# Quantum Molecular Similarity. 3. QTMS Descriptors

S. E. O'Brien and P. L. A. Popelier*

Department of Chemistry, U.M.I.S.T., 88 Sackville Street, Manchester M60 1QD, GB

Received November 15, 2000

Building on the ideas of a previous paper [part 1, *J. Phys. Chem. A* **1999**, *103*, 2883] we present a new molecular similarity method based on the topology of the electron density. This method is directly applicable to QSARs and is called quantum topological molecular similarity (QTMS). It has been tested for five sets of carboxylic systems including *para*- and *meta*-benzoic acid, *para*-phenylacetic acid, 4-X-bicyclo[2.2.2]-octane-1-carboxylic acids, and polysubstituted benzoic acids. In combination with the partial least squares (PLS) procedure QTMS is able to produce excellent and statistically valid regressions. It is shown that QTMS avoids certain challenges of traditional Carbó-like similarity indices. Finally, QTMS is able to suggest a molecular fragment that contains the active center or the part of the molecule that is responsible for the QSAR.

## INTRODUCTION

With the advent of more powerful computers and improved algorithms the field of quantitative structure−activity relationships (QSAR) increasingly benefits from quantum chemical calculations, both a*b initio* and semiempirical. In paper 1[1] of a series of papers we introduced a new molecular similarity method based on the topology of the electron density $\rho$.[2,3] This method shares the basic philosophy with the Carbó similarity index, which is that in principle $\rho$ contains all the information that can be known about a molecule. This statement is justified by the first Hohenberg−Kohn[4] theorem, which forms the basis of modern Density Functional Theory.[5] Furthermore $\rho$ is an entity that exists in real 3D space and that can be determined via X-ray diffraction or calculated using ab initio methods.

An appropriate name that covers the essential elements of our proposed method is quantum topological molecular similarity (QTMS). The QTMS method is “quantum” since it draws its data from computational schemes that explicitly incorporate the quantum nature of molecules.[6] It is “topological” since it uses the ideas of the theory of “atoms in molecules” (AIM) pioneered by Bader[3] to discretize the quantum information contained in a molecular system. In other words, in this paper QTMS uses the properties of special points in space called *critical points* (CP) to compactly represent a molecule. Furthermore, the keyword “topological” also refers to the possibility of a discrete molecular representation based on atomic properties, obtained after the partitioning of $\rho$ via its gradient vector field.[2,7] However, this extension is not discussed here but is the subject of ongoing work in our laboratory. Finally, the label “molecular similarity” refers to the general ability of QTMS to compare molecules and assess their similarity.

The basic idea of QTMS was proposed some time ago[8] and developed in paper 1. In the current paper we fully incorporate QTMS in a firm statistical framework currently employed in modern QSAR, making use of the partial least squares (PLS)[9] procedure. We show here that the early results on *para*-benzoic acids presented in paper 1 survive the current rigorous statistical treatment. Furthermore we show that its range of applications extends to other carboxylic acid systems, such as *para*- and *meta*-benzoic acid, *para*-phenylacetic acid, 4-X-bicyclo[2.2.2]octane-1-carboxylic acids, and polysubstituted benzoic acids. We have collected evidence[10] that the method presented here is also applicable to more complicated activities, but those results will be published elsewhere. This paper focuses on the details of the current stage of the QTMS method.

First we review the topology of the electron density. Subsequently we describe the steps taken in a QTMS analysis. Then we apply the QTMS analysis to five carboxylic acid systems. Finally we comment extensively on the strengths and future challenges of QTMS compared to molecular quantum similarity as developed by the Carbó group.

## THE TOPOLOGY OF THE ELECTRON DENSITY

The topology of the electron density is a full part of the theory of AIM,[2,3] which has been developed over the last three decades. In our group AIM has evolved into a research program to bridge the gap between modern ab initio wave functions and chemical insight. This theory is the most elaborately researched and documented way of partitioning a quantum system[11] (e.g. a molecule, van der Waals complex or crystal) into atomic constituents, and reviewing it here is beyond the scope of this article. AIM provides a consistent way of partitioning and hence localizing chemical information, irrespective of the particular mathematical representation of the electron density, including molecular orbitals, basis sets (plane waves, Gaussians, Slaters), crystallographic structure factors, or grids.

For the purpose of this paper we only focus on the so-called *bond critical points* (BCP). These are points in real 3D space where the gradient of the electron distribution vanishes (or $\nabla\rho = \mathbf{0}$) and where the Hessian of $\rho$ (or $\nabla\nabla\rho$)

* Corresponding author phone: +44-161-200 4511; e-mail: pla@umist.ac.uk.

10.1021/ci0004661 CCC: $20.00 © 2001 American Chemical Society
Published on Web 04/21/2001

has two negative eigenvalues and one positive one. These points occur roughly between two bonded nuclei and are part of a more complex topology of $\rho$, which is not relevant for our current application.

The Hessian $\nabla\nabla\rho$ is a matrix describing all possible second derivatives of $\rho$ with respect to position coordinates *x, y,* and *z*. We introduce the Hessian eigenvalues $\lambda_i$ is because they express the local curvature of $\rho$ in a point *independent* of the choice of molecular coordinate system. By convention they are ordered as follows: $\lambda_1 < \lambda_2 < \lambda_3$. Consequently, at a BCP, $\lambda_1 < \lambda_2 < 0$ and $\lambda_3 > 0$. The latter positive curvature is associated with an eigenvector that is tangent to the *bond path*. The bond path is a curve in real space linking two bonded nuclei along which $\rho$ is a maximum with respect to any neighboring line. The Laplacian of $\rho$, or $\nabla^2\rho$, is a measure of how much $\rho$ is concentrated or depleted in a point.

Another quantity derived from the Hessian eigenvalues is the ellipticity at the BCP, denoted by $\epsilon_b$ or simply $\epsilon$. The ellipticity is defined as $(\lambda_1/\lambda_2) -1$ and is always positive because $\lambda_1 < \lambda_2 < 0$ at the BCP. Since $|\lambda_1| > |\lambda_2|$ the latter corresponds to the "soft" curvature. A contour diagram of $\rho$ in the plane of the eigenvectors corresponding to $\lambda_1$ and $\lambda_2$ shows a set of nested ellipses (or circles if $\lambda_1 = \lambda_2$). Clearly $\lambda_2$ corresponds to the major axis because in this direction less contour lines are crossed per unit length as a result of the soft curvature. Bonds can further be characterized by evaluating two types of kinetic energy densities, denoted by K(**r**) and G(**r**)², at the BCP. They are defined as $K(\mathbf{r}) = -{}^1/_4 N\int d\tau'[\psi^*\nabla^2\psi + \psi\nabla^2\psi^*]$ and $G(\mathbf{r}) = {}^1/_2 N\int d\tau'\nabla\psi^* \bullet \nabla\psi$, where $\int d\tau'$ denotes an integration over the spin coordinates of all *N* electrons except one.

The quantities introduced above can be packed together in a vector serving as a chemical descriptor for a bond. In other words, $\mathbf{P} = (\rho, \lambda_1, \lambda_2, \lambda_3, \nabla^2\rho, \epsilon, \mathrm{K}, \mathrm{G})_b$ is an eight-dimensional vector of property components evaluated at the BCP, indicated by the subscript *b*. Within congeneric classes of bonds, some of these components can be related to a chemical interpretation.[2] For example, the ellipticity measures the susceptibility of ring bonds to rupture and provides a quantitative generalization of the $\pi$ character of a bond. In this work the descriptor **P** spans an eight-dimensional hyperspace called BCP space. Of course the dimensionality of BCP space can be increased if more types of properties are evaluated at each BCP. The main idea behind BCP space is that it is an abstract space of a number of quantum mechanical properties evaluated at topologically special points in the molecules, i.e., BCPs. Currently the number of descriptors included in **P** is arbitrary, which probably warrants a systematic study to justify a recommended number. In Paper 1 for example a three-dimensional BCP space, consisting of the components $\rho_b$, $\nabla^2\rho_b$, and $\epsilon_b$, was successful in the construction of a *para*- and *meta*-benzoic acid QSAR. Furthermore, in future work redundancies from linear combinations such as $\nabla^2\rho = \lambda_1 + \lambda_2 + \lambda_3$ will be avoided, since this and other QTMS studies have shown that the "raw" variables (i.e. the components of **P**) do not feature in the interpretation of the QSAR. In the next section it will become clear that we are using principal component analysis (PCA) instead at the interpretation stage of the QSAR.

Paper 2[12] of the present series of papers focused on the stability of the property vector **P** with respect to the level of calculation. We will discuss the influence of the level of calculation upon the results of the PLS analysis. Paper 2 also investigated the dependence of **P** on the equilibrium bond length $R_e$. The existence of local linear relationships between some components of **P** with $R_e$ was confirmed provided the bonds vary little in their chemical surroundings. It was concluded that BCP properties cannot be trivially recovered or even predicted via $R_e$ alone. As a result $R_e$ is added as a ninth component to **P**. The latter extension of **P** illustrates the flexibility of the quantum topological description. In principle it is possible to transcend BCP space and add atomic properties as descriptors. These atomic properties are also defined[2,13] within the topological framework of the AIM theory and encompass atomic charge, dipole moment, volume, energy, etc. Furthermore, critical points from other scalar fields such as $\nabla^2\rho$ can be included in **P**. As a result QTMS extracts in a consistent and unified way discrete information from molecular wave functions.

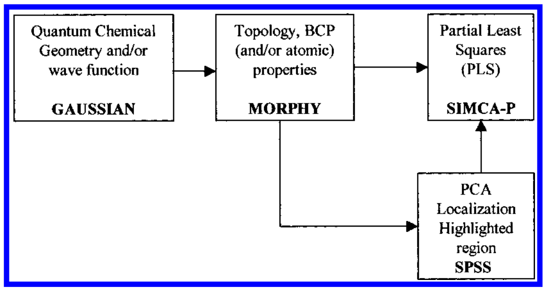
Quantum Chemical Geometry and/or wave function
GAUSSIAN
Topology, BCP (and/or atomic) properties
MORPHY
Partial Least Squares (PLS)
SIMCA-P
PCA Localization Highlighted region
SPSS

**Figure 1.** Chart representing the main computational modules involved in a QTMS analysis. The bold text represents the names of the programs used in this work.

## THE QTMS ANALYSIS

Figure 1 summarizes the main computational modules involved in a QTMS analysis together with the corresponding names of the computer programs we have used in this work. The restrictions upon the choice of series of molecules are the usual ones found in QSAR: one must have a common set of experimentally obtained data (the activities) against which the structure descriptors are regressed, the molecules should all produce the measured response in a mechanistically similar way, and the data set should be large enough to avoid unsafe correlations caused by chance factors.

In the first step a guess for the geometry of each molecule in the QSAR set chosen for study is offered to the ab initio program GAUSSIAN98.[14] A theoretical method (e.g. HF, MP2, B3LYP) and basis set has to be chosen depending on the computer resources available. As usual in any ab initio study the level of calculation needs to be varied in order to obtain reliable results. The choice of calculation levels is determined by the computer resource at hand and is perhaps guided by the trends discovered in this work. In addition, for our current methodology to be practically applicable, the series must share a common molecular skeleton. This is not a fundamental requirement of the method but is due to the computationally unfeasibly large number of variables that would be generated if no bond-to-bond matching was decided a priori. In practical terms, a common molecular framework is used in much conventional Hansch[15] and QSAR analysis.

In the second step the electronic wave function is passed on to (a local version of) the program MORPHY98,[16] which locates the BCPs using an automatic and robust algorithm.[17]

**Table 1.** Substituents for *Para*-Substituted Benzoic Acids with Associated $\sigma_p$ Values[a]

| X | $NMe_2$ | NHMe | $NH_2$ | OMe | Me | $CH_2Me$ | $CHCH_2$ | H | F | SH | Br | Cl | $CF_3$ | CN | $NO_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma_p$ | −0.83 | −0.7 | −0.66 | −0.27 | −0.17 | −0.15 | −0.04 | 0 | 0.06 | 0.15 | 0.23 | 0.23 | 0.54 | 0.66 | 0.78 |

[a] Me = methyl.

At this stage the property vectors **P** for each BCP are constructed yielding a discrete quantum fingerprint for each molecule in the QSAR set.

The third stage is the PLS analysis, which we have rigorously applied using recommended values[18] of the SIMCA-P program.[19] Throughout this paper we report four well-known statistics, denoted in this paper by $r^2$, $q^2$, $r^2$(int), and $q^2$(int). These statistics make us decide upon the quality of the regression, more precisely, on its internal predictive power (cross-validation) as well as its validity against high correlation by pure chance (randomization). We discuss these statistics in more detail below. It is important to realize that at this stage PLS operates on the "raw" variables, i.e., the BCP properties themselves.

The fourth and final stage involves a data reorganization to reduce the number of variables. Rather than discarding certain data, we compress them via a principle component analysis (PCA).[20] Since we evaluate BCP properties at specific points in space it is natural to apply PCA on the components of a property vector **P**. By extracting the PCs associated with each group of variables, we summarize the electronic properties of that particular bond in fewer variables. By reducing the available data without making choices about which pieces of information to keep and which to reject,[21] we are able to use the PCs to carry out our QSAR analysis. Only the PCs which have eigenvalues greater than one will be extracted, using the program SPSS.[22] The percentage variance explained (or the "total information kept") by each PC at every BCP is typically around 90% or higher. The purpose of the PCA is to interpret the mode of action of the QSAR. Here the topology helps us to localize the chemical information that is important to explain the measured response. It is important to realize that PLS is carried out again, this time on the extracted PCs rather than on the "raw" variables (see Figure 1). In summary, the PLS analysis is used twice, once on the actual BCP properties, and once on the PCs. The former PLS analysis yields the regression statistics ("quality and validity" of fit), whereas the second PLS analysis focuses on the interpretation of the regression.

There is a temptation to reduce the number of descriptors, but this removes information and can also lead to spurious models. Hierarchical PLS models have been suggested as a way to combat this problem.[9,21,23] Instead of using the raw variables it is suggested that descriptors can be grouped according to type, region, or function. A "super-variable", which incorporates the main factors of each of these groups, can then be used to develop a relationship. The model can be divided into two blocks in which the upper level models the activity data and the lower level models the "super-variables". There are certain programs and algorithms[21,24,25] that have been used for hierarchical PLS, but they all differ and are fairly application specific.

The first statistic $r^2$ is the commonly used correlation coefficient or "goodness of fit", while the second, $q^2$, is called the cross-validated $r^2$.[20] This statistic is dependent on the PRESS[26] score, defined within the "leave one out" cross-validation technique. The more similar $r^2$ and $q^2$, the more valid the correlation. The third and fourth statistics are formulated in the context of the data permutation or randomization. This test estimates the probability that a good fit is obtained if the Y variables ("activities") are randomly reordered. The program SIMCA-P produces a graph following each randomization and subsequent PLS analysis. In this graph the $r^2$ and $q^2$ values are plotted against the absolute value of the correlation coefficient between the original response variable and its permutation. A line is fitted through the $r^2$ values and another through the $q^2$ values, and the intercepts of these two lines are scrutinized. It is recommended[9] that for valid models one must find intercepts of $r^2 < 0.4$ and $q^2 < 0.05$. In other words, if $r^2(\text{int}) \geq 0.4$ or $q^2(\text{int}) \geq 0.05$, the model should be discarded on the suspicion that it is due to pure chance (where "int" refers to "intercept").

The PCs arising from PLS are known as latent variables (LV). Their construction fulfils the criteria stipulated for PCA but adds an additional constraint, namely that each LV must be maximally correlated with the dependent variable "Y". The program SIMCA-P prescribes a criterion for the significance of an LV, i.e., if $q^2 < 0.097$ the LV is not significant and no more LVs are computed. The PLS regression is then deemed complete. We also use the variable importance in the projection (VIP) values. The VIP gives the relative importance of each independent variable ("X") in the regression. Hence factors that contribute substantially to the fit have high VIP scores.

It is a working hypothesis of QTMS that the "active center" of a molecule constitutes the BCPs associated with the highest VIPs. Both the test cases looked at in this work and more complicated QSARs[10] suggest that the highest VIPs may contain other parts of the molecule that cannot be readily associated with the mode of activity. However, based on our experience so far we believe that QTMS has sufficient suggestive power to highlight the part of the molecule responsible for the measured response. In some cases this highlighting is rather sharp, in some more diffuse with some contamination. It was a gratifying to notice[27] that the "active center" changes with different responses. However, at the moment we have no rigorous criterion to isolate the "active center" from the VIP plot.

## *PARA*-BENZOIC ACID

Having explained the QTMS methodology we are now in a position to apply it to a well-studied QSAR, namely Hammett's[28] $\sigma_p$ values for *para*-substituted benzoic acids. We illustrate the QTMS method in more detail than the subsequent systems. The currently preferred $\sigma_p$ values[29] are listed in Table 1. The use of this set of molecules is advantageous for outlining the method due to its structural simplicity, the extensive work previously carried out on it,[15,30]

**Table 2.** Notation for the Different Levels of Calculation[a]

| label | level of calculation |
|---|---|
| A | AM1//AM1 |
| B | HF/3-21G(d)//HF/3-21G(d) |
| C | HF/6-31G(d)//HF/6-31G(d) |
| D | B3LYP/6-311+G(2d,p)//HF/6-31G(d) |
| E | B3LYP/6-311+G(2d,p)//B3LYP/6-311+G(2d,p) |

[a] We follow the standard notation of Gaussian basis sets.[67] The level of calculation at the right-hand side of the double slash is that used during geometry optimization and the level at the left side is the one used for the wave function generation. The acronym HF stands for the Hartree−Fock computational scheme and B3LYP[68] is a popular standard hybrid DFT method.

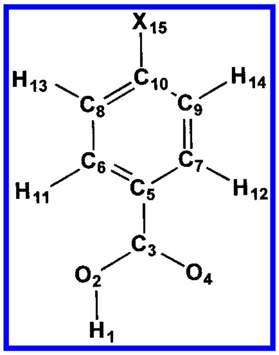

**Figure 2.** Labeling scheme of the common molecular skeleton of the *para*-substituted benzoic acids. **X** represents the substituents given in Table 1.

**Table 3.** VIPs for the 24 PCs Involved in the PLS Analysis of *p*-Benzoic Acids versus the Hammett $\sigma_p$ Constant[a]

| component | VIP | component | VIP |
|---|---|---|---|
| PC1_2a | 1.440 | PC8_10 | 0.950 |
| PC2_3 | 1.438 | PC9_10 | 0.880 |
| PC3_4 | 1.426 | PC5_6b | 0.781 |
| PC3_5 | 1.394 | PC10_15a | 0.705 |
| PC7_12 | 1.323 | PC9_14b | 0.655 |
| PC6_11 | 1.303 | PC7_9 | 0.582 |
| PC8_13 | 1.285 | PC8_13b | 0.537 |
| PC9_14 | 1.252 | PC6_8b | 0.527 |
| PC5_7b | 1.216 | PC5_7 | 0.414 |
| PC6_8 | 1.068 | PC10_15c | 0.387 |
| PC7_9b | 1.022 | PC10_15b | 0.261 |
| PC5_6 | 1.019 | PC1_2b | 0.085 |

[a] See Figure 3.

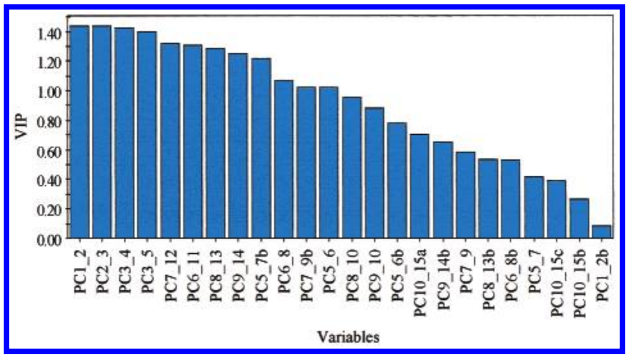

**Figure 3.** VIP plot for PLS analysis of p-benzoic acids versus Hammett sigma constants, using 24 PCs as descriptor variables (see Table 3).

and hence the degree to which the reaction is understood. The specific choice of substituents was designed to give an equal spread of positive and negative $\sigma_p$ values.

As the molecules are generally rigid the added complication of conformation does not arise. Where any geometrical flexibility does occur, the lowest energy conformation has been used. All the molecules were computed as having $C_s$ symmetry with the exception of the $NR_2$ substituted benzoic acids in which the amino groups were around 20 degrees out of the plane. The common molecular skeleton and atomic numbering scheme are given in Figure 2.

The systems looked at in this work were investigated at each of the five levels of calculation shown in Table 2. We use the abbreviations (A,B,C,D,E) throughout the following text and tables. Although BCP properties vary significantly from level to level, it has been shown that relative trends in the properties are preserved between sets.[12,31] The degree to which the resultant QSAR is level dependent will be investigated. However for this initial investigation we utilize only the B3LYP/6-311+G(2d,p)//B3LYP/6-311+G(2d,p) level of calculation (E).

The PLS analysis was applied on the 15 × 9 = 135 raw variables, i.e., the 9 original BCP properties as found in the 15 **P** descriptor vectors, one vector for each BCP. From this analysis emerged a model with two LVs and (cumulative) $r^2$ and $q^2$ values of 0.98 and 0.96, respectively. Although this model is valid it does not lend itself very well to interpretation, which is why we resort to PCA after PLS.

The VIPs for all 24 PCs involved in the PLS analysis are given in Table 3 and visually represented in Figure 3. The three most important variables are PC 1_2a, PC 2_3, and PC 3_4, are all very similar VIP scores. Referring to the molecular skeleton in Figure 3 these variables correspond to the area of $\rho$ comprising the carboxylic group. The next PC (PC 3_5) comes from the C−C bond attached to the ring. Unsurprisingly, this too has relatively high importance as it is directly affected by the carbonyl carbon. Although the BCPs are localized in space, they reflect the density in their near vicinity. These four PCs can be associated with the carboxylic group's bonds and so can link directly to the chemist's intuitive ideas of molecular reactivity. We label the molecular area consisting of the PC variables with the highest VIP score as the *highlighted region*, which we surmise to contain the active center of the molecule. We note that the changes in the properties of $\rho$ in this region reflect the observed variations in activity. This information fits in with knowledge acquired about benzoic acid acidities from traditional methods.

As the variables' VIP scores decrease we observe that the C−H BCPs are picked out after the carboxylic group and only then the aromatic ring. Whether this qualitatively reflects the mechanisms associated with benzoic acid acidity is unclear. As mentioned before there is no obvious point at which the highlighted region can be deemed to end. The graph in Figure 3 clearly shows that the decrease in VIP scores is gradual and does not suggest a natural cutoff. However we can identify those areas of the electron density that have more influence in the model than others.

We now examine how much these results depend on the level of calculation and report the results in Tables 4 and 5. Note that the results obtained at level A (Table 2) only use $R_e$ as a property, because it is beyond the scope of this work to unambiguously extract total molecular electron densities from AM1. Tables 4 and 5 show that in terms of the ability to reproduce and predict activities all the levels of calculation

**Table 4.** Correlation Coefficients for a PLS Analysis of $\sigma_p$ Regressed against the BCP Properties of *Para*-Substituted Benzoic Acids

| level | $r^2$ | $q^2$ | $r^2$ (int) | $q^2$ (int) | #LVs[b] |
|---|---|---|---|---|---|
| A | 0.96 | 0.92 | OK[a] | OK | 3 |
| B | 0.96 | 0.95 | OK | OK | 1 |
| C | 0.96 | 0.95 | OK | OK | 1 |
| D | 0.96 | 0.95 | OK | OK | 1 |
| E | 0.98 | 0.95 | OK | OK | 2 |

[a] If the randomization test (data permutation) is passed then the label "OK" is given, otherwise the numerical values of $r^2$(int) and/or $q^2$(int) would be listed. [b] Number of latent variables (LVs) in regression.

**Table 5.** Correlation Coefficients for a PLS Analysis of $\sigma_p$ Regressed against the PCs of *Para*-Substituted Benzoic Acids

| level | $r^2$ | $q^2$ | important PCs |
|---|---|---|---|
| A | 0.96 | 0.92 | 3−4, 2−3, 3−5, 1−2 |
| B | 0.97 | 0.97 | 2−3, 3−4, 1−2, 3−5a |
| C | 0.99 | 0.97 | 1−2, 2−3, 3−4, 3−5a |
| D | 0.97 | 0.96 | 2−3, 3−4, 1−2, 3−5 |
| E | 0.96 | 0.96 | 1−2a, 2−3, 3−4, 3−5 |

[a] The four most important PCs are shown.

produce very similar results. In Table 5 we do not report the validation statistics explicitly because the $r^2$(int) and $q^2$(int) values are expected to be lower confirming better validation in view of the smaller number of X variables. We see that in both tables all the correlations are very high and the randomization validated the results. Encouragingly the results found at the semiempirical AM1 level (achieved in a fraction of the time) are extremely good but required three latent variables.

There are some differences in the ordering of the PCs. Although the four most important PCs were the same at all five levels each set of calculations produced conflicting sequences of VIPs. Qualitatively there is little difference between the basis sets but there is quantitatively. We can certainly see that the order of importance is definitely basis set dependent (B and C differ) and also dependent on the level of calculation. One must be wary of overinterpreting small differences in VIPs. However at all levels the highlighted region is similar and corresponds to the region of the molecule responsible for the observed activity.

## *PARA*-PHENYLACETIC ACID

We now focus on a QSAR, in which resonance interactions have been eliminated. For that purpose the $\sigma^0$ constant has been evaluated from the acidities of phenylacetic acids.[32] The $CH_2$ group provides insulation of the carboxyl group from any resonance interaction. The labeling scheme for the QTMS analysis of the set of *para*-substituted phenylacetic acids is given in Figure 4. The set of 15 substituents and their $\sigma_p{}^0$ values[29] are given in Table 6. As before all molecules were computed in their lowest energy conformations at the five levels of calculation (Table 2).

The results of the PLS analysis can be seen in Table 7. It is clear that all the levels reproduce $\sigma_p{}^0$ values ($r^2$) and provide good predictability ($q^2$). All the regressions are valid. It appears that the results tend to improve as the level of calculation increases, but this trend was not apparent in the benzoic acid set. To summarize the PLS analysis of the PCs

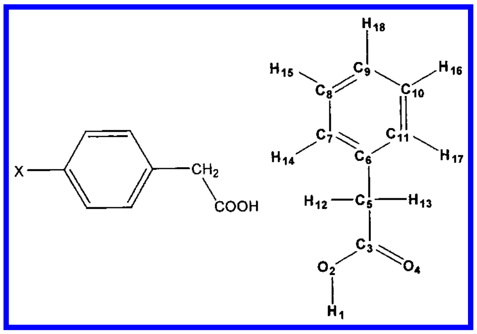

**Figure 4.** Labeling scheme of the common molecular skeleton of the *para*-substituted phenyl acetic acids. **X** represents the substituents given in Table 6.

based on BCP space we use a color code to mark the bonds with the highest VIP scores, as shown in Figure 5. Numerical data can be found in ref 10, showing the relative importance of each variable.

In Figure 5 the bonds are color-coded according to their importance in the projection. The code is shown in the diagram; the red bond being the most relevant for explaining the activity, followed by the yellow one, etc. At all levels of calculation BCP 1−2 is the most important. According to the VIP scores this region of the molecule is more prominent in the regression than the others. Conventional chemistry also tells us that this bond is the one which is broken in the process of acidity.

The ordering of the next most important factors differs at each level of calculation. In general the carboxylic group is shown to have substantial significance in the regression but levels A and C do not highlight BCP 3−4. There are many zones which are picked out as being important in the PLS analysis that one would not expect to be relevant to the mechanism of acidity. In particular the importance of the C−H BCPs situated *ortho* to the substituents are not anticipated. This is an example of the "contamination" mentioned before.

## 4-SUBSTITUTED BICYCLO[2.2.2]OCTANE-1-CARBOXYLIC ACID

There appear to be two effects included within the Hammett $\sigma$. One is a resonance contribution and the other a field-inductive element. For saturated systems the resonance is negligible or nonexistent so $\sigma$ as defined in the resonating benzoic acid system is not applicable. A different parameter that describes through bond and through space effects is known as $\sigma^I$.

The system used to measure $\sigma^I$ is the 4-X-bicyclo[2.2.2]octane-1-carboxylic acids[33] (Figure 6). In these molecules the substituent is not able to have any resonance interaction with the carboxyl group. We model this series using 11 different substituents (Table 8). The choice of substituent was governed by the availability of $\sigma^I$ constants. As the system is difficult to synthesize, direct measurement of the field-inductive parameter is not very common. All values were selected from Hansch, Leo, and Hoekman.[29] Every molecule except the nitro-substituted acid possesses $C_s$ symmetry.

The results of the PLS analysis as carried out with all the BCP variables are shown in Table 9. The *ab initio* calculations reproduce $\sigma^I$ very well except for the semiempirical

**Table 6.** Substituents for *Para*-Substituted Phenyl Acetic Acids with Associated $\sigma_p^0$ Values[a]

| X | $NMe_2$ | NHMe | $NH_2$ | OMe | Me | $CH_2Me$ | $CHCH_2$ | H | F | SH | Br | Cl | $CF_3$ | CN | $NO_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^0_p$ | −0.48 | −0.43 | −0.36 | −0.11 | −0.12 | −0.13 | 0 | 0 | 0.21 | 0.07 | 0.30 | 0.28 | 0.54 | 0.68 | 0.80 |

[a] Me = methyl.

**Table 7.** Correlation Coefficients for PLS Analysis of $\sigma_p^0$ Regressed against the BCP Properties of *Para*-Substituted Phenylacetic Acids

| level | $r^2$ | $q^2$ | $r^2$ (int) | $q^2$ (int) | #LVs |
|---|---|---|---|---|---|
| A | 0.91 | 0.88 | OK | OK | 1 |
| B | 0.93 | 0.93 | OK | OK | 1 |
| C | 0.95 | 0.91 | OK | OK | 2 |
| D | 0.95 | 0.92 | OK | OK | 2 |
| E | 0.99 | 0.97 | OK | OK | 2 |

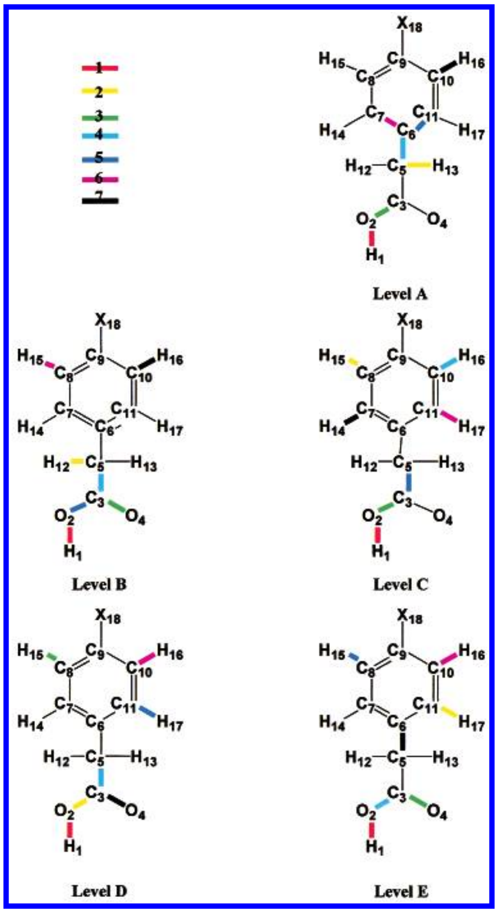

**Figure 5.** Color coding to denote the most important bonds in the PLS regression of $\sigma_p^0$ against the PCs drawn from the BCP properties of *para*-substituted phenylacetic acids at each level of calculation (A,B,C,D,E, Table 2).

method (level A). The values of the $r^2$ intercepts upon randomization of the response data indicate that the answers obtained at levels C and D are unsafe because the intercepts are slightly higher than the default threshold limit of 0.4. The $r^2$ intercept obtained with level E is in fact just below the limit at 0.39. This fact and the similarity of the other regression statistics to levels C and D indicate that the analysis provides comparable results at the three different basis sets. We decided to proceed with the QTMS analysis for all five sets.

Upon analysis of the extracted PCs we represent the ordering of the components in color-coded form in Figure 7. As we have seen before, different levels of calculation produce various orderings of the VIP scores, which gradually decrease in value except for level A. This level includes only distances and is the only level that unequivocally points at the COOH group as the highlighted region, despite the mediocre regression statistics. It should be noted that there is substantial drop in the value of the VIPs beyond the COOH group. The highlighted regions at each level of calculation show substantial overlap and are biased toward the COOH group and away from the side of the molecule where the substituent is attached. The carboxyl group (except level E) and the BCPs 6−13, 6−14, 10−19, and 10−20 are all shown to be important. As with the phenylacetic acids we find some unexpected BCPs being highlighted.

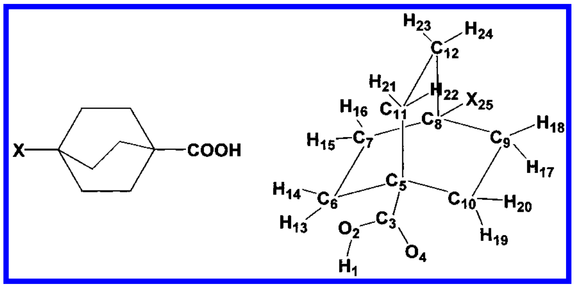

**Figure 6.** Labeling scheme of the common molecular skeleton of the 4-substituted bicyclo[2.2.2]octane-1-carboxylic acid. X represents the substituents given in Table 8.

**Table 8.** Substituents for *Para*-Substituted Phenyl Acetic Acids with Associated $\sigma^I$ Values[a]

| X | OMe | Me | $CH_2Me$ | H | F | Br | Cl | $CF_3$ | CN | $NO_2$ | OH |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma^I$ | 0.29 | 0 | −0.01 | 0 | 0.43 | 0.45 | 0.45 | 0.38 | 0.58 | 0.64 | 0.28 |

[a] Me = methyl.

**Table 9.** Correlation Coefficients for PLS Analysis of $\sigma^I$ Regressed against the BCP Properties of 4-X-Bicyclo[2.2.2]octane-1-carboxylic Acids

| level | $r^2$ | $q^2$ | $r^2$ (int) | $q^2$ (int) | #LVs |
|---|---|---|---|---|---|
| A | 0.62 | 0.54 | OK | OK | 1 |
| B | 0.94 | 0.92 | OK | OK | 1 |
| C | 0.99 | 0.95 | 0.45 | OK | 2 |
| D | 0.99 | 0.96 | 0.45 | OK | 2 |
| E | 0.99 | 0.96 | OK | OK | 2 |

It is unclear why levels C and D are statistically not as well validated as the others yet produce comparable results. It is probable that the use of two LVs in the PLS analysis increases the likelihood of chance correlation. The fact that level E, also with two LVs, is close to the validation cutoff point (0.39) indicates that this premise is correct. Although the problems encountered with linear least squares regression when more variables than cases are included are generally overcome by PLS, the ratio of cases to LVs can affect the validation. As a rough rule of thumb one should only employ one LV for every five molecules in the series. More LVs than this can lead to poor validity. The range of $\sigma^I$ values used in the regressions above was not very large, although the cases-to-LVs ratio is satisfactory.

**Table 10.** Substituents for *Meta*-Substituted Benzoic Acids with Associated $\sigma_m$ Values[a]

| X | $NMe_2$ | NHMe | $NH_2$ | OMe | Me | $CH_2Me$ | $CHCH_2$ | H | F | SH | Br | Cl | $CF_3$ | CN | $NO_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma_m$ | −0.16 | −0.21 | −0.16 | 0.12 | −0.07 | −0.07 | 0.06 | 0 | 0.34 | 0.25 | 0.39 | 0.37 | 0.43 | 0.56 | 0.71 |

[a] Me = methyl.

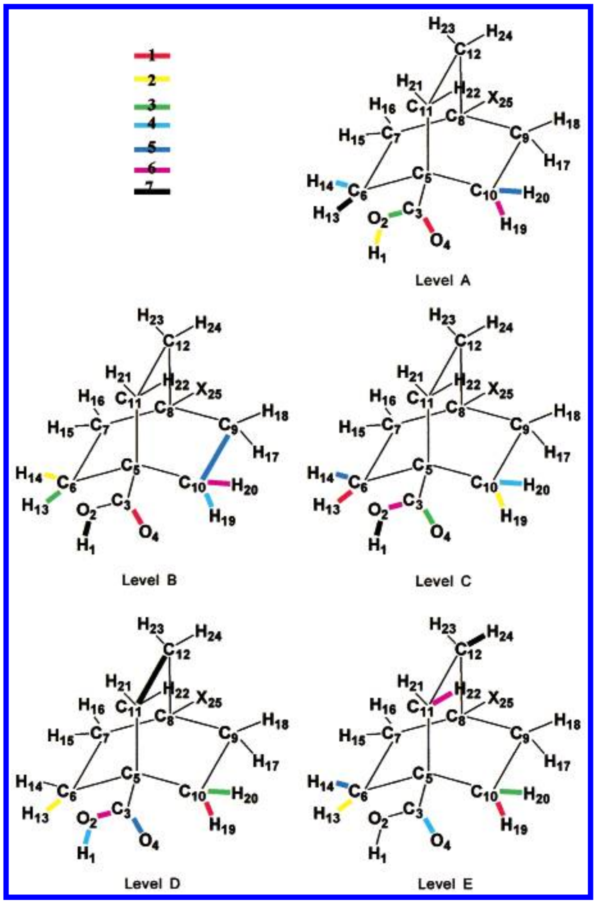

**Figure 7.** Color coding to denote the most important bonds in the PLS regression of $\sigma^I$ against the PCs drawn from the BCP properties of 4-X-bicyclo[2.2.2]octane-1-carboxylic acids at each level of calculation (A,B,C,D,E, Table 2).

Validation criteria are absolutely necessary if a QSAR is to be in any way useful. Because all other factors (other than the randomized $r^2$ intercept) indicate a good QSAR we can be fairly sure that levels C and D provide significant answers. However, in the absence of other levels of calculation with which to compare the results, it is prudent to disregard QSARs that produce insufficient validation.

## *META*-BENZOIC ACID

*Meta*-substitution does not allow significant through resonance[28,34] and hence certain functional groups induce significantly different acidities in *meta*-substituted benzoic acids from those in the *para*-substituted forms. This fact has given rise to $\sigma_m$ constants. We have computed the wave functions of 15 different *meta*-substituted benzoic acids (labeling scheme in Figure 8) at all five levels (Table 2). The substituent list, along with the appropriate $\sigma_m$ constants, is shown in Table 10. As the molecules are rigid the issue of conformation does not arise. If the substituent is geometrically flexible the lowest energy conformation is used.

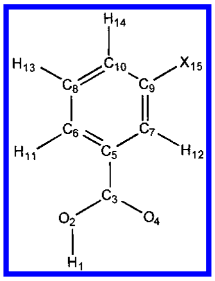

**Figure 8.** Labeling scheme of the common molecular skeleton of the *meta*-substituted benzoic acids. X represents the substituents given in Table 10.

**Table 11.** Correlation Coefficients for PLS Analysis of $\sigma_m$ Regressed against the BCP Properties of *Meta*-Substituted Benzoic Acids

| level | $r^2$ | $q^2$ | $r^2$ (int) | $q^2$ (int) | #LVs |
|---|---|---|---|---|---|
| A | 0.96 | 0.90 | OK | OK | 4 |
| B | 0.98 | 0.96 | OK | OK | 2 |
| C | 0.99 | 0.97 | OK | OK | 1 |
| D | 0.99 | 0.97 | OK | OK | 2 |
| E | 0.99 | 0.98 | OK | OK | 2 |

The results of the PLS analysis when carried out with all the BCP variables are shown in Table 11. All levels of calculation reproduce $\sigma_m$ and are valid. The color-coded results of the PLS analysis using the PCs are shown in Figure 9. Although the ordering of the BCP importance varies with the level of calculation the highlighted region remains fairly constant. The carboxyl group and the immediately adjacent regions of electron density maintain high importance. In addition the BCP 8−13 is consistently shown to be significant in the regression. The reason this BCP is a member of the highlighted region is unclear in view of conventional interpretation of acidity.

## POLYSUBSTITUTED BENZOIC ACIDS − ADDITIVITY OF SIGMAS

The $\sigma$ constants examined so far have all resulted from single substitutions. It has been suggested that one can reproduce the combined electronic effect of two (or more) substituents by simply summing the individual substituent constants.[35] If there is no interaction between the moieties this assumption is valid; however, there are cases where this does not hold.[36] Steric interactions can twist the substituents out of the ring plane and inhibit resonance or other effects. Table 12 shows a selection where the addition of $\sigma$s does *not* match experimental measurements. These were obtained from ref 36. The reference system is again the acidity of substituted benzoic acids.

We have computed the set of 15 poly-substituted benzoic acids (Table 12, labeling scheme in Figure 10). Wherever possible the substituents were modeled as having $C_s$ symmetry. The regression results, shown in Table 13, have been obtained after omission of an outlier, i.e., the 3,4,5-tri-$OCH_3$ substituted molecule. The VIP plot calculated at level E is shown in Figure 11. Unlike in the previous cases the three most important factors are the same for all the levels of calculation. In addition to this they clearly contribute much more to the regression than any other variables.

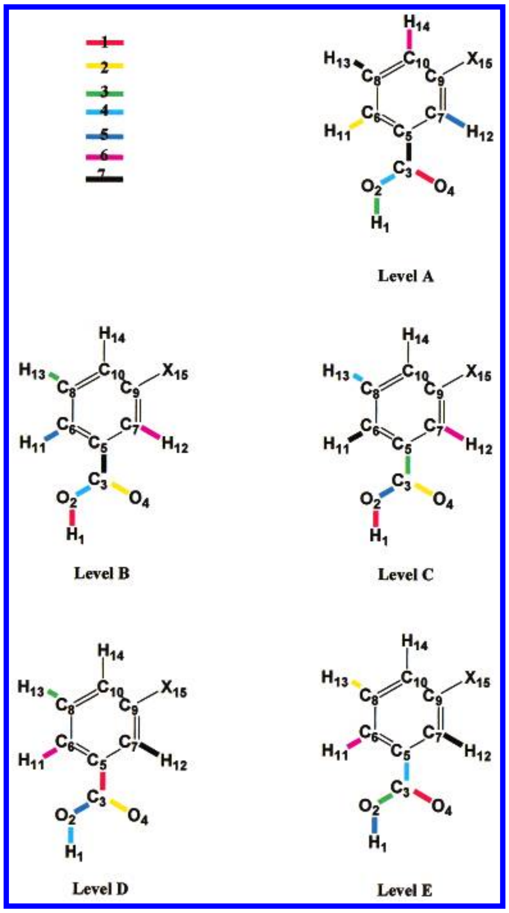

**Figure 9.** Color coding to denote the most important bonds in the PLS regression of $\sigma^I$ against the PCs drawn from the BCP properties of *meta*-substituted benzoic acids at each level of calculation (A,B,C,D,E, Table 2).

**Table 12.** Substituents with Their Observed and Combined Substituent Constants

| substituents | observed $\sigma$ | $\Sigma\sigma$ | $\lvert\Delta\sigma\rvert$ |
|---|---|---|---|
| H | 0 | 0 | 0 |
| 3,4-di-Cl | 0.52 | 0.60 | 0.08 |
| 3-Cl, 4-$OCH_3$ | 0.27 | 0.10 | 0.17 |
| 3-Br, 4-$CH_3$ | 0.15 | 0.22 | 0.07 |
| 3-$CH_3$, 4-$OCH_3$ | −0.26 | −0.34 | 0.08 |
| 3-$CH_3$, 4-$N(CH_3)_2$ | −0.30 | −0.90 | −0.60 |
| 3-$OCH_3$, 4-OH | −0.33 | −0.25 | 0.08 |
| 3-$NO_2$, 4-$NO_2$ | 1.38 | 1.49 | 0.11 |
| 3-$NO_2$, 4-Br | 0.83 | 0.94 | 0.11 |
| 3-$NH_2$, 4-$CH_3$ | −0.21 | −0.33 | 0.12 |
| 3-$N(CH_3)_2$, 4-$CH_3$ | −0.18 | −0.32 | 0.14 |
| 3-$OCH_3$, 5-$OCH_3$ | 0.05 | 0.24 | 0.19 |
| 3-OH, 5-OH | 0.16 | 0.24 | 0.08 |
| 3,4,5-tri-$OCH_3$ | 0.07 | −0.03 | 0.10 |
| 3-OH, 4-$OCH_3$, 5-$NO_2$ | 0.63 | 0.56 | 0.07 |

In both the graphs BCPs 1−2, 2−3, and 3−4 are evidently considerably more important than the other factors. At level E BCP 3−5 also stands clear of the rest of the variables. The prominence of BCP 3−5 is similarly seen at level B but not at levels A, C, or D. In this series of molecules, far more than in any other so far examined, the highlighted region is clearly indicated, because it encompasses predominantly the carboxyl group. The obvious distinction of this region of electron density as being relevant in describing the molecular acidities is most encouraging. However, there is no indication as to why the highlighted region is so strictly defined in these molecules and not in others, where we witness a gradual drop in VIP scores.

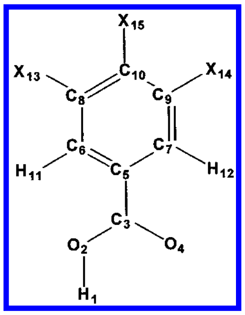

**Figure 10.** Labeling scheme of the common molecular skeleton of a poly-substituted benzoic acids system. If X is not specified in Table 12 it is a hydrogen atom.

**Table 13.** Correlation Coefficients for PLS Analysis of the Observed $\sigma$ Values Regressed against the BCP Properties of Substituted Benzoic Acids, after Omission of the 3,4,5-tri-$OCH_3$ Substituted Molecule

| level | $r^2$ | $q^2$ | $r^2$ (int) | $q^2$ (int) | #LVs |
|---|---|---|---|---|---|
| A | 0.96 | 0.88 | OK | OK | 2 |
| B | 0.96 | 0.92 | OK | OK | 1 |
| C | 0.96 | 0.90 | OK | OK | 1 |
| D | 0.95 | 0.86 | OK | OK | 1 |
| E | 0.94 | 0.88 | OK | OK | 1 |

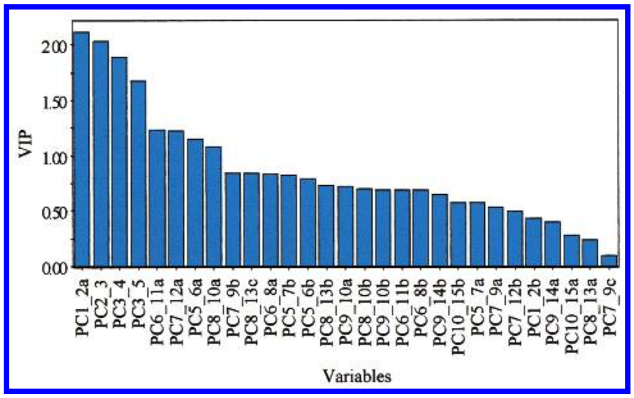

**Figure 11.** VIP Plot for PLS analysis of trisubstituted benzoic acids versus observed activities, using PCs as descriptor variables, calculated at level E (see Table 2).

Since $\sigma$ constants alter between systems several different such parameters have been defined and measured to reflect the variability of substituent effects. In 1968 Swain and Lupton had found 43 different $\sigma$s.[37] In each case the constant had to be measured on a given reference system and then transferability to similar systems assumed. As we have demonstrated, BCP properties provide a method of reproducing and predicting a substituent effect that can be applied universally. We do not have to transfer $\sigma$ constants between comparable systems, but, in essence, we measure the slightly differing substituent constants as they appear in the molecules of interest. The plethora of $\sigma$s can be replaced by one measure. Moreover, given the excellent regression the $\sigma$ value of a substituent for which no value is tabulated can be estimated.

## MOLECULAR QUANTUM SIMILARITY

In this section we contrast our work with the area of molecular quantum similarity (MQS)[38] pioneered and largely developed by the Carbó group. It should be emphasized that our method is *not* a variant of the Carbó method since it differs from it in several key points, as will become clear. As stated in the Introduction however QTMS shares the same philosophy as the Carbó method but turns out to be computationally less demanding. Moreover QTMS *avoids* (rather than solves) typical problems confronting Carbó's MQS.

The following formula is the definition of a general molecular quantum similarity measure (MQSM) used in QSAR as well as QSPR[39]

$$Z_{AB}(\Omega) = \int\int \rho_A(\mathbf{r}_1)\Omega(\mathbf{r}_1,\mathbf{r}_2)\rho_B(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 \qquad (1)$$

where $\Omega$ is a positive definite operator and $\rho_A$ and $\rho_B$ are the first-order electron densities corresponding to molecules *A* and *B*, respectively. General surveys of molecular quantum similarity have been published recently.[40,41] The original Carbó similarity index[42] is a so-called MQS index (MQSI) and measures the degree of overlap between two electron densities, $\rho_\mathrm{A}$ and $\rho_\mathrm{B}$. It can be expressed in terms of MQSM as follows, setting $\Omega$ to the Dirac delta function $\delta(\mathbf{r}_1 - \mathbf{r}_2)$:

$$C_{AB} = \frac{Z_{AB}}{\sqrt{Z_{AA}Z_{BB}}} \qquad (2)$$

Normalization factors appear in the denominator in order to ensure that the index takes values between 0 and 1, where 1 signifies perfect similarity. The application of MQSI certainly has merit in the light of the considerable attention it has received over the years,[43−46] but unfortunately it poses some challenges, which can be avoided by using the alternative route proposed by QTMS.

We briefly review four well-known difficulties and proposed remedies that arise when using the Carbó index or the more recent MQSM approach. We discuss each problem in the following order: the molecular superposition problem, the computational cost, the dominance of electron density near the nuclei, and the molecular fragment problem.

**Molecular Superposition.** The integrals appearing in eqs 1 or 2 must be maximized in order to obtain a meaningful index. This maximization invokes a 3D superposition of each pair of molecules under comparison, which is a complicated problem that has frequently been addressed.[47−49] However several successful MQSM studies[50−52] using only a self-similarity measure ($Z_{AA}$) make an optimization depending on relative position unnecessary. In the general case however, reliably finding the global maximum when aligning molecules is inherently complex. Each set of superpositions contains several local maxima, all but one is undesirable. The issue becomes not only finding the global maximum but its unbiased and reproducible identification.[53] The location of these maxima involves the calculation of a huge number of multicenter integrals.[54] The issue of molecular superposition has been tackled in a much wider context than merely quantum molecular similarity.[48] The dependence of activity upon conformation is not generally explored in quantum molecular similarity and most studies aim to compare either lowest energy structures or structures obtained from X-ray diffraction experiments.[55,56] Many methods have been used to improve[53] both the robustness and the speed of superposition, such as simple least-squares fits,[45] the Simplex method,[57] Monte Carlo optimization[49] and application of Fourier transforms.[47] Each technique has advantages and drawbacks, but no single method appears to dominate present day quantum molecular similarity studies.

**Computational Cost.** For a set of molecular wave functions that are calculated at a reasonable level the expense of directly comparing each electron density function can be prohibitive.[43,45,58] Some groups have examined semiempirical methods for obtaining densities,[45] while Carbó's group has used a series of spherical approximations to the density in their atomic shell approximation (ASA) approach.[58] However the ASA approximation is not an inherent feature of MQSM, and indeed calculations are performed at full ab initio level, which is not that time-consuming for self-similarity measures. The parametrization of ASA's densities certainly shortens the whole process considerably. It is claimed that results obtained with these promolecular densities are comparable to those found using the full density, which may be due to the overemphasis placed on the nuclear electron densities.

**Dominance of Core Electron Density.** This observation leads to a third problem with the Carbó index, namely the degree to which the measure is biased by core densities. The nuclear electron densities dominate the total density, whereas it may be the valence electron regions that are involved in binding or reactions. This feature of $\rho$ also contributes to the superposition problem. When maximizing the overlap between two densities we are locked into emphasizing those regions that contribute most to the density,[59] namely the nuclei. The degree of bias produced by core densities has been tackled in different ways. Bowen-Jenkins and Richards[46] sought to use valence electrons, however, probably due to the unphysical splitting of core and valence densities, further results were not so good. Momentum densities[60,61] were used which emphasize the variation of the outer-valence electron density that is chemically interesting. This approach retains the full molecular electron density and does not involve arbitrary cutoffs. Use of momentum densities has been also been extended and incorporated into studies involving Shannon entropies.[62]

**Molecular Fragment Problem.** Finally, as with most molecular similarity techniques, it is not clear which region(s) of the molecules to include in the overlap maximization. Some time ago Lee and Smithline[63] proposed a method based on arbitrary and nonunique fragment densities. But the following question remains: should the whole molecule always be studied or only those parts that we recognize[64] as being involved in a reaction.[50] Ponec et al. examine specific parts of a molecule when the reaction center is known.[51,65] However, they state that this is "possible only when the reaction center can be unambiguously determined". In the majority of cases, and certainly when novel drugs are being considered, the mode of action is not very well established. However, very recently MQSM applications have appeared (ref 66 and pages 73−83 in ref 38) where subsets of molecular fragments and their combinations are systematically incorporated in a given QSAR and assessed for relevance by statistical methods.

## HOW QTMS AVOIDS SOME MQSM ISSUES

**Molecular Superposition.** For the studied sets of carboxylic acids one does not need the superposition procedure in order to obtain excellent QSARs. This is not to say that conformation has no influence in our approach. Molecular structure and geometry, and hence conformation, express themselves entirely in the abstract space spanned by BCP properties. The actual comparison between molecules against a measured response occurs in that abstract space, not in real 3D space. This means that if a QSAR model is good enough for a set of rigid systems, the superposition was avoided at the level of comparison. If the QSAR is not good enough, conformational flexibility may be added, creating more refined data in the abstract property space.

**Computational Cost.** The computational cost is severely reduced by two factors: there are no costly integrals to be computed appearing in eqs 1 or 2, and the molecular superposition is avoided. Note that the location of critical points requires only a fraction of the time needed to compute the wave function.

**Dominance of Core Electron Density.** QTMS is not dominated by core densities. The high densities near the nuclei are not ignored in our approach since we use the *total* molecular density but implicitly influence the properties at the BCPs.

**Molecular Fragment Problem.** Finally, we do not determine a priori the molecular fragment that is responsible for a given activity. Instead this fragment follows naturally from the PLS analysis, although it can be more diffuse than desired. In other words, the VIP analysis ranks the variables according to their importance in a continuous way, without introducing any cutoff. In an earlier version of QTMS, published in paper 1 and applied to substituted benzoic acids, which did not resort to PLS, it was shown that the regression worsened considerably if the whole common molecular skeleton was included, rather than just the COOH group. So again the relevant molecular fragment, responsible for reproducing the studied activity is dictated by the activity itself, rather than by arbitrary *a priori* assumptions.

## CURRENT LIMITATIONS AND FUTURE WORK

In view of its ab initio character QTMS will increasingly take advantage of faster computers, such that larger sets of large molecules can be tackled. The work presented here benefited from the common skeleton found in the benzoic acids. Although *para*- and *meta*-subtituted benzoic acids have been treated separately recent unpublished work shows that they can be regressed together in a single set. This constitutes the first extension of QTMS beyond a common skeleton with a priori matched nuclei of the same atomic number (Z). Further extensions and generalizations of the matching procedure are planned. At the moment there is no firm statistically based cutoff criterion to isolate the highlighted region (which contains the active center) from the VIP plot. Preliminary work suggests that larger QSAR sets make the active center stand out more. In other words, the profile of the VIP plot decreases fairly abruptly after the atoms of the active center (e.g. COOH). Currently we have no satisfactory explanation for the spurious fragments (such as C−H) occurring in the highlighted region, but there is a possibility that their importance in the VIP plot will decrease when the number of substituents in the QSAR is increased. Furthermore a rigorous criterion for the dimensionality of BCP space has not been put forward. In the absence of any underlying theory that would fix this number (if possible) only experience can suggest an appropriate dimensionality. Finally, QTMS has not been extended to nuclear quantum similarity measures, which already exist in MQSM. However, QTMS' main development priority is the area of biologically and ecologically relevant molecules. In summary, we see QTMS as a new molecular similarity method with the generation and partitioning of quantum information as centerpieces. The interpretation of this information is subject to known methodologies such as PLS, neural networks, cluster analysis, or genetic algorithms. A considerable amount of work is planned to investigate the best choices in the three segments of QTMS: generation, partitioning, and interpretation.

## CONCLUSION

We have presented a new molecular similarity method based on the topology of the electron density, which is directly applicable to QSARs. This method is called quantum topological molecular similarity (QTMS) and is under continuous development. Although it shares the use of the electron density $\rho$ with the original Carbó index and its generalization to quantum molecular similarity measures (MQSM), its practical implementation differs substantially from it and its variants. As a consequence QTMS *avoids* four important challenges of MQSM. The QTMS method directly compares discrete topological representations of molecules without 3D superposition, using properties evaluated at the bond critical points (BCP). The rate-determining step of the method is the quantum chemical calculation since the actual molecular comparison is very fast. Moreover QTMS does not suffer from core density dominance. Currently QTMS is coupled to a rigorous statistical analysis based on partial least squares (PLS). This chemometric method is used to assess the significance and quality of the regressions as well as to suggest the so-called highlighted region. This region is the molecular fragment corresponding to the BCPs that are most significant in explaining the activity.

We have proven the applicability of QTMS to five carboxylic acid systems at five different levels of calculation. Each level benefits from the geometry optimization of the lower level since successively updated geometries are obtained. All levels of calculation provided very good regressions. In most cases good results were even seen with AM1 derived bond lengths. This level (A) is approximately 3 orders of magnitude faster than the highest ab initio level (level E). However, very poor results are also seen with the AM1 method. Ab initio derived BCP space does not fail in the same way, either because BCP space provides a much richer description of $\rho$ or because of the inadequacies of the semiempirical method itself. It is proposed that ab initio level calculations should be employed in conjunction with AM1 to ensure the accuracy of the method for a given data set. If AM1 provides consistently good results further investigation can be done at lower cost and on larger variants of the system.

A bonus to QTMS's predictive power of the measured response is that it also provides a highlighted region, which contains the active center provided one has several levels of calculations to draw from. The highlighted zone is not decided a priori but follows from the VIP scores. Unfortunately the VIP profile decreases gradually thereby suggesting a rather diffuse highlighted zone. This means that the active center may lie buried in larger molecular fragment, which contains unexpected bonds.

In summary, we propose novel quantum chemical descriptors that can replace electronic parameters in a QSAR. These descriptors are discrete and are drawn from quantum chemical calculations via the topology of the electron density. We do not make any assumptions about the underlying mechanism of activity in the statistical treatment.

In future communications we will publish QTMS applied to biologically relevant QSARs as most encouraging results have been obtained with it. We plan to extend the range of AIM topological descriptors with atomic properties and critical points from scalar fields other than the electron density, for systems where bond breaking is not part of the activity. Equally we are aware that QTMS may suffer from the inherent drawbacks and limitations of PLS.

## ACKNOWLEDGMENT

Gratitude is expressed to EPSRC for a studentship (GR/L65895). The referees are thanked for their extensive comments, which allowed us to make the comparison with MQSM more balanced.

## REFERENCES AND NOTES

(1) Popelier, P. L. A. Quantum molecular similarity. 1. BCP space. *J. Phys. Chem. A* **1999**, *103*, 2883−2890.

(2) Popelier, P. L. A. *Atoms in Molecules: an introduction*; Pearson, H., Ed.; London, 2000.

(3) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Clarendon Press: 1990.

(4) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864−B871.

(5) Ziegler, T. Approximate Density Functional Theory As a Practical Tool in Molecular Energetics and Dynamics. *Chem. Rev.* **1991**, *91*, 651−667.

(6) As shown in this paper QTSM does not extract its data from pure ab initio calculations alone. QTSM may rely on semiempirical results, which draw from experimental data, or from so-called hybrid density functionals such as the popular B3LYP, which contain parameters fine-tuned with experimental data.

(7) Bader, R. F. W.; Anderson, S. G.; Duke, A. J. Quantum Topology of Molecular Charge Distributions. 1 *J. Am. Chem. Soc.* **1979**, *101*, 1389−1395.

(8) Popelier, P. L. A. In *Molecular Similarity and complimentarity based on the theory of atoms in molecules*; Dean, P. M., Ed.; London, 1995.

(9) Wold, S.; Sjostrom, M. In *Partial Least Squares Projections to Latent Structures (PLS) in Chemistry*; Schleyer, P. v. R., Ed.; New York, 1998.

(10) O'Brien, S. E. Ph.D. Thesis, Department of Chemistry, UMIST, Manchester, 2000.

(11) Popelier, P. L. A.; Aicken F. M.; O'Brien S. E. *Specialist Periodical Report, Chemical Modelling: Applications & Theory*; Royal Society of Chemistry: Hinchliffe, A., Ed.; 2000; Vol. 3, pp 143−198.

(12) O'Brien, S. E.; Popelier, P. L. A. Quantum molecular similarity. Part 2: The relation between properties in BCP space and bond length. *Can. J. Chem.* **1999**, *77*, 28−36.

(13) Aicken, F. M.; Popelier P. L. A. Atomic properties of selected biomolecules. Part 1. The interpretation of atomic integration errors. *Can. J. Chem.* **2000**, *78*, 415−426.

(14) Frisch, M. J. T.; Schlegel, G. W.; Scuseria, H. B. G. E.; Robb, M. A.; J. R. C.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; J. C. B.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; K. N. K.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; M. C.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; J. O.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; D. K. M.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; J. C.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; G. L.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; R. L. M.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; A. N.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; B. J.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; M. H.-G.; Replogle, E. S.; Pople, J. A. In GAUSSIAN98; Pittsburgh, 1998.

(15) Hansch, C. Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; Heller, S. R., Ed.; ACS Professional Reference Books, 1995.

(16) Popelier, P. L. A. In Morphy98; MORPHY98 − a program written by P. L. A. Popelier with a contribution from R. G. A. Bone; UMIST, Manchester, England, 1998.

(17) Popelier, P. L. A. A Robust Algorithm to Locate Automatically All Types of Critical- Points in the Charge-Density and Its Laplacian. *Chem. Phys. Lett.* **1994**, *228*, 160−164.

(18) UMETRICS, A. SIMCA-P 8.0 User Guide and Tutorial; 1999.

(19) UMETRICS. In SIMCA-P 8.0; 1998; http://www.umetrics.com.

(20) Livingstone, L. *Data Analysis for Chemists*; Oxford University Press: 1995.

(21) Wold, S.; Kettaneh, N.; Tjessem, K. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *J. Chemometrics* **1996**, *10*, 463−482.

(22) SPSS Inc., version 10.0.7; 2000; http://www.spss.com.

(23) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J. The Collinearity Problem in Linear-Regression − the Partial Least- Squares (PLS) Approach to Generalized Inverses. *Siam J. Sci. Statist. Comput.* **1984**, *5*, 735−743.

(24) Berglund, A.; Wold. S. A serial extension of multiblock PLS. *J. Chemometrics* **1999**, *13*, 461−471.

(25) Westerhuis, J. A.; Kourti, T.; MacGregor, J. F. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemometrics* **1998**, *12*, 301−321.

(26) Leach, A. *Molecular modelling: principles and applications*; Longman; 1996.

(27) O'Brien, S. E.; Popelier, P. L. A. *Quantum Molecular Similarity: Use of Atoms in Molecules derived quantities as QSAR variables*; ECCOMAS: Barcelona, Spain, 2000.

(28) Hammett, L. P. *Physical Organic Chemistry*; McGraw-Hill: 1970.

(29) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*; Heller, S. R., Ed.; American Chemical Society: 1995.

(30) Ludwig, M.; Wold, S.; Exner, O. The Role of Meta-Benzene and Para-Benzene Derivatives in the Evaluation of Substituent Effects − a Multivariate Data-Analysis. *Acta Chem. Scand.* **1992**, *46*, 549−554.

(31) Stutchbury, N. C. J.; Cooper, D. L. Charge Partitioning By Zero-Flux Surfaces − the Acidities and Basicities of Simple Aliphatic-Alcohols and Amines. *J. Chem. Phys.* **1983**, *79*, 4967−4972.

(32) Taft, R. W.; Ehrenson, S.; Lewis, I. C.; Glick, R E. *J. Am. Chem. Soc.* **1959**, *81*, 5352−5355.

(33) Roberts, J. D. Moreland, W. T. *J. Am. Chem. Soc.* **1953**, *75*, 2167−2170.

(34) Johnson, C. D. *The Hammett Equation*; Cambridge University Press: 1973.

(35) Jaffe, H. H. *Chem. Rev.* **1953**, *53*, 191−222.

(36) Hansch, C.; Leo, A.; Unger, S. H.; Kim, K. H.; Nikaitani, D.; Lein, E. J. *J. Med. Chem.* **1973**, *16*, 1207−1213.

(37) Swain, C. G.; Lupton, E. C. *J. Am. Chem. Soc.* **1968**, *90*, 4328−4331.

(38) Carbo-Dorca, R.; Robert, D.; Amat, L.; Girones, X.; Besalu, E. *Molecular Similarity in QSAR and Drug Design*; Springer: 2000.

(39) Girones, X.; Amat, L.; Robert, D.; Carbo-Dorca, R. *J. Comput.-Aided Mol. Design* **2000**, *14*, 477−485.

(40) Carbo-Dorca, R.; Besalu, E. *J. Mol. Struct (THEOCHEM)* **1998**, *451*, 11−23.

(41) Carbo-Dorca, R.; Amat, L.; Besalu, E.; Girones, X.; Robert, D. *J. Mol. Struct. (THEOCHEM)* **2000**, *504*, 181−228.

(42) Carbo, R.; Leyda, L.; Arnau, M. How Similar is a Molecule to Another? An Electron Density Measure of Similarity between Two Molecular Structures. *Intl. J. Quantum Chem.* **1980**, *17*, 1185−1189.

(43) Richards, W. G.; Hodgkin, E. E. Molecular Similarity. *Chem. Br.* **1988**, *24*, 1141−1143.

(44) Ponec, R. Similarity Models in the Theory of Pericyclic Macromolecules. *Top. Curr. Chem.* **1995**, *174*, 1−26.

(45) Hodgkin, E. E.; Richards, W. G. A Semiempirical Method For Calculating Molecular Similarity. *J. Chem. Soc., Chem. Commun.* **1986**, 1342−1344.

(46) Bowen-Jenkins, P. E.; Richards, W. G. Molecular Similarity in Terms of Valence Electron-Density. *J. Chem. Soc., Chem. Commun.* **1986**, 133−135.

(47) Nissink, J. W. M.; Verdonk, M. L.; Kroon, J.; Mietzner, T.; Klebe, G. Superposition of molecules: Electron density fitting by application of Fourier transforms. *J. Comput. Chem.* **1997**, *18*, 638−645.

(48) Miller, M. D. In *Molecular Superposition*; Encycl. of Comput. Chem.; Schleyer, P. v. R., Ed.; Wiley: New York, 1998.

(49) Parretti, M. F.; Kroemer, R. T.; Rothman, J. H.; Richards, W. G. Alignment of molecules by the Monte Carlo optimization of molecular similarity indices. *J. Comput. Chem.* **1997**, *18*, 1344−1353.

(50) Amat, L.; Carbo-Dorca, R.; Ponec, R. Simple linear QSAR models based on quantum similarity measures. *J. Med. Chem.* **1999**, *42*, 5169−5180.

(51) Ponec, R.; Amat, L.; Carbo-Dorca, R. Molecular basis of quantitative structure-properties relationships (QSPR): A quantum similarity approach. *J. Comput.-Aided Mol. Design* **1999**, *13*, 259−270.

(52) Ponec, R.; Amat, L.; Carbo-Dorca, R. Quantum Similarity approach to LFER: substituent and solvent effects on the acidities of carboxylic acids. *J. Phys. Org. Chem.* **1999**, *12*, 447−454.

(53) Constans, P.; Amat, L.; Carbo-Dorca, R. Toward a global maximization of the molecular similarity function: Superposition of two molecules. *J. Comput. Chem.* **1997**, *18*, 826−846.

(54) Bowen-Jenkins, P. E.; Richards, W. G. Quantitative Measures of Similarity Between Pharmacologically Active Compounds. *Intl. J. Quantum Chem.* **1986**, *30*, 763−768.

(55) Leherte, L.; Allen, F. H. Shape Information From a Critical-Point Analysis of Calculated Electron-Density Maps − Application to DNA-Drug Systems. *J. Comput.-Aided Mol. Design* **1994**, *8*, 257−272.

(56) Leherte, L.; Fortier, S.; Glasgow, J.; Allen, F. H. Molecular Scene Analysis − Application of a Topological Approach to the Automated Interpretation of Protein Electron-Density Maps. *Acta Crystallogr. Sect. D* **1994**, *50*, 155−166.

(57) Hodgkin, E. E.; Richards, W. G. Molecular Similarity Based On Electrostatic Potential and Electric-Field. *Intl. J. Quantum Chem.* **1987**, *14*, 105−110.

(58) Amat, L.; Carbo- Dorca, R. Quantum similarity measures under atomic shell approximation: First-order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, *18*, 2023−2039.

(59) Kestner, N. R.; Combariza, J. E. In *Basis Set Superposition Errors: Theory and Practice*; Lipkowitz, K. B., Boyd, D. B., Eds.; Chichester, 1999.

(60) Allan, N. L.; Cooper, D. L. Momentum-Space Electron-Densities and Quantum Molecular Similarity. *Top. Curr. Chem.* **1995**, *173*, 85−111.

(61) Measures, P. T.; Mort, K. A.; Cooper, D. L.; Allan, N. L., A quantum molecular similarity approach to anti-HIV activity. *Theochem-J. Mol. Struct.* **1998**, *423*, 113−123.

(62) Ho, M.; Smith, V. H.; Weaver, D. F.; Gatti, C.; Sagar, R. P.; Esquivel, R. O. Molecular similarity based on information entropies and distances. *J. Chem. Phys.* **1998**, *108*, 5469−5475.

(63) Lee, C.; Smithline, S. An Approach to Molecular Similarity Using Density-Functional Theory. *J. Phys. Chem.* **1994**, *98*, 1135−1138.

(64) In the same vein assumptions have been made about the role the core electrons play in activity suggesting one could concentrate on valence electrons alone (ref 46).

(65) However, it was pointed out by one of the referees that a recent study submitted to this journal that a priori determination of a molecular fragment was eliminated.

(66) Robert, D.; Amata, L.; Carbo-Dorca, R. *Intl. J. Quantum Chem.* **2000**, *80*, 265−282.

(67) Foresman, J. B. Frisch, A. *Exploring Chemistry with Electronic Structure Methods*; Gaussian Inc.: 1996.

(68) Becke, A. D. Density-Functional Thermochemistry. 3. the Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

CI0004661