Available online at www.sciencedirect.com
SCIENCE DIRECT
Chemical Physics Letters 416 (2005) 376–380
CHEMICAL PHYSICS LETTERS
www.elsevier.com/locate/cplett

The quantitative use of momentum-space descriptors

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, Avon BS8 1TS, UK

Received 2 September 2005; in final form 13 September 2005
Available online 19 October 2005

Abstract

We explore the possible use of various momentum-space quantities as molecular descriptors in QSAR and QSPR studies. It is found that three- or four-descriptor models, that include molecular weight, provide useful correlations for a range of property and activity data, such as gas-chromatography retention times, gas-hexadecane partition coefficients and tadpole narcosis concentrations. The underlying reason for this success suggests the utility of hybrid models that take account also of the atomic character, albeit corrected for the number of bonds.

1. Introduction

On the whole, most chemists feel much more 'comfortable' when visualizing molecular electronic structure in terms of position space (r-space) electron densities rather than via the complementary momentum space (p-space) representation. Nonetheless, studies of p-space densities do go back to the early days of molecular quantum mechanics, and there has been extensive literature over the intervening decades, motivated in part also by various experimental techniques [1,2] for which p-space interpretations are particularly convenient and/or appropriate. Furthermore, similarity and dissimilarity indices based on quantitative comparisons of p-space electron densities $\rho(\mathbf{p})$ have proved useful in a range of applications, mostly related to the rationalization and prediction of drug activity. This has led ourselves, and also others, to explore the possible use of appropriate p-space quantities as molecular descriptors in QSAR and QSPR studies. This short contribution reports some of our initial findings in this area. For the purposes of comparison, we have used exactly the same experimental data sets as McCoy and Sykes [3–5] in their exploration of possible p-space strategies. In spite of some initial surprises, the prognosis does indeed appear to be favourable.

2. Momentum-space descriptors

Our procedures for generating p-space densities from conventional electronic structure calculations and for calculating appropriate p-space integrals have been described many times, and so we have chosen here to concentrate only on recent developments. A useful introduction, with references to much of the previous literature, is provided in [6–11] Our primary concern in the present work is with the possible direct use of moments of momentum

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

as descriptors in quantitative structure activity/property relations. We consider values of n from -2 to $+2$, where decreasing values of n place increasing emphasis on the slower-moving valence electrons. We note that values of $\langle p^n \rangle$ have already proved useful for the qualitative classification of reaction pathways [12] and for the quantitative prediction of Hammett σ constants [13].

* Corresponding authors. Fax: +44 151 794 3588 (D.L. Cooper), +44 117 925 0612 (N.L. Allan).
E-mail addresses: dlc@liv.ac.uk (D.L. Cooper), n.l.allan@bris.ac.uk (N.L. Allan).

0009-2614/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2005.09.055

An alternative p-space strategy, pursued by McCoy and Sykes [3–5], involves the angular integration

$$D_2(p) = \int_0^{\pi} \int_0^{2\pi} p^2 \sin\theta \, \rho(\mathbf{p}) \, \mathrm{d}\phi \, \mathrm{d}\theta \tag{2}$$

and then fitting of the low-momentum portion of $D_2(p)$ to a quartic in p. The resulting expansion coefficients were used as parameters in various multiple regression models, but incorporating also more conventional descriptors. Useful correlations, with good R^2 values and relatively low standard errors, were established for various molecular properties.

Instead of $D_2(p)$, we examined the radial integral

$$I_n(p_{\max}) = \int_0^{p_{\max}} p^n \rho(\mathbf{p}) \, \mathrm{d}\mathbf{p} \tag{3}$$

and found that the dependence of $I_n(p_{\max})$ on $p_{\max}$, for a given value of n, has much the same shape for a wide range of molecules. Indeed, simple linear rescaling of the two axes practically superimposes the curves for a given value of n, at least up to 60% or so of the limiting value of I_n, i.e. $\langle p^n \rangle$ in Eq. (1). This suggests that there is somewhat less information available than one might suppose from multi-parameter fitting of the low-p region of $D_2(p)$. As such, it does seems particularly worthwhile to investigate whether one could achieve useful correlations with small numbers of p-space expectation values, thereby avoiding the fitting (and, perhaps, over-fitting) of somewhat arbitrary segments of $D_2(p)$.

For this exploratory study, we considered many of the same properties as McCoy and Sykes [3–5], adopting the same experimental data for gas chromatography retention times, gas-hexadecane partition coefficients, liquid densities, tadpole narcosis concentrations, fathead minnow toxicities, electric polarizabilities, and diamagnetic susceptibilities.

The generation of momentum densities from conventional r-space wavefunctions is, in general, entirely straightforward, as is the subsequent integration so as to obtain $\langle p^n \rangle$. However, given that an eventual aim is to use appropriate p-space descriptors for the relatively rapid processing of fairly large sets of molecules, it is important to be able to use fairly inexpensive electronic structure calculations. With this in mind, we chose to work here with the AM1 procedure in MOPAC [14,15], and we also used geometries that were optimized only at this low level of theory. Realistically, one might in due course need to be able to use something even cheaper for large-scale routine studies of (typically) many thousands of molecules. On the other hand, such a scheme is very unlikely indeed to distinguish properly between close-lying conformers.

3. Results and discussion

3.1. Gas-chromatography retention times

We considered the same set of 23 aliphatic and aromatic molecules as McCoy and Sykes [4,5]. We had expected to use a number of values of $\langle p^n \rangle$, corresponding to various (non-integer) values of n from -2 to $+2$, but we soon discovered rather good correlations between expectation values calculated with values of n that are reasonably close. As such, it makes sense to restrict n to a very small number of values. After various tests, also of correlations for other properties, we settled on the integer values -2, 0 and $+2$. Of course, $\langle p^0 \rangle$ is simply equal to the number of electrons and $\langle p^2 \rangle$ is twice the kinetic energy. Bearing in mind the virial theorem, which relates the kinetic energy to the total energy, it seemed appropriate to select as descriptors our computed value of $\langle p^{-2} \rangle$, as well as the number of electrons N and the total energy E in the AM1 calculations. To this trio, we added the relative molecular mass or 'molecular weight' M. All of our models in this Letter are linear in the various descriptors, as was also the case for the work of Sykes and McCoy [3–5].

We used linear regression facilities in a PC version of SPSS, and subsequently obtained additional estimates of the quality of the various models by using Microsoft Excel to compare the 'predicted' and experimental quantities. In assessing the relative merits of our different models, we monitored the standard SPSS analysis of variance F statistic (which should be large) as well as the significance value of the F statistic (which should be small). A conventional criterion for the significance value of the F statistic is that it should be less than 0.05. We found in the present work that this quantity was typically in the range 10^{-10}–10^{-30}, indicating that the success of a given model in reproducing the variation in the data is rather unlikely to be due to chance. An alternative method for evaluating the statistical importance of correlations in QSAR studies, derived using geometric arguments, was subsequently shown to be equivalent to the classical F significance method [16,17].

The correlation between observed and predicted gas-chromatography retention times was found to be fairly good, as can be seen from Fig. 1a: we find $R^2 = 0.97$ and a standard error of $\Delta = 0.49$. There is certainly still room for improvement, but the model is already of a quality that is comparable to those described in various previous studies. For example, McCoy and Sykes [4,5] found $R^2 = 0.97$ and $\Delta = 0.51$ using five descriptors for this data set. We assessed the importance of our four descriptors ($\langle p^{-2} \rangle$, N, E and M) simply by noting the change to R^2 and Δ on leaving out one or more of them. The worst of the three-descriptor models was found to be the one that excludes molecular weight: it gives $R^2 = 0.86$ and $\Delta = 1.06$. The importance of including M turned out to be a general finding for almost all of the properties we considered. For the present case, the least important of the four descriptors is the total energy: its exclusion from the model results in $R^2 = 0.95$ and $\Delta = 0.61$. As measured by the significance value of the F statistic, this three-descriptor model is statistically less significant than the four-descriptor model that also includes the total energy.

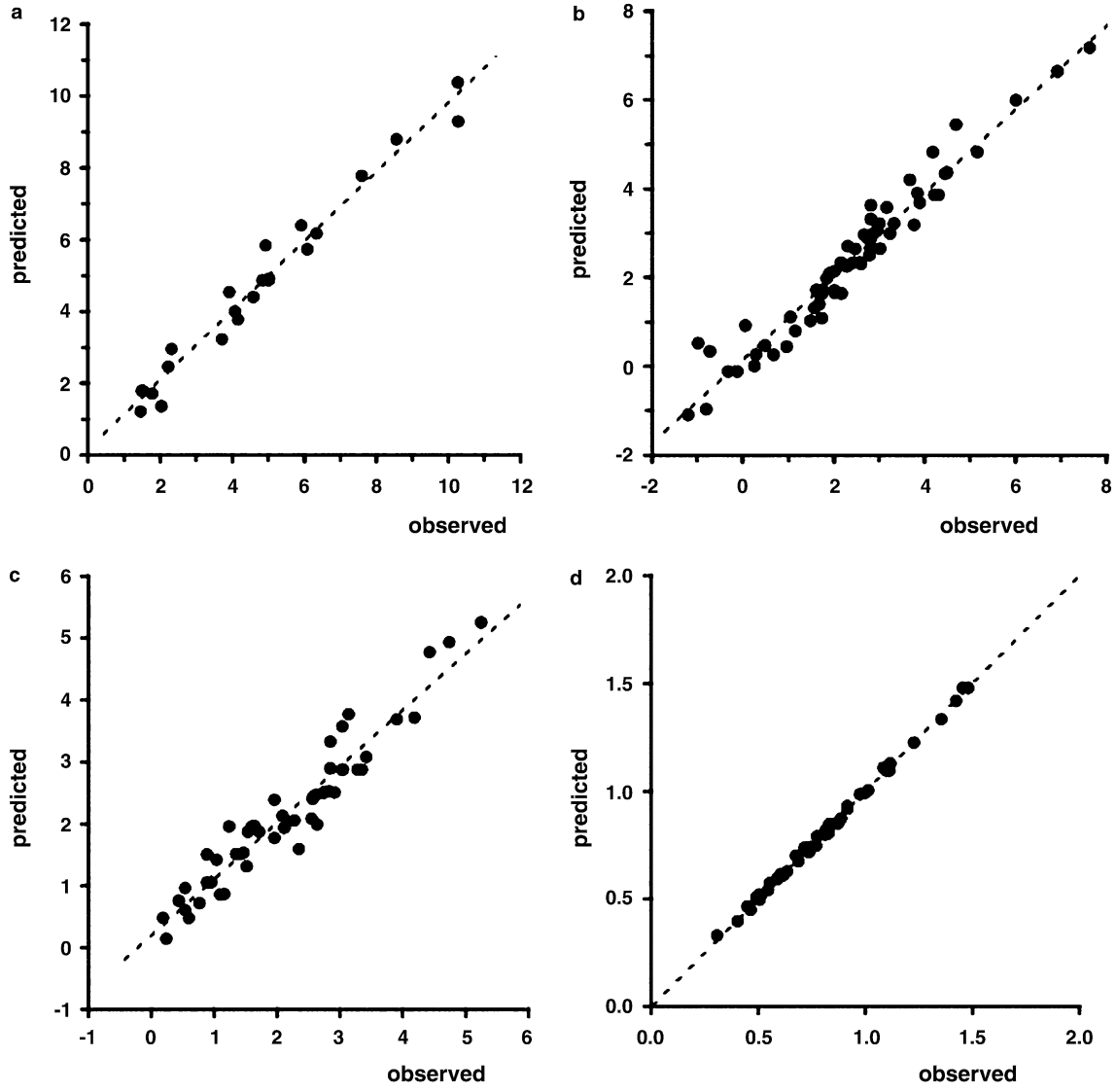

Fig. 1. Correlation between predicted and observed data: (a) gas-chromatography retention times (in minutes); (b) gas-hexadecane partition coefficients, expressed as $\log(L^{16})$; (c) tadpole narcosis concentrations, expressed as $\log(1/C_{\text{nar}})$; (d) McGowan's volume (arbitrary units). Experimental data are the same as in [4,5].

3.2. Gas-hexadecane partition coefficients

Gas-hexadecane partition coefficients, expressed as $\log(L^{16})$ are of course widely used in structure activity studies. For the same set of 63 molecules as considered by McCoy and Sykes [4,5], we find that a model based on just four descriptors, namely $\langle p^{-2}\rangle$, N, E and M, gives $R^2 = 0.95$ and $\Delta = 0.43$ (see Fig. 1b), with M being the most important of the four descriptors. McCoy and Sykes [4,5] do not present statistics for this full set of molecules because four fluorine-containing systems were obvious outliers in their work. Using four descriptors for the reduced set of 59 molecules, they found $R^2 = 0.95$ and $\Delta = 0.39$.

3.3. Tadpole narcosis concentrations

Tadpole narcosis concentrations, expressed as $\log(1/C_{\text{nar}})$, were considered for the same 52 molecules as considered by McCoy and Sykes [4,5]. A three-descriptor model, using $\langle p^{-2}\rangle$, E and M, gives a useful correlation (see Fig. 1c) with $R^2 = 0.92$ and $\Delta = 0.35$. As before, M was found to be the most important of the descriptors: its exclusion from the model results in $R^2 = 0.62$ and $\Delta = 0.74$. For these data, very little was gained by using all four of our descriptors. With five descriptors, McCoy and Sykes [4,5] obtained $R^2 = 0.93$ and $\Delta = 0.33$.

With the possible exception of the fathead minnow data, which exhibited a little more scatter, we found analogous high quality three- or four-parameter models for all of the sets of experimental data have been considered by McCoy and Sykes [3–5]. This fuelled a growing uneasy feeling that both we and they were, perhaps, doing *too* well. This suspicion was confirmed, to a large extent, when we considered empirical molecular volumes.

3.4. McGowan volume

Given the obvious importance of the size of a molecular system to many of its properties, it comes as no surprise that molecular volumes have often proved useful in QSAR and QSPR studies. A particularly straightforward empirical scheme for estimating molecular volumes, typically for use in such studies, was proposed by Mellors and McGowan [18]. In essence, one simply adds together a set of fixed atomic values and then reduces the resulting total by a fixed correction for each bond. In this context, no distinction is made for the bond order or for the type of bonding. The McGowan volume of ethyne, for example, is simply the sum of the values for two H atoms and for two C atoms, corrected for a total of just three 'bonds'.

We examined the McGowan volumes of the same 54 molecules as considered by McCoy and Sykes [4,5]. Using only two descriptors, namely $\langle p^{-2}\rangle$ and M, the quality of the resulting model is remarkable (see Fig. 1d), with a value of R^2 that is practically unity. A similarly high value of R^2 was obtained by McCoy and Sykes [4,5], using four descriptors. By any objective measure, these results are simply too good, and they are suggestive of an alternative explanation for the apparent success (for a range of molecular property and activity data) of both our models and the scheme used by McCoy and Sykes [3–5].

It now appears that many of these various properties are much more 'atomic' than we might have supposed, except that we must take account of the number of bonds (in the sense that this term is used when calculating McGowan volumes). With this in mind we checked for all of the sets of molecular property and activity data that we could establish just as good regression models (as measured by R^2 and Δ) when using *only* the numbers of atoms of each type and the numbers of 'bonds'. The increased number of descriptors did tend to lead to some increase in the significance value of the F statistic, indicating that the new models are statistically less important, but all of these values do remain very small indeed. We also established that we could make reasonable predictions of our calculated values of $\langle p^{-2}\rangle$ and E when using these very simple descriptors. It seems almost inevitable that there must also be reasonable correlations with atomic quantities, corrected for number of 'bonds', for combinations of the quartic coefficients used by McCoy and Sykes [3–5].

4. Summary and conclusions

It is now clear that moments of momentum $\langle p^n\rangle$ $(-2 \leqslant n \leqslant +2)$, when augmented with the relative molecular mass or 'molecular weight' M, may be used to construct useful three- or four-descriptor models for a range of experimental molecular property and activity data. As well as the cases for which results have been presented here, we successfully used the same basic approach for electric polarizabilities, diamagnetic susceptibilities, liquid densities and fathead minnow toxicities, again using the same molecular data sets as McCoy and Sykes [3–5].

Our initial enthusiasm has been tempered by the realization that there appears to be a rather simple underlying explanation for the success of our models and of those of McCoy and Sykes [3–5]. In general terms, one can probably do just as well (as measured by R^2 and Δ) with a much simpler scheme that uses as descriptors only the numbers of atoms of each type and the number of 'bonds'. Of course, one needs to be careful in this context to ensure that there is a sufficient range of molecules, so as to avoid situations in which there are just a few rather similar molecules that contain a particular element. For all of the cases that we considered, there is a significant penalty on excluding the number of 'bonds', which is the *only* 'molecular' aspect of this rather trivial set of descriptors.

Many types of descriptors, with some of them being fairly complex, have been employed for the successful prediction of wide ranges of molecular property and activity data. Based on our results using values of $\langle p^n\rangle$, it is tempting to wonder whether the exclusion of molecular weight from some of those models has led to an unnecessary increase in the number of descriptors. The present study also provides a salutary lesson: one might do well to investigate *why* some of those models do so well. In particular, it could be fruitful to ascertain the reliability, or otherwise, of the much simpler atom-like scheme, based only on numbers of atoms and bonds.

We find that our best results (as measured by R^2 and Δ) are obtained from hybrid models in which the numbers of atoms and bonds are augmented with the values of $\langle p^{-2}\rangle$. Although all descriptors are used simultaneously in the multiple regression, there is a sense in which the anticipated role of $\langle p^{-2}\rangle$ is to help to describe the 'residue', i.e. that part of the data set that cannot be predicted only from the simple atom-like model. With this in mind, we have started to search for further p-space descriptors that might prove quantitatively useful for refining the prediction of molecular properties and activity data. Our initial results [19] for a generalized 'entropy-like' quantity appear to be particularly promising, and further results will be reported in due course. An ultimate aim is to identify p-space quantities that could be useful additional descriptors for the improved modelling of more challenging problems, such as the brain–blood partitioning of organic solutes.

References

[1] B.G. Williams (Ed.), Compton Scattering, McGraw-Hill, New York, USA, 1977.
[2] E. Weigold, I.E. McCarthy, Electron Momentum Spectroscopy, Kluwer/Plenum, New York, USA, 1999.
[3] E.F. McCoy, M.J. Sykes, Chem. Phys. Lett. 313 (1999) 707.
[4] M.J. Sykes, Estimation of molecular properties with momentum space wavefunctions, Ph.D. thesis, The Flinders University of South Australia, 2000.

[5] E.F. McCoy, M.J. Sykes, J. Chem. Inf. Comp. Sci. 43 (2003) 545.
[6] For example N.L. Allan, D.L. Cooper, Top. Curr. Chem. 173 (1995) 85.
[7] D.L. Cooper, N.L. Allan, Molecular similarity and momentum space, in: R. Carbó (Ed.), Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches, Kluwer, Netherlands, 1995, p. 31.
[8] P.T. Measures, K.A. Mort, N.L. Allan, D.L. Cooper, J. Comput. Aided Mol. Des. 9 (1995) 331.
[9] P.T. Measures, N.L. Allan, D.L. Cooper, Adv. Mol. Similarity 1 (1996) 61.
[10] N.L. Allan, D.L. Cooper, J. Math. Chem. 23 (1998) 51.
[11] D.L. Cooper, N.L. Allan, P.B. Karadakov, Confronting modern valence bond theory with momentum-space quantum similarity and with pair density analysis, in: R. Carbó-Dorca (Ed.), The Fundamentals of Molecular Similarity, Kluwer/Plenum, New York, USA, 2001, p. 169.
[12] Ll. Amat, R. Carbó-Dorca, D.L. Cooper, N.L. Allan, Chem. Phys. Lett. 367 (2003) 207.
[13] Ll. Amat, R. Carbó-Dorca, D.L. Cooper, N.L. Allan, R. Ponec, Mol. Phys. 101 (2003) 3159.
[14] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, J. Am. Chem. Soc. 107 (1985) 3902.
[15] J.P. Stewart, J. Comput. Aided Mol. Des. 4 (1990) 1, and references therein.
[16] J. Pecka, R. Ponec, J. Math. Chem. 27 (2000) 13.
[17] E. Besalú, J.V. de Julián-Ortiz, J. Math. Chem. 36 (2004) 361.
[18] A. Mellors, J.C. McGowan, Biochem. Pharmacol. 34 (1985) 2413.
[19] J.H. Al-Fahemi, D.L. Cooper, N.L. Allan, J. Mol. Struct. (Theochem) 727 (2005) 57.