J Math Chem (2009) 45:544–549
DOI 10.1007/s10910-008-9428-x

ORIGINAL PAPER

# QSAR and the ultimate molecular descriptor: the shape of electron density clouds

**Paul G. Mezey**

Published online: 8 August 2008

**Abstract** The advantages of electron density representations of molecules in QSAR studies and the related shape descriptors and shape similarity measures are discussed, with special emphasis on molecules involving pi-bonded systems and aromatic rings.

**Keywords** QSAR · Molecular descriptors · Shape · Electron density

## 1 Introduction

Quantitative structure–activity relations, QSAR, interconnect two subfields of molecular informatics: information on molecular structure and information on molecular properties, as manifested in molecular activities. Hence, quantitative structure–activity relations are fundamentally dependent on the information content of molecules. Traditionally, there have been several approaches employed to extract, represent, and analyze such information, prominent among the used representations are the molecular graph describing the essence of bonding pattern, and the so-called space-filling models, such as fused-sphere models, representing in a simplistic way the spatial relations of molecular fragments within and between molecules. These traditional approaches have experienced considerable developments and are widely used, yet in recent years it has become increasingly evident, as well as practically useful, that molecular electron densities fully contain all molecular information, hence it is natural

P. G. Mezey (✉)
Canada Research Chair in Scientific Modeling and Simulation, Department of Chemistry
and Department of Physics and Physical Oceanography, Memorial University of Newfoundland,
283 Prince Philip Drive, St. John's, NL, Canada A1B 3X7
e-mail: paul.mezey@gmail.com

P. G. Mezey
Institute for Advanced Study, Collegium Budapest, Szentháromság u. 2, 1014 Budapest, Hungary

to consider their direct use in QSAR studies. Since molecules contain nothing else than atomic nuclei and an electron density cloud, the latter representing the actual, fuzzy bodies of molecules, the use of electron density as an actual molecular descriptor in QSAR guards against oversimplification and offers some guaranties for quality. Specifically, it is only the limitations to the accuracy of the actual electronic density description that may be responsible if the model is missing some aspects of relevant molecular information. Hence, it is not surprising that electron density modeling has been shown to serve the needs of molecular informatics better than most of the earlier models. In this report one relevant aspect of this development, the marked differences in the quality of the representation of molecules involving pi bond (and by extrapolation, delta bond) and aromatic ring contributions to the molecular body are discussed, if the representations involve the actual electronic cloud of the molecular electron density, or if it is based on the simpler, fused-sphere models.

The usefulness of electron density representation as the source of information on molecular structure is rather evident by exclusion: as mentioned, molecules contain nothing else but atomic nuclei and the electron density cloud, where the latter also gives full information about the location and nature of nuclei. Hence, all molecular information must be present in the electron density. A precise formulation of this principle is found in the Hohenberg-Kohn theorem of density functional theory [1] and for molecular parts, in the holographic electron density theorem [2,3], asserting the global and local information carrying properties of molecular electron densities. It is this central role of electron density that makes it an excellent tool for the study of QSAR.

In addition to the comparisons and the chemically significant differences between electron density and fused sphere representations, two electron density approaches will also be discussed briefly, one focusing on functional groups [4] within molecules, the other providing a direct link between traditional, molecular graph approaches and one based on the algebraic-topological shape analysis approach [5] to QSAR.

## 2 Electron density clouds and fused-sphere models as molecular representations

Various quantum chemical computational methods, for example, traditional Hartree-Fock computations employing atomic orbital basis sets of various quality, composed from a few or, alternatively, a large number of atomic basis functions, or higher level, post-Hartree-Fock correlated wavefunctions, or coupled cluster methods, provide various levels of accuracy, well demonstrated in the quality of computed molecular wavefunctions, and even more importantly, in energy calculations. Nevertheless, if one is concerned with the quality of shape representations of molecules in terms of electron density, for example, by a series of molecular isodensity contours, MIDCO's, even relatively low level (e.g. small basis set Hartree-Fock) computations provide shape representations of electron densities acceptable for most shape analysis purposes, especially in the low density regions.

A molecular isodensity contour surface, MIDCO G(K,a) for some electron density threshold value a and a specified nuclear configuration K is defined as the collection of all points r of the 3D space where the electronic density $\rho$(r) of the molecule of nuclear configuration K is equal to the electron density threshold value a. If it is essential to specify the nuclear configuration K of the molecule when we discuss the electron density cloud, then the notation $\rho$(K,r) will be used for the electron density. Keeping these considerations in mind, the expression

$$\mathrm{G(K,a)} = \{\mathrm{r} : \rho(\mathrm{K,r}) = \mathrm{a}\} \tag{1}$$

defines the corresponding MIDCO G(K,a).

As it has been shown in detail, ab initio quality but smaller basis set calculations using the Hartree-Fock method do show noticeable errors in the electron density ranges of larger values, near the nuclei and even in the bonding regions of molecules. Nevertheless, in the lower ranges of chemically relevant density values these computations provide useful results. This is the case, for example, in the outer isodensity contours of valence regions, typically represented by MIDCO's of electron density threshold values falling within the threshold interval

$$0.001\,\mathrm{a.u.} \leq \mathrm{a} \leq 0.01\,\mathrm{a.u.} \tag{2}$$

(where 1 a.u. = 1 e/bohr$^3$ is the atomic unit for electron density), where even lower level computations using small basis sets, for example, the 4–31G Gaussian basis set, provide reasonably good quality shape representations of the electron density.

In quantum chemical calculations the angular parts of the atomic orbitals are well represented, since the approximations usually affect more the radial parts of these functions, especially, if instead of Slater type orbitals of correct cusp condition (correct behavior of the radial derivative near the nuclei) are replaced by a set of Gaussian functions, which by their very definition do not fulfill the cusp condition. Similarly, the long range behavior of Gaussian functions also deviates from the correct solutions of atomic wavefunctions. Nevertheless, the angular representations are correct, hence most ab initio quality quantum chemistry methods provide proper approximations to account for both the sigma and pi type bonding contributions within molecules.

It follows, that these and higher quality quantum chemical electron density representations also provide a valid picture of the contributions of pi-bonds, conjugated bond systems, and aromatic rings to both the local and the overall molecular shape and the general spatial structure of molecules.

The situation is rather different if one deals with the still very popular and widely used fused-sphere representations of molecular bodies. Here individual atoms are represented by spheres, and the radii of these spheres are chosen traditionally to provide some level of interpenetration if two such spheres are positioned with their centers having a distance close to the typical bond distance for the given atom pair. In this sense, the spheres are considered fused, and the resulting fused object represents some approximate image of a local molecular range for some intermediate threshold MIDCO of the molecule. However, it is immediately obvious that a pair of spheres placed in any manner becomes an object of cylindrical symmetry, with the special case of even higher, spherical symmetry for the chemically irrelevant positioning of coincident centers, when one obtains spherical symmetry and the size of the larger sphere. Consequently, only sigma type, that is, locally cylindrical bonding symmetry can be described by such fused-sphere models. This, evidently, excludes a major, rather important aspect of chemical bonding: all pi bonding and aromatic bonding features, as well as the less significant delta bond features. This rather dramatic shortcoming of the fused sphere models still often used for the shape representations of molecules is evident but seldom mentioned, and it is troubling that such a simplistic approach, completely ignoring the spatial influence of pi bonds is often used for interpreting molecular interactions and biochemical shape conditions, for example, the spatial conditions for enzyme activity.

Today the reasonably accurate quantum chemical computation of electronic density is no longer an excessively time-consuming task, and such, ab initio quality computations are available even for proteins of well over a thousand atoms, for example, by employing the linear-scaling Adjustable Density Matrix Assembler (ADMA) method [6–11], based on the earlier numerical approach of the Molecular Electron Density Loge Approach (MEDLA) [12–14].

Whereas fused-sphere models are still widely used, and they do convey a crude representation of the fuzzy molecular body of the electron density cloud, the fact that they miss very significant bonding contributions, such as those due to pi-bonds and aromatic rings, is often ignored when interpreting these models. The electron density considerations discussed above strongly suggest that it is advantageous and also feasible to replace the traditional fused-sphere representations of molecular bodies with actual, quantum-chemically computed electron density representations, especially, if, as it is most often the case, pi-bonded and aromatic ring contributions are relevant to the chemical problem. With the linear scaling quantum chemistry methods such as the ADMA technique [6–11], electron density calculations are feasible even for biological macromolecules.

## 3 Shape similarity measures of electron densities as input information for QSAR

For the shape comparisons of electron density clouds within the QSAR framework one may follow essentially two main approaches: either considering the shapes of the complete molecules, or focusing on some local ranges which are likely to be responsible for the activity considered. In the latter case the local range is the vicinity of one or several functional groups.

The shape similarity measures the most extensively tested for fuzzy electron density clouds are based on the shape group methods [5] which are providing numerical shape codes obtained from an algebraic-topological analysis of curvature properties for an entire range of MIDCO's. As it has been shown, these shape group methods and the derived shape similarity measures are equally applicable for complete molecules or for molecular fragments, such as functional groups within molecular series [15].

In electron density analysis the definition of functional groups is related to the definition of density domains. A Density Domain DD(K,a) is a subset of the molecular electron density cloud enclosed by a MIDCO G(K,a)

$$\mathrm{DD(K,a)} = \{\mathrm{r} : \rho(\mathrm{K,r}) \geq \mathrm{a}\}. \tag{3}$$

The density domain approach has been suggested for a quantum chemical representation of formal *functional groups*. One may start with a useful analogy: consider two separate but slightly interacting molecules near one another. As long as these molecules have separate identity, each must have some Density Domain containing all the nuclei of the molecule, but none of the nuclei of the other molecule. Separate identity is in fact manifested by the presence of such density domains.

Consider now a molecule and one of its connected density domains DD(K,a) and the nuclei enclosed by it. This subset of the nuclei of the molecule is separated from the rest of the nuclei by the boundary of the density domain DD(K,a), that is, by the MIDCO G(K,a). This very fact indicates that these nuclei, together with the local electronic density cloud surrounding them, represent a sub-entity of the molecule, with limited autonomy, and some degree of individual identity. Of course, this separate identity is not as marked as it is for the previous example of two, only slightly interacting molecules, but the principles are the same for the separation condition, although at some different level, indicating some limited autonomy.

Consequently, it is natural to regard the existence of such a density domain DD(K,a) as a criterion and a representative of a formal *functional group*. In practice, the above condition is used only to identify a subset of nuclei that belongs to such a density domain, and the entire associated fuzzy electron density contribution, as obtained from the fuzzy density fragmentation technique of the ADMA method, is regarded as the local, fuzzy electron density fragment of the functional group.

The electron density shape analysis and similarity evaluation methods applied to complete molecules or to fuzzy density fragments, such as those of functional groups, will not be described here and the reader is referred to the literature quoted above. Instead, the following discussion will be focused on the choice of reference shapes within a series, against which all other shapes are compared.

We are concerned with the continuum nature of electron densities, where these continua are reduced to a simpler model which lead to numerical shape codes, which themselves can be considered as high-dimensional vectors. In the typical applications of the shape group methods, to either complete molecules or to molecular fragments such as functional groups, the shape codes can be represented as $41 \times 21$ matrices or as vectors of 861 components.

One may follow two strategies, depending on the actual QSAR problem. If there is a lead compound, or a compound with extreme level of activity with reference to the actual biological effect, then it is useful to use this compound, or, the local range of the relevant functional group of this molecule as shape reference, and then the correlations are generated in terms of differences in shape codes as well as differences in activities, using the reference molecule for each comparison. It is likely that the shape feature characteristic to the given type of activity is most prominent in the molecule, or in the local range of molecule with the highest level of activity.

On the other hand, if the task is to generate a survey of a family of molecules for a whole set of potential biological activities, then it is apparently advantageous if the shape comparisons are obtained with reference to some "average" molecule within the family. Whereas in most situations the molecules show considerable shape diversity within the family studied, for a given functional group likely to be involved in several biochemical reaction the task is simpler, demonstrating the advantage of local shape analysis. It is well known that molecules with rather diverse global shapes are likely to show local shape similarities, e.g. shape similarities of some local functional groups, if they show similar activities in a given biochemical process. Yet, if the survey is extended to several types of activities, it is useful to take some "average" shape for the given type of functional group as reference.

Note that this can be achieved either by finding an actual molecule that shows a functional group shape that is approximately "average" according to shape code comparisons, or by generating an actual average shape for the given functional group, that is the actual average with respect to the given molecular family, all possessing the same functional group, locally distorted by the rest of each molecule. Whereas methods for the actual generation of such average conformations and the associated average electron densities are described elsewhere [16,17], the special requirements of QSAR provide motivation for further potential developments in this field.

**Acknowledgements** The financial support of the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Canada Research Chairs Program, Canada Foundation for Innovation, Memorial University of Newfoundland, RET Program of Eötvös University of Budapest, Hungary, and the Institute for Advanced Study, Collegium Budapest, Hungary are gratefully acknowledged.

## References

1. P. Hohenberg, W. Kohn, Inhomogeneous electron gas. Phys. Rev. **136**, B864–B871 (1964)
2. P.G. Mezey, The holographic electron density theorem and quantum similarity measures. Mol. Phys. **96**, 169–178 (1999)
3. P.G. Mezey, Holographic Electron Density Shape Theorem and Its Role in Drug Design and Toxicological Risk Assessment, J. Chem. Inf. Comp. Sci. **39**, 224–230 (1999).
4. P.G. Mezey, Functional groups in quantum chemistry. Adv. Quantum Chem. **27**, 163–222 (1996)
5. P.G. Mezey, Shape in chemistry: an introduction to molecular shape and topology (VCH Publishers, New York, 1993)
6. P.G. Mezey, Macromolecular density matrices and electron densities with adjustable nuclear geometries. J. Math. Chem. **18**, 141–168 (1995)
7. P.G. Mezey, Quantum similarity measures and Löwdin's transform for approximate density matrices and macromolecular forces. Int. J. Quantum Chem. **63**, 39–48 (1997)
8. P.G. Mezey, Quantum chemistry of macromolecular shape. Int. Rev. Phys. Chem. **16**, 361–388 (1997)
9. P.G. Mezey, Computational microscopy: pictures of proteins. Pharm. News **4**, 29–34 (1997)
10. T.E. Exner, P.G. Mezey, Ab initio quality properties for macromolecules using the ADMA approach. J. Comput. Chem. **24**, 1980–1986 (2003)
11. Zs. Szekeres, T.E. Exner, P.G. Mezey, Fuzzy fragment selection strategies, basis set dependence, and HF–DFT comparisons in the applications of the ADMA method of macromolecular quantum chemistry, Internat. J. Quantum Chem. **104**, 847–860 (2005)
12. P.D. Walker, P.G. Mezey, Ab initio quality electron densities for proteins: a MEDLA approach. J. Am. Chem. Soc. **116**, 12022–12032 (1994)
13. P.D. Walker, P.G. Mezey, realistic, detailed images of proteins and tertiary structure elements: ab initio quality electron density calculations for bovine insulin. Can. J. Chem. **72**, 2531–2536 (1994)
14. P.D. Walker, P.G. Mezey, A new computational microscope for molecules: high resolution MEDLA images of taxol and HIV-1 protease, using additive electron density fragmentation principles and fuzzy set methods. J. Math. Chem. **17**, 203–234 (1995)
15. P.G. Mezey, Functional groups in quantum chemistry. Adv. Quantum Chem, **27**, 163–222 (1996)
16. P.G. Mezey, Averaged electron densities for averaged conformations. J. Comput. Chem. **19**, 1337–1344 (1998)
17. P.G. Mezey, Distributions and averages of molecular conformations. Comput. Chem. **25**, 69–75 (2001)