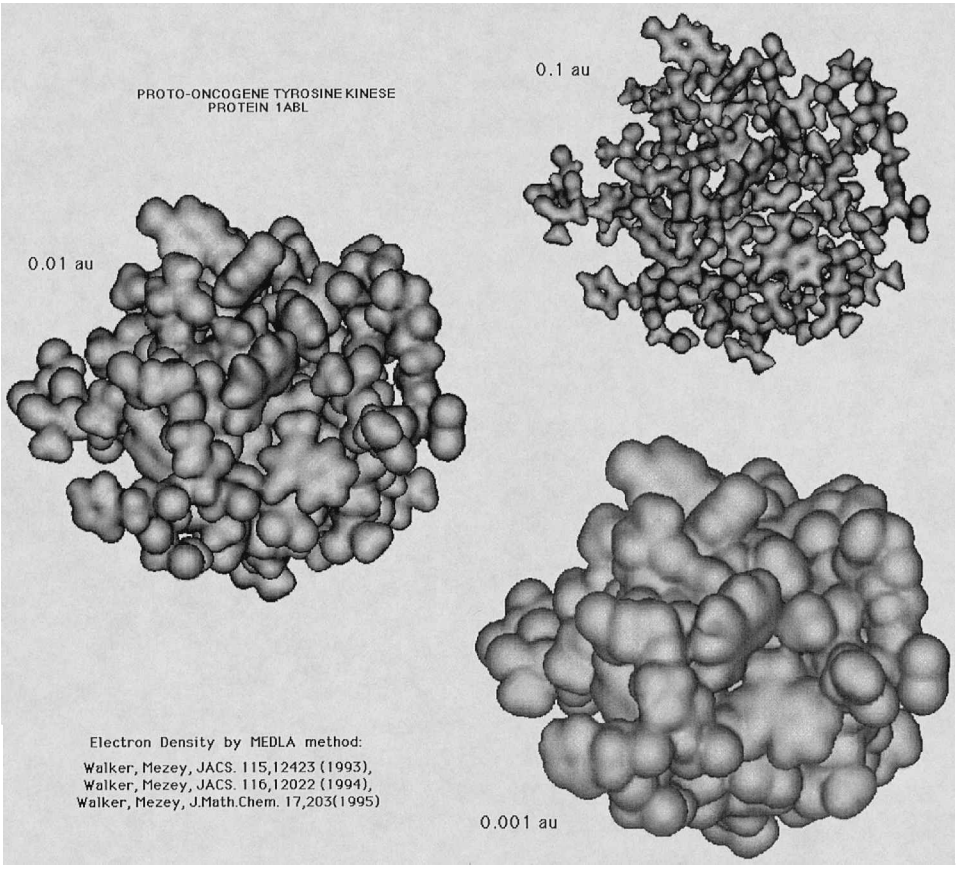

Figure 3. Three MIDCOs $G(0{\cdot}1)$, $G(0{\cdot}01)$ and $G(0{\cdot}001)$ of the calculated *ab initio* quality AFDF electron density of a protein, the proto-oncogene tyrosine kinese protein 1ABL containing 873 atoms. Electron density thresholds are given in atomic units. Macromolecular electron density images computed by the MEDLA or ADMA methods may provide interesting insight. For a reproducible local or global shape analysis, avoiding the subjective elements of visual inspection, non-visual algorithmic methods, such as the SGMs are suggested.

integer vector **C** of 861 components, and the shape similarity measure $s(\mathrm{A},\mathrm{B})$ can be written as

$$s(\mathrm{A},\mathrm{B}) = \sum_{i=1}^{861} \frac{\delta_{j(i),\,k(i)}}{861}, \tag{68}$$

where $\delta_{j,\,k}$ is the Kroenecker delta, and the integers $j(i)$ and $k(i)$ are defined as

$$j(i) = C_i(\mathrm{A}) \tag{69}$$

and

$$k(i) = C_i(B) \tag{70}$$

respectively.

Numerical evaluation and comparison of molecular shape features have been applied to several molecular families and useful shape–property correlations have been established using this methodology [85–89].

4.4. *Application of the shape group methods to global and local shape problems of macromolecules*

Visual inspection of computed electron density contours of macromolecules is a rather subjective tool for the evaluation of local and global shape features. In figure 3, the calculated *ab initio* quality AFDF electron density of a protein, the proto-oncogene tyrosine kinese protein 1ABL containing 873 atoms [90] is displayed at three density thresholds as the MIDCOs $G(0{\cdot}1)$, $G(0{\cdot}01)$ and $G(0{\cdot}001)$. The threshold values are expressed in atomic units. Whereas such images of reasonably accurate electron densities may provide interesting insight, nevertheless, for detailed and reproducible shape evaluations, non-visual algorithmic methods are preferred.

One pragmatic feature of the SGM of topological shape description is the combination of the advantages of geometry and topology. This approach follows the spirit of the principle of geometrical similarity as topological equivalence (GSTE) [65]. On the one hand, the local geometrical curvature properties and the associated geometrical classification of points of the MIDCO surfaces lead to the local shape domains and to the truncated MIDCOs $G(K, a, \mu)$ are defined in terms of geometry. On the other hand, the truncated surfaces $G(K, a, \mu)$ are characterized topologically using the shape groups and their Betti numbers, which are topological invariants within ranges of electron density threshold a and reference curvature parameter b.

One may focus on the global shape features of macromolecular electron densities obtained using the MEDLA or ADMA methods by restricting the range of electron density thresholds a to low densities. This also allows one to reduce the range of curvature parameter b to the intermediate values, since extreme negative or positive local curvatures are not likely to occur at all for MIDCOs of low-density thresholds. Such limited shape analysis ignores many of the fine details of macromolecular shapes evident only at high densities and focuses on the most prominent global features. This approach does not require any specific modification of the shape group methodology beyond the special choice of the range within the (a, b) parameter maps.

A more challenging task is the local shape analysis of molecular moieties within macromolecules [48–52, 65].

The SGMs, implemented as fuzzy electron density shape analysis methods are applicable to molecular fragments, as has been first pointed out in [71]. The fuzzy aspect of this methodology implies that the molecular fragments so analysed are chosen as fuzzy entities, where the electron density of the fragment has no definite boundary, a fuzzy property analogous to that of complete molecules.

This fuzzy aspect of molecular fragments is well reflected in the definition of DDs [64–66], which were the first practical implementation of the fuzzy molecular fragment concept within this framework. Note, however, that two DDs of separate identities at some high-density range may merge into a single DD within a lower-density range, and the individual identities of the fuzzy objects of high density are no longer preserved at a low density. Nevertheless, the overall density can be partitioned into fuzzy fragments even within the density range where merger occurs. This partitioning can be obtained in an additive manner. One such AFDF can be obtained as a generalization of the density domains using the Mulliken–Mezey AFDF method [48–52].

An early adaptation of the SGM for the local shape analysis of fuzzy molecular fragments, such as the local cavity region of an enzyme or the reactive region of another large molecule, has been suggested in [72], using a series of nested contour surfaces representing the local shape of fuzzy electron density fragments. Whereas for

an important range of density thresholds the relevant pieces of the actual contour surfaces of the complete macromolecule do not form closed surfaces, nevertheless these local contours can be completed by suitably extending them to form closed surfaces [72]. These closed surfaces become the actual representations of local fuzzy molecular fragments [72].

The AFDF approach provides a natural and practical tool for the construction of such local fuzzy fragments of the macromolecular electron density. Useful practical implementations of the AFDF technique are the MEDLA method of Walker and Mezey [53–57] and the more recent ADMA macromolecular density matrix technique [50, 52, 60–63].

For a given electron density threshold a, the set of functional groups $F_1, F_2, \ldots, F_m$ of a molecule M is determined by those density domains which appear as separate entities $DD_1(a, K), DD_2(a, K), \ldots, DD_m(a, K)$ at this threshold. It is important to keep in mind that, for a different threshold value a', a different set of DDs and a different set of functional groups might be identified within the same molecule M; however, there are only finite number of topologically different DDs, and a complete accounting of all possible functional groups is possible using only a finite number of density thresholds. In the following discussion we shall consider just one density threshold a; however, the treatment can be extended easily for multiple thresholds. Note that the functional groups $F_1, F_2, \ldots, F_m$ are identified at a single threshold, but the corresponding fragment electron densities $\rho_{F_1}(\mathbf{r}), \rho_{F_2}(\mathbf{r}), \ldots, \rho_{F_m}(\mathbf{r})$ are not restricted to any specific threshold value a.

By taking the nuclear set of serial index k for each fuzzy fragment density as the nuclear set embedded in the corresponding $DD_k(a, K)$ representing the functional group F_k, and by carrying out the fuzzy density fragmentation procedure, the electron density contribution $\rho^k(\mathbf{r})$ of each functional group F_k can be determined using the AFDF scheme. The fuzzy fragment electron density contributions $\rho_{F_1}(\mathbf{r}), \rho_{F_2}(\mathbf{r}), \ldots, \rho_{F_k}(\mathbf{r}), \ldots, \rho_{F_m}(\mathbf{r})$ represent the formal 'share' of each functional group F_k within the total macromolecular electron density $\rho_M(\mathbf{r})$. As a consequence of the exact additivity of the Mulliken–Mezey fragmentation scheme, the total electron density $\rho_M(\mathbf{r})$ of molecule M is the sum of functional group electron densities at each point $\mathbf{r}$:

$$\rho_M(\mathbf{r}) = \sum_k \rho_{F_k}(\mathbf{r}). \tag{71}$$

One may consider the fragment electron density $\rho_{F_i}(\mathbf{r})$ of each individual functional group F_k as a separate individual fuzzy object within the fuzzy body $\rho_M(\mathbf{r})$ of the macromolecule M.

4.5. *Local shape analysis of isolated functional groups*

Since molecular fragments are described by fuzzy electron densities analogous to densities of complete molecules, the local shape analysis of functional groups follows the same principles as the shape analysis of complete molecules. However, the terminology 'fragment isodensity contour' (FIDCO) surface is used instead of MIDCO surface.

The notation F is used for the actual fragment or functional group selected for study and M' denotes the rest of the macromolecule M. This fragment M' is possibly composed from several fragments $F_1, F_2, \ldots, F_{m-1}$, and the fragment F is assumed to correspond to the last fragment in the series: $F = F_m$.

If the molecular density fragment F is regarded as an entity on which the influence of the rest of the molecule is unimportant, then it is meaningful to generate contours for F where the density threshold a is compared only with the actual fragment density $\rho_F(\mathbf{r})$, and the FIDCOs themselves are not influenced by the additional density contributions from the rest of the molecule.

In this case, a FIDCO for a fragment F in a molecule FM' is defined as follows:

$$G_{F\backslash M'}(a) = \{\mathbf{r}: \rho_F(\mathbf{r}) = a,\ \rho_F(\mathbf{r}) \geqslant \rho_{F_k}(\mathbf{r}),\ k = 1, \ldots, m-1\}. \tag{72}$$

The following two alternative definitions are equivalent to that given by equation (72):

$$G_{F\backslash M'}(a) = G_F(a) \cap \{\mathbf{r}: \rho_F(\mathbf{r}) \geqslant \rho_{F_k}(\mathbf{r}),\ k = 1, \ldots, m-1\}, \tag{73}$$

and

$$G_{F\backslash M'}(a) = G_F(a) \backslash \{\mathbf{r}: \exists\, k \in \{1, \ldots, m-1\}: \rho_F(\mathbf{r}) < \rho_{F_k}(\mathbf{r})\}. \tag{74}$$

The FIDCO $G_{F\backslash M'}(a)$ of fragment F in macromolecule $M = FM'$ is the set of all those points $\mathbf{r}$ where the electron density contribution $\rho_F(\mathbf{r})$ of fragment F is dominant within the macromolecule FM'.

The series of such FIDCOs for a whole range of density thresholds a can be analysed using the standard SGM, with one modification: an additional domain type D_{-1} is introduced, representing the connection of fragment F to the rest of the molecule within the actual FM' system:

$$D_{-1}(G_{F\backslash M'}(a)) = \{\mathbf{r}: \mathbf{r} \in G_F(a),\ \exists\, k \in \{1, \ldots, m-1\}: \rho_F(\mathbf{r}) < \rho_{F_k}(\mathbf{r})\}. \tag{75}$$

It is only the boundary $\Delta D_{-1}(G_{F\backslash M'}(a))$, defined as

$$\Delta D_{-1}(G_{F\backslash M'}(a)) = \{\mathbf{r}: \mathbf{r} \in G_{F\backslash M'}(a),\ \exists\, k' \in \{1, \ldots, m-1\}: \rho_F(\mathbf{r}) = \rho_{F_{k'}}(\mathbf{r}),\ \rho_{F_k}(\mathbf{r}) \geqslant \rho_{F_k}(\mathbf{r}),\ k = 1, \ldots, m-1\}, \tag{76}$$

that can be found on the FIDCO $G_{F\backslash M'}(a)$, since the domain $D_{-1}(G_{F\backslash M'}(a))$ itself exists only on the intact $G_F(a)$ contour surface. The actual formal domain $D_{-1}(G_{F\backslash M'}(a))$ appears only as 'cover(s)' over the hole(s) of the FIDCO $G_{F\backslash M'}(a)$ in the macromolecule FM'.

A simpler representation of fragment F in molecule FM' is obtained if the composite M' of all the remaining fragments $F_1, F_2, \ldots, F_{m-1}$ is used. This definition of contours is given as

$$G_{F\backslash \Sigma M'}(a) = \{\mathbf{r}: \rho_F(\mathbf{r}) = a,\ \rho_F(\mathbf{r}) \geqslant \rho_{M'}(\mathbf{r})\}, \tag{77}$$

where $\rho_{M'}(\mathbf{r})$ denotes the composite density given by

$$\rho_{M'}(\mathbf{r}) = \rho_{F_1}(\mathbf{r}) + \rho_{F_2}(\mathbf{r}) + \ldots + \rho_{F_{m-1}}(\mathbf{r}). \tag{78}$$

Within this approach, new local domains appear at locations where the connections occur between the fragment F and the rest M' of the molecule FM':

$$D_{-1}(G_{F\backslash \Sigma M'}(a)) = \{\mathbf{r}: \mathbf{r} \in G_F(a),\ \rho_F(\mathbf{r}) \leqslant \rho_{M'}(\mathbf{r})\}. \tag{79}$$

The boundaries of these additional domains are defined as

$$\Delta D_{-1}(G_{F\backslash \Sigma M'}(a)) = \{\mathbf{r}: \mathbf{r} \in G_{F\backslash \Sigma M'}(a),\ \rho_F(\mathbf{r}) = \rho_{M'}(\mathbf{r})\} \tag{80}$$

and can be computed by locating first all points $\mathbf{r}$ where $\rho_F(\mathbf{r}) = \rho_{M'}(\mathbf{r})$.

4.6. *Local shape analysis of interacting functional groups*

If the interactions of various molecular fragments in a macromolecule FM' are of interest, then the local shape analysis can no longer be carried out on an 'isolated'

FIDCO $G_F(a)$. In order to account for these interactions, a new contour calculation is required. The corresponding 'interactive' FIDCO $G_{F(M')}(a)$ is defined as

$$G_{F(M')}(a) = \{\mathbf{r}: \rho_F(\mathbf{r}) + \rho_{M'}(\mathbf{r}) = a,\ \rho_F(\mathbf{r}) \geqslant \rho_{M'}(\mathbf{r})\}. \tag{81}$$

No domains $D_{-1}(G_{F(M')}(a))$ are defined, since no surface is defined where a formal 'cover' of a domain representing a hole of the FIDCO $G_{F(M')}(a)$ would be found. Nevertheless, the notation $\Delta D_{-1}(G_{F(M')}(a))$ is used for the boundaries of the holes on $G_{F(M')}(a)$:

$$\Delta D_{-1}(G_{F(M')}(a)) = \{\mathbf{r}: \mathbf{r} \in G_{F(M')}(a),\ \rho_F(\mathbf{r}) = \rho_{M'}(\mathbf{r})\}. \tag{82}$$

The computation and shape analysis of the interactive FIDCOs of a fragment F in a macromolecule FM' require additional contour calculations; hence, this approach is computationally more expensive than the shape analysis of the non-interactive FIDCOs $G_{F\setminus M'}(a)$.

Using the additional domains or domain boundaries, the standard shape group approach of electron density shape analysis is applicable, providing numerical shape codes and shape similarity measures for functional groups and other local moieties of macromolecules.

5. Summary

The AFDF methods provide the basic tools for an application of the SGMs to the study of the global and local shape features of macromolecules. Complementing the topological shape analysis techniques, the SWAT method provides a simple approximate technique for the study of electron density deformations and the associated shape changes accompanying small nuclear displacements.

Acknowledgement

This work was supported by the Natural Science and Engineering Research Council of Canada.

References

[1] Hartree, D. R., 1928, *Proc. Camb. phil. Soc. math. phys. Sci.*, **24,** 111, 426; 1929, *ibid.*, **25,** 225, 310.
[2] Fock, V., 1930, *Z. physik*, **61,** 126.
[3] Hall, G. G., 1951, *Proc. R. Soc.* A, **205,** 541.
[4] Roothaan, C. C., 1951, *Rev. mod. Phys.*, **23,** 69; 1960, *ibid.*, **32,** 179.
[5] Mulliken, R. S., 1955, *J. chem. Phys.*, **23,** 1833, 1841, 2338, 2343.
[6] Mulliken, R. S., 1962, *J. chem. Phys.*, **36,** 3428.
[7] Gillespie, R. J., 1972, *Molecular Geometry* (London: Van Nostrand Reinhold).
[8] Gillespie, R. J., and Hargittai, I., 1991, *The VSEPR Model of Molecular Geometry* (Boston, Massachusetts: Allyn and Bacon).
[9] Szabo, A., and Ostlund, N. S., 1982, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory* (London: Macmillan).
[10] Frisch, M. J., Head-Gordon, M., Trucks, G. W., Foresman, J. B., Schlegel, H. B., Raghavachari, K., Robb, M. A., Binkley, J. S., González, C., DeFries, D. J., Fox, D. J., Whiteside, R. A., Seeger, R., Melius, C. F., Baker, J., Martin, R., Kahn, L. R., Stewart, J. J. P., Topiol, S., and Pople, J. A., 1990, *GAUSSIAN 90* (Pittsburgh, Pennsylvania: Gaussian, Inc.).
[11] Coppens, P., and Hall, M. B. (editors), 1982, *Electron Distribution and the Chemical Bond* (New York: Plenum).
[12] March, N. H., 1989, *Electron Density Theory of Atoms and Molecules* (New York: Academic Press).
[13] Karle, J., 1991, *Proc. natn. Acad. Sci. USA*, **88,** 10099.
[14] Pichon-Pesme, V., Lecomte, C., Wiest, R., and Benard, M., 1992, *J. Am. chem. Soc.*, **114,** 2713.

[15] WIEST, R., PICHON-PESME, V., BENARD, M., and LECOMTE, C., 1994, *J. phys. Chem.*, **98,** 1351.
[16] COLLARD, K., and HALL, G. G., 1977, *Int. J. quant. Chem.*, **12,** 623.
[17] TAL, Y., BADER, R. F. W., NGUYEN-DANG, T. T., OJHA, M., and ANDERSON, S. G., 1981, *J. chem. Phys.*, **74,** 5162.
[18] BADER, R. F. W., NGUYEN-DANG, T. T., 1981, *Adv. quant. Chem.*, **14,** 63.
[19] CIOSLOWSKI, J., 1990, *J. phys. Chem.*, **94,** 5496.
[20] CIOSLOWSKI, J., MIXON, S. T., and EDWARDS, W. D., 1991, *J. Am. chem. Soc.*, **113,** 1083.
[21] CIOSLOWSKI, J., and FLEISCHMANN, E. D., 1991, *J. chem. Phys.*, **94,** 3730.
[22] CIOSLOWSKI, J., O'CONNOR, P. B., and FLEISCHMANN, E. D., 1991, *J. Am. chem. Soc.*, **113,** 1086.
[23] CIOSLOWSKI, J., MIXON, S. T., and FLEISCHMANN, E. D., 1991, *J. Am. chem. Soc.*, **113,** 4751.
[24] CIOSLOWSKI, J., and MIXON, S. T., 1992, *Can. J. Chem.*, **70,** 443.
[25] HOHENBERG, P., and KOHN, W., 1964, *Phys. Rev.*, **136,** B864.
[26] KOHN, W., and SHAM, L. J., 1965, *Phys. Rev.*, **140,** A1133.
[27] PARR, R. G., 1975, *Proc. natn. Acad. Sci. USA*, **72,** 763.
[28] LEVY, M., 1979, *Proc. natn. Acad. Sci. USA*, **76,** 6062.
[29] LEVY, M., 1982, *Phys. Rev.* A, **26,** 1200.
[30] LUDENA, E. V., 1983, *J. chem. Phys.*, **79,** 6174.
[31] PERDEW, J. P., 1986, *Phys. Rev.* B, **33,** 8822.
[32] BECKE, A., 1986, *Phys. Rev.* A, **33,** 2786.
[33] BECKE, A., 1986, *J. chem. Phys.*, **84,** 4524.
[34] POLITZER, P., 1987, *J. chem. Phys.*, **86,** 1072.
[35] SALAHUB, D. R., 1987, *Adv. chem. Phys.*, **69,** 447.
[36] BECKE, A., 1988, *J. chem. Phys.*, **88,** 1053.
[37] BECKE, A., 1988, *Phys. Rev.* A, **38,** 3098.
[38] PARR, R. G., 1988, *J. phys. Chem.*, **92,** 3060.
[39] TACHIBANA, A., 1988, *Int. J. quant. Chem.*, **34,** 309.
[40] TACHIBANA, A., 1988, *High Temperature Superconducting Materials*, edited by W. E. Hatfield and J. H. Miller (New York: Dekker), pp. 86–96.
[41] PARR, R. G., and YANG, W., 1989, *Density Functional Theory of Atoms and Molecules* (Oxford: Clarendon).
[42] KRYACHKO, E. S., and LUDENA, E. V., 1989, *Density Functional Theory of Many-Electron Systems* (Dordrecht: Kluwer).
[43] ZIEGLER, T., 1991, *Chem. Rev.*, **91,** 651.
[44] PÁPAI, I., GOURSOT, A., ST-AMANT, A., and SALAHUB, D. R., 1992, *Theor. chim. Acta*, **84,** 217.
[45] LABANOWSKI, J. K., and ANDZELM, J. (editors), 1991, *Density Functional Methods in Chemistry* (New York: Springer).
[46] ANDZELM, J., and WIMMER, E., 1992, *J. chem. Phys.*, **96,** 1280.
[47] SEMINARIO, J. M., and POLITZER, P., 1992, *Int. J. quant. Chem. Symp.*, **26,** 497.
[48] MEZEY, P. G., 1995, *Molecular Similarity*, Topics in Current Chemistry, Vol. 173, edited by K. Sen (Berlin: Springer), pp. 63–83.
[49] MEZEY, P. G., 1995, *Molecular Similarity in Drug Design*, edited by P. M. Dean (Glasgow: Chapman & Hall–Blackie), pp. 241–268.
[50] MEZEY, P. G., 1995, *Structural Chem.*, **6**, 261.
[51] MEZEY, P. G., 1996, *Adv. Quant. Chem.*, **27,** 163.
[52] MEZEY, P. G., 1996, *Computational Chemistry: Reviews and Current Trends*, Vol. 1 edited by J. Leszczynski (Singapore: World Scientific), pp. 109–137.
[53] WALKER, P. D., and MEZEY, P. G., 1993, *J. Am. chem. Soc.*, **115**, 12423.
[54] WALKER, P. D., and MEZEY, P. G., 1994, *J. Am. chem. Soc.*, **116**, 12022.
[55] WALKER, P. D., and MEZEY, P. G., 1994, *Can. J. Chem.*, **72,** 2531.
[56] WALKER, P. D., and MEZEY, P. G., 1995, *J. math. Chem.*, **17,** 203.
[57] WALKER, P. D., and MEZEY, P. G., 1995, *J. comput. Chem.*, **16,** 1238.
[58] MEZEY, P. G., ZIMPEL, Z., WARBURTON, P., WALKER, P. D., IRVINE, D. G., DIXON, D. G., and GREENBERG, B., 1996, *J. chem. Inf. Comput. Sci.*, **36,** 602.
[59] WALKER, P. D., and MEZEY, P. G., 1993, *Program MEDLA 93* (Mathematical Chemistry Research Unit, University of Saskatchewan).

[60] MEZEY, P. G., 1995, *J. math. Chem.*, **18,** 141.
[61] MEZEY, P. G., 1996, *Adv. molec. Similarity*, **1,** 89.
[62] MEZEY, P. G., 1997, *Int. J. quant. Chem.*, **63**, 39.
[63] MEZEY, P. G., 1995, *Program ADMA 95* (Mathematical Chemistry Research Unit, University of Saskatchewan).
[64] MEZEY, P. G., 1992, *J. chem. Inf. Comp. Sci.*, **32,** 650.
[65] MEZEY, P. G., 1993, *Shape in Chemistry: An Introduction to Molecular Shape and Topology* (New York: VCH).
[66] MEZEY, P. G., 1994, *Can. J. Chem.*, **72,** 928.
[67] MEZEY, P. G., 1990, *Reviews in Computational Chemistry*, edited by K. B. Lipkowitz and D. B. Boyd (New York: VCH), pp. 265–294.
[68] MEZEY, P. G., 1996, *Program SWAT 96* (Mathematical Chemistry Research Unit, University of Saskatchewan).
[69] MEZEY, P. G., 1986, *Int. J. quant. Chem., quant. Biol. Symp.*, **12,** 113.
[70] MEZEY, P. G., 1987, *J. comput. Chem.*, **8,** 462.
[71] MEZEY, P. G., 1987, *Int. J. quant. Chem., quant. Biol. Symp.*, **14,** 127.
[72] MEZEY, P. G., 1988, *J. math. Chem.*, **2,** 325.
[73] ZIMPEL, Z., and MEZEY, P. G., 1996, *Int. J. quant. Chem.*, **59,** 379.
[74] PILAR, F. L., 1968, *Elementary Quantum Chemistry* (New York: McGraw-Hill).
[75] MCWEENY, R., and SUTCLIFFE, B. T., 1969, *Methods of Molecular Quantum Mechanics* (New York: Academic Press).
[76] LÖWDIN, P. O., 1950, *J. chem. Phys.*, **18,** 365.
[77] LÖWDIN, P. O., 1956, *Adv. Phys.*, **5,** 1.
[78] LÖWDIN, P. O., 1970, *Adv. quant. Chem.*, **5,** 185.
[79] MASSA, L., HUANG, L., and KARLE, J., 1995, *Int. J. quant. Chem., quant. Biol. Symp.*, **29,** 371.
[80] HELLMANN, H., 1937, *Einführung in die Quantenchemie* (Leipzig: Deuticke), Section 54.
[81] FEYNMAN, R. P., 1939, *Phys. Rev.*, **56,** 340.
[82] EPSTEIN, S. T., 1981, *The Force Concept in Chemistry*, edited by B. M. Deb, (New York: Van Nostrand Reinhold).
[83] MEZEY, P. G., 1990, *Reports in Molecular Theory* edited by H. Weinstein and G. Naray-Szabo (Boca Raton, Florida: CRC Press), pp. 165–183.
[84] MEZEY, P. G., 1990, *Concepts and Applications of Molecular Similarity* edited by M. A. Johnson and G. M. Maggiora (New York: Wiley), pp. 321–368.
[85] ARTECA, G. A., and MEZEY, P. G., 1992, *Chem. Phys.*, **161,** 1.
[86] WALKER, P. D., ARTECA, G. A., and MEZEY, P. G., 1993, *J. comput. Chem.*, **14,** 1172.
[87] WALKER, P. D., MAGGIORA, G. M., JOHNSON, M. A., PETKE, J. D., and *Mezey, P. G.*, 1995, *J. chem. Inf. Comput. Sci.*, **35,** 568.
[88] WALKER, P. D., MEZEY, P. G., MAGGIORA, G. M., JOHNSON, M. A., and PETKE, J. D., 1995, *J. comput Chem.*, **16,** 1474.
[89] HEAL, G. A., WALKER, P. D., RAMEK, M., and MEZEY, P. G., 1996, *Canad. J. Chem.*, **74,** 1660.
[90] MEZEY, P. G., and WALKER, P. D., 1997, *Drug Discovery Today*, **2**, 6.