ELSEVIER

Journal of Molecular Structure: THEOCHEM 727 (2005) 57–61

THEO CHEM

www.elsevier.com/locate/theochem

The use of momentum-space descriptors for predicting octanol–water partition coefficients

Jabir H. Al-Fahemi[a], David L. Cooper[a,*], Neil L. Allan[b]

[a]*Department of Chemistry, University of Liverpool, Liverpool L69 7ZD, UK*
[b]*School of Chemistry, University of Bristol, Cantocks Close, Bristol BS8 1TS, UK*

Received 26 October 2004; accepted 10 November 2004
Available online 28 June 2005

Abstract

We extend previous work on quantitative structure–activity and structure–property relationships using molecular descriptors based on quantities determined from the momentum-space (*p*-space) electron density. In particular, we introduce new molecular descriptors that are related to the *p*-space information entropy. We also examine the use of a simple molecular shape descriptor, *X*. For the LC_{50} toxicity data of a series of saturated alcohols, the simple shape descriptor is found to be remarkably successful, and a correlation is observed between *X* and the entropy-like *p*-space descriptors. We develop a promising 13-descriptor regression model for the log *P* values of a set of 76 chemically diverse molecules. Similar approaches, combining simple classical parameters with *p*-space quantities, are likely to prove useful for more complex QSAR/QSPR problems.

Keywords: Momentum-space; Octanol–water; Partition coefficient

1. Introduction

Octanol–water partition coefficients, expressed as log *P*, are a mainstay of quantitative structure–activity (QSAR) and structure–property (QSPR) relations, and so a wide variety of reliable schemes have been developed for estimating log *P*. These are of course especially important when no reliable experimental data are available, as is particularly likely to be the case for newly synthesized molecules and, indeed, for systems that have not yet been synthesized. Amongst the most successful of these predictive schemes are those based on fragment decomposition, summing fixed values for molecular fragments, together with various correction factors. In parallel with such widely used approaches there have been various attempts to use information derived from position-space quantum mechanical calculations, such as various classes of computed molecular quantities and even quantum molecular similarities. The present work is concerned with a new strategy of this type, but based instead on momentum-space quantities.

As is well known, momentum-space (*p*-space) electron densities may be readily obtained from conventional position-space (*r*-space) wave functions. Not only do such *p*-space descriptions provide interesting alternative representations of the bonding in small and large molecules, but they have also proved useful in a wide range of molecular similarity studies. The success of those studies motivated McCoy and Sykes [1] to investigate the utility for QSAR/QSPR work of molecular descriptors derived from *p*-space electron densities. Following on from their papers, we have successfully employed expectation values of powers of *p* in QSAR/QSPR studies of gas chromatography retention times, gas–hexadecane partition coefficients, various toxicity data, and so on [2]. An important finding, however, of such work was that values of $\langle p^n \rangle$, augmented with the molecular weight, did not perform significantly better than rather trivial one-dimensional conventional descriptors that are linked to the molecular composition and connectivity. Nonetheless, our best results were obtained with hybrid models that combined conventional and *p*-space descriptors.

* Corresponding author. Tel.: +44 151 794 3532; fax: +44 151 794 3588.
E-mail address: dlc@liv.ac.uk (D.L. Cooper).

0166-1280/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.theochem.2005.02.016

The focus of the present work is an investigation of the utility of augmenting our existing descriptors with a molecular shape descriptor and with two new classes of *p*-space quantities that are linked to the molecular information entropy. Our main aim is the identification of suitable new *p*-space descriptors that might prove useful in more challenging QSAR studies, such as the partitioning of organic solutes across the blood–brain barrier. We recognize, of course, that the prediction of log *P* is, to a large extent, already a solved problem, but it does provide a useful testing ground for our proposed new descriptors.

2. Calculations

We used geometry optimisations with the AM1 procedure in MOPAC [3,4] to generate conventional semi-empirical *r*-space wave functions. As we have described many times before [5], it is then entirely straightforward to generate from these the corresponding momentum-space electron density $\rho(\mathbf{p})$. All of the *p*-space quantities required for our QSAR/QSPR or similarity studies may then be generated using appropriate numerical integration schemes. We could of course have used somewhat more sophisticated wave functions, as in some of our previous work, but we are mindful here of our ultimate goal of processing reasonably rapidly large numbers of relatively large molecules.

Two genres of *p*-space descriptor were assessed. As in previous work [2], we considered first of all expectation values of power of *p*, defined according to:

$$\langle p^n \rangle = \int p^n \rho(\mathbf{p}) \mathrm{d}\mathbf{p} \tag{1}$$

Of course, $\langle p^2 \rangle$ is determined by the kinetic energy which, according to the virial theorem, can be related to the total molecular energy. Similarly, $\langle p^0 \rangle$ should coincide with the total number of electrons treated explicitly in the electronic structure calculations, i.e. with the integral of the electron density over all space. As such, it makes sense to use instead of those two integrals the total energy *E* from the AM1 calculation and the total number of (valence) electrons. The only moment of momentum which we evaluate explicitly is $\langle p^{-2} \rangle$. These three quantities, augmented by the molecular weight, were found to be sufficient to generate useful regression models for a number of properties and activities [2].

While considering the plausible definitions of additional *p*-space descriptors that might provide new information, we took account of the work of Gadre [6], and subsequently of various other authors, including ourselves [7,8], on so-called information 'entropies' of the form:

$$S = -\int \rho(\mathbf{p}) \ln \rho(\mathbf{p}) \mathrm{d}\mathbf{p} \tag{2}$$

For the purpose of generating a *family* of descriptors, it makes sense to generalize this expression by means of the incorporation of a power of *p*. Our new 'entropy-like' descriptors may thus be defined according to:

$$S_n = -\int p^n \rho(\mathbf{p}) \ln \rho(\mathbf{p}) \mathrm{d}\mathbf{p} \tag{3}$$

An obvious alternative could be to scale $\rho(\mathbf{p})$ by N^{-1}, thereby reducing somewhat the overall dependence of such quantities on the total number of electrons, *N*. With this in mind, we define also a further family of entropy-like descriptors according to:

$$\acute{S}_n = -\int p^n \left(\frac{1}{N}\rho(\mathbf{p})\right) \ln\left(\frac{1}{N}\rho(\mathbf{p})\right) \mathrm{d}\mathbf{p} \tag{4}$$

These two families of descriptors are of course trivially related to one another, according to

$$\acute{S}_n = \frac{1}{N}(\langle p^n \rangle \ln N + S_n) \tag{5}$$

but it seemed appropriate to investigate the utility, or otherwise, of both sets, choosing our usual *n* values of -2, 0 and $+2$.

It seems almost inevitable that the shape [9] of a given molecule could have a significant bearing on many of the types of properties and activities that we wish to be able to study. Accordingly, in addition to our various new *p*-space descriptors, it seemed appropriate to consider a very simple shape descriptor. For a given molecular volume *V*, the surface area σ achieves its smallest value for a sphere and so the dimensionless ratio

$$X = 36\pi\left(\frac{V^2}{\sigma^3}\right) \tag{6}$$

could be a useful measure of the deviation from a spherical shape. Various simple strategies are available for estimating approximate molecular volumes and surface areas, and so the evaluation of such an index is in principle very straightforward. Numerical values of *X* lie in the range from unity, for a sphere, down towards zero.

3. Results and discussion

3.1. Saturated alcohols

Romanelli et al. [10] modelled log LC_{50} toxicity data for fathead minnows of a series of 12 saturated alcohols. As descriptors, they used calculated values of molecular surface area, molecular volume, log *P*, molar refractivity and polarizability. They examined more than a dozen linear, quadratic and cubic models, achieving correlation coefficients R^2 which ranged from 0.993 to 0.999.

As a first step towards developing useful regression models based on our own descriptors, we decided to examine how well (or otherwise) we can predict log LC_{50} for this series of molecules using nothing more than the simple shape descriptor, *X*, defined in Eq. (6). We were somewhat surprised by the relatively high quality of

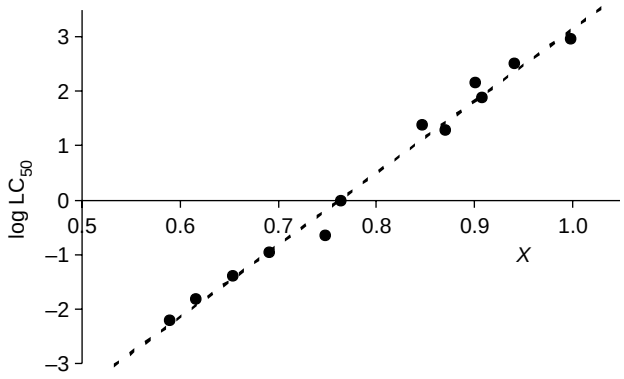

Fig. 1. Apparent correlation of log LC_{50} with the simple shape descriptor X, for 12 saturated alcohols.

the resulting fit (see Fig. 1), which is characterized by $R^2 = 0.987$ and a standard error $\Delta = 0.218$. We found that X also provides useful one-descriptor fits to the log P values used by Romanelli et al. This, however, turns out mostly to be an artefact: some of their calculated log P values differ somewhat from the experimental ones, and we find that we do not achieve such a high quality one-descriptor fit to the experimental data. Even so, the apparent utility of X for modelling the log LC_{50} data is quite remarkable.

We wondered, of course, whether there might be any significant correlations between our simple shape descriptor and the various families of p-space descriptors that we intend to use. After a little experimentation, we found for this series of 12 saturated alcohols that the following linear model based on four of the entropy-like p-space descriptors predicts X with $R^2 = 0.976$ and $\Delta = 0.026$:

$$X = b_1 + b_2\acute{S}_0 + b_3\acute{S}_{-2} + b_4 S_{+2} + b_5\acute{S}_{+2} \tag{7}$$

Indeed, reasonable fits can also be achieved with smaller numbers of these S_n and $\acute{S}_n$ quantities, but we found that none of the individual descriptors varies linearly with X or with, say, the number of carbon atoms. A downside of these various observations is that we are unlikely to achieve very much by incorporating X into QSAR/QSPR models that already include our various S_n and $\acute{S}_n$ descriptors.

3.2. Homologous series

Amat et al. [11] evaluated the overlap integral

$$Z_{AA} = \int \rho_A^{oct}(\mathbf{r})\rho_A^{aq}(\mathbf{r})d\mathbf{r} \tag{8}$$

in which $\rho_A^{oct}(\mathbf{r})$ and $\rho_A^{aq}(\mathbf{r})$ are the r-space electrons densities of molecule A in models of octanol and water, respectively. Those authors did not, however, reoptimize the geometries in the ab initio calculations that were done for different 'solvents'. For each of 12 different series of molecules, they found excellent correlations between experimental values of log P and their calculated values of Z_{AA}. The various R^2 values are reported in Table 1. This apparent success is, however, marred by the observation that there appears to be no useful correlation when one considers together all 58 molecules (see Table 1).

Table 1
Correlation coefficients R^2 between Z_{AA} and log P for 12 series of molecules, as reported in Ref. [11]

Series	N_{mol}	R^2
A: RH	6	0.996
B: RNH_2	4	0.996
C: ROH	6	0.996
D: RCH(OH)R′	6	0.996
E: RCOR′	6	0.993
F: CH_3COOR	5	0.996
G: RCOOH	5	0.998
H: HCONHR	5	0.999
I: $RCONH_2$	5	0.999
J: RCl	4	0.998
K: $Cl(CH_2)_nOH$	3	0.981
L: $RCH(NH_2)COOH$	4	0.986
All of the above	58	0.005

N_{mol} is the number of molecules in a given series. Data for these same series of molecules are illustrated in Fig. 2.

The underlying reasons for the success of the individual correlations become much clearer when one plots the numerical values of Z_{AA} from Ref. [11] against the number of carbon atoms, n_C, as is done in Fig. 2a. In general, the different homologous series have different starting points, but there is then a fairly constant increment to the value of Z_{AA} of ca. 17 units for each additional carbon atom; each of the separate correlations of Z_{AA} with n_C is characterized by an R^2 value of practically unity. The corresponding plot of log P against n_C (Fig. 2b) also shows linear behaviour for each separate class of molecule, with R^2 values that are very close to those reported by Amat et al. [11]. It is now easy to see that the separate good correlations between log P and Z_{AA} were really the outcome of the way that each of these quantities increases in a regular fashion with increasing chain length for a given homologous series.

It seems that these Z_{AA} integrals are far more sensitive to the overall 'size' of the system than they are to any differences between $\rho_A^{oct}(\mathbf{r})$ and $\rho_A^{aq}(\mathbf{r})$. It could be interesting in due course to try to use Z_{AA} integrals, or their p-space counterparts, to investigate normalized similarity measures, or suitably defined dissimilarity indices, that really measure the differences between $\rho_A^{oct}(\mathbf{r})$ and $\rho_A^{aq}(\mathbf{r})$. However, it remains possible that these differences could be too small, for fixed geometry, for useful correlations to be established. Circumstantial evidence for this is provided by the outcome of an attempt to predict all the values of the 'raw' Z_{AA} integrals that were reported in Ref. [11]. For this purpose we used as descriptors not only the total number of carbon atoms, n_C, but also the corresponding numbers (n_X) of atoms of all the different elements. The resulting linear regression

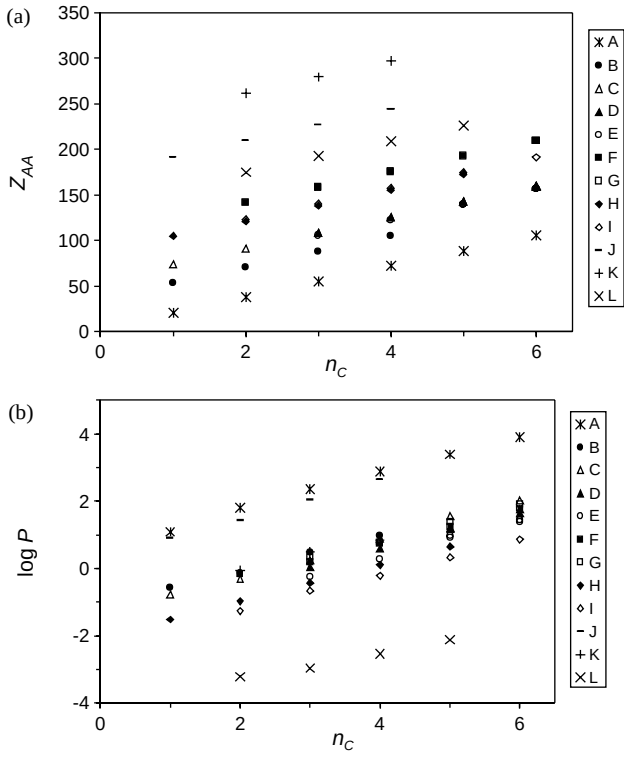

Fig. 2. Variation of Z_{AA} and of log P with the total number of carbon atoms, n_C, for the series of molecules listed in Table 1. All data were taken from Ref. [11].

model is characterized by $R^2=1.000$ and $\Delta=0.560$.[1] Various quantities such as the electron–electron repulsion energy (V_{ee}) have also been suggested as descriptors for QSAR studies, but as noted by Gironés et al. [12] in a study of a series of amides, molecular values of V_{ee} are determined to a large extent by the types and numbers of atoms. This is certainly also true of the values of Z_{AA}.

It is clear that we should be very cautious whenever rather good results for individual homologous series do not translate into a single reliable model for treating together *all* of the different series. Our results for the log LC_{50} of saturated alcohols, discussed earlier, are open to similar criticism. In order to test the reliability, or otherwise, of our various families of *p*-space descriptors it is essential that we examine a single set of structurally-diverse molecules.

3.3. Structurally-diverse molecules

Our starting point for the construction of a suitable data set was the work of Bodor et al. [13], who modelled values of log P for 118 organic compounds, achieving a correlation coefficient $R^2=0.939$ and a standard error $\Delta=0.296$. They used a regression model based on σ^2, σ, O^2, O, I_{alkane}, molecular weight, calculated dipole moment, Q_{ON}, Q_N^4, Q_N^2, Q_N, Q_O^4, Q_O^2 and Q_O, where Q_N and Q_O are the square root of the sum of squared charges on nitrogen atoms and oxygen atoms, respectively, and Q_{ON} is the sum of absolute values of atomic charges on nitrogen and oxygen atoms. I_{alkane} is an indicator variable for alkanes (its value is unity if the molecule is an alkane, otherwise zero) and the quantity O, which they called the 'ovality', is a shape descriptor defined in terms of the volume V and surface area σ. We find that O is closely related to our own simple shape descriptor (Eq. (6)), according to $X=O^{-3}$.

We selected 42 molecules from the set considered by Bodor et al. [13], namely propane, pentane, cyclohexane, methanol, ethanol, propan-1-ol, butan-1-ol, pentan-1-ol, hexan-1-ol, octan-1-ol, dimethyl ether, diethyl ether, acetone, acetic acid, propanoic acid, butanoic acid, hexanoic acid, methyl acetate, ethyl acetate, acetonitrile, furan, pyrrole, pyridine, fluoromethane, chloromethane, dichloromethane, trichloromethane, tetrachloromethane, chloroethane, phenol, aniline, methylamine, butylamine, diphenylamine, benzene, toluene, naphthalene, benzaldehyde, biphenyl, butanone, acetophenone and hydroquinone. This set, which we will call 'set A', does not include any amino acids. It is our expectation that such systems could behave somewhat differently from the others on account of the likelihood of zwitterionic character in aqueous solution. Some of the molecules in our set A already appear in the list considered by Amat et al. [11]. Merging the two lists, but excluding the four amino acids considered in Ref. [11], produces a set of 76 molecules, which we call 'set B'.

In building our various regression models, we considered various families of descriptors and then various combinations of those families. However, we report here only a summary of our various numerical experiments, concentrating on those sets of descriptors which led eventually to the best of our models. Our findings were much the same whether we considered the smaller list of molecules, set A, or the larger one, set B.

The simplest useful group of descriptors consists primarily of n_C, n_H, n_O, n_N, n_{Cl}, which count the total numbers of atoms of these types. Inspired by the utility of McGowan's simple scheme [14] for estimating molecular volumes, we also included n_{bond}, which is the total number of pairs of directly bonded atoms in the conventional structural formula, with no account taken of bond multiplicity. For set B, a linear regression model based on just these six descriptors gives a correlation coefficient of $R^2=0.856$ and a standard error $\Delta=0.482$. Augmenting these descriptors with S_n and $\acute{S}_n$ ($n=-2$, 0, $+2$) leads to an improved model, with $R^2=0.928$ and $\Delta=0.356$, and including also the total calculated energy E leads to $R^2=0.965$ and $\Delta=0.251$. Attempts to include also the expectation value $\langle p^{-2}\rangle$ and/or the molecular weight and/or the total number of (valence) electrons N were unsuccessful, on account of linear dependence

[1] The corresponding n_X-based model for the log P values of this set of 58 molecules is less good ($R^2=0.846$ and $\Delta=0.580$).

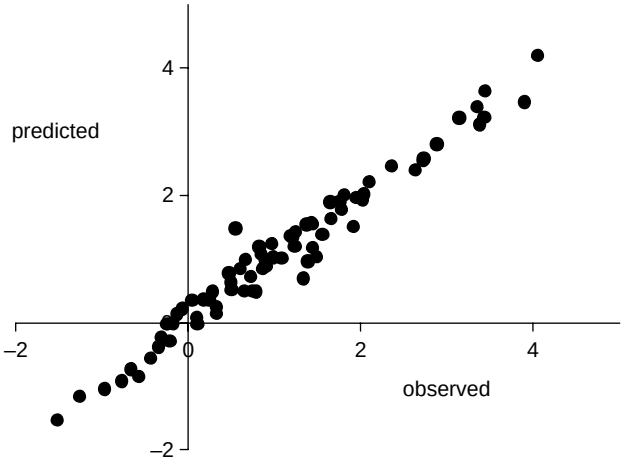

Fig. 3. Predicted and observed values of log P for 76 molecules.

between the various descriptors. A small further improvement could, however, be achieved by including also ln N. There was then very little change to the quality of the overall model if we excluded S_0.

A comparison of experimental and predicted values of log P from our preferred linear model is shown in Fig. 3 for the 76 molecules of set B. This regression model, characterized by $R^2 = 0.964$ and $\Delta = 0.256$, includes 13 descriptors: n_C, n_H, n_O, n_N, n_{Cl}, n_{bond}, E, ln N, $\acute{S}_0$, S_{-2}, $\acute{S}_{-2}$, S_{+2}, and $\acute{S}_{+2}$. As an internal test of the quality of our 13-descriptor model, we checked using the regression coefficients for set A that we could indeed make useful predictions for all of the 34 molecules that were added on going to set B.

It is likely that the incorporation of powers of our 13 descriptors could lead to even smaller errors. However, we have actively chosen not to pursue such a route because of our concerns that we might become involved in overfitting the data. Something of an outlier in many of the models that we investigated is hydroquinone. Excluding that molecule does leads to more flattering values of R^2 and Δ, but there is of course no justification for such an a posteriori reselection of the data set.

As mentioned earlier, we were hesitant to include any amino acids in our test sets. Nonetheless, it seemed worthwhile to investigate briefly the utility of an index $I_{zwitter}$ for the anticipated number of 'zwitterionic units' in the amino acids. As such, we augmented set B with the four amino acids (glycine, alanine, α-aminobutyric, α-aminovaleric) considered by Amat et al. [11], and we expanded our 13-descriptor model to include also $I_{zwitter}$. For this expanded set, consisting of 80 molecules, the 14-descriptor model gives $R^2 = 0.976$ and $\Delta = 0.250$. It could be interesting in further work to consider more complicated amino acid examples, with $I_{zwitter} > 1$. Direct comparison of R^2 and Δ values with the work of Bodor et al. [13] would be inappropriate, on account of the different data sets, but it is nonetheless clear that the models described here are performing rather well.

4. Conclusions

In this paper, we have extended previous work on quantitative structure–activity and structure–property relationships using molecular descriptors based on quantities determined from the momentum-space (p-space) electron density. In particular, we introduced new molecular descriptors that are related to the p-space information entropy. We also examined the use of a simple molecular shape descriptor, X. For the LC_{50} toxicity data of a series of saturated alcohols, the simple shape descriptor was found to be remarkably successful, and a correlation was observed between X and the entropy-like p-space descriptors. We developed a promising 13-descriptor regression model for the log P values of a set of 76 chemically diverse molecules, combining simple classical parameters (numbers of atoms and the number of bonds) with p-space quantities. Similar approaches are likely to prove useful for more complex QSAR/QSPR problems.

References

[1] E.F. McCoy, M.J. Sykes, Chem. Phys. Lett. 313 (1999) 707; E.F. McCoy, M.J. Sykes, J. Chem. Inf. Comput. Sci. 43 (2003) 545.
[2] J.H. Al-Fahemi, D.L. Cooper, N.L. Allan, submitted for publication.
[3] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, J. Am. Chem. Soc. 107 (1985) 3902.
[4] J.J.P. Stewart, J. Comput. Aided Mol. Des. 4 (1990) 1 and references therein.
[5] N.L. Allan, D.L. Cooper, Top. Curr. Chem. 173 (1995) 85; D.L. Cooper, N.L. Allan, in: R. Carbó (Ed.), Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches, Kluwer Academic Publishers, Netherlands, 1995; P.T. Measures, K.A. Mort, N.L. Allan, D.L. Cooper, J. Comput. Aided Mol. Des. 9 (1995) 331; P.T. Measures, N.L. Allan, D.L. Cooper, Adv. Mol. Similarity 1 (1996) 61; N.L. Allan, D.L. Cooper, J. Math. Chem. 23 (1998) 51; D.L. Cooper, N.L. Allan, P.B. Karadakov, in: R. Carbó-Dorca (Ed.), The Fundamentals of Molecular Similarity, Kluwer/Plenum Press, New York, 2001; Ll. Amat, R. Carbó-Dorca, D.L. Cooper, N.L. Allan, Chem. Phys. Lett. 367 (2003) 207.
[6] S.R. Gadre, Phys. Rev. A 30 (1984) 620; S.R. Gadre, R.D. Bendale, S.P. Gejji, Chem. Phys. Lett. 117 (1985) 138; S.R. Gadre, R.D. Bendale, Int. J. Quant. Chem. 28 (1985) 311; S.R. Gadre, S.B. Sears, S.J. Chakravorty, R.D. Bendale, Phys. Rev. A 32 (1985) 2602.
[7] M. Hô, B.J. Clark, V.H. Smith Jr., D.F. Weaver, C. Gatti, R.P. Sagar, R.O. Esquivel, J. Chem. Phys. 112 (2000) 7572; R.P. Sagar, J.C. Ramírez, R.O. Esquivel, M. Hô, V.H. Smith Jr., Phys. Rev. A 63 (2001) 022509.
[8] N.L. Allan, D.L. Cooper, J. Chem. Phys. 84 (1986) 5594.
[9] P.G. Mezey, Shape in Chemistry: An Introduction to Molecular Shape and Topology, VCH Publishers, New York, 1993.
[10] G.P. Romanelli, L.F.R. Cafferata, E.A. Castro, Theochem 504 (2000) 261.
[11] L. Amat, R. Carbó-Dorca, R. Ponec, J. Comput. Chem. 19 (1998) 1575.
[12] X. Gironés, L. Amat, D. Robert, R. Carbó-Dorca, J. Comput. Aided Mol. Des. 14 (2000) 477.
[13] N. Bodor, Z. Gabanyi, C.-K. Wong, J. Am. Chem. Soc. 111 (1989) 3783.
[14] A. Mellors, J.C. McGowan, Biochem. Pharm. 34 (1985) 2413.