Quantum topological molecular similarity. Part 5. Further development with an application to the toxicity of polychlorinated dibenzo-*p*-dioxins† (PCDDs)

P. L. A. Popelier,* U. A. Chaudry and P. J. Smith

Dept. of Chemistry, UMIST, 88 Sackville Street, Manchester, UK M60 1QD.
E-mail: pla@umist.ac.uk

Received (in Cambridge, UK) 8th April 2002, Accepted 30th April 2002
First published as an Advance Article on the web 23rd May 2002

A new method called quantum topological molecular similarity (QTMS), which was previously introduced, is further developed and applied. An excellent and statistically validated QSAR is obtained for the Hammett acidity constants of a set of 68 carboxylic acids including *p*- and *m*-benzoic acids, *p*-phenylacetic acid, 4-X-bicyclo[2.2.2]octane-1-carboxylic acids and poly-substituted benzoic acids. This investigation shows that the previously imposed condition for a minimal and restricted common molecular skeleton can be relaxed. The O–H and the C–O bonds are recovered as the active center as expected. The first time use of atomic properties instead of bond properties leads to valid QSARs. Finally QTMS is applied to predict three different activities (pEC_{50}) of the ecologically relevant polychlorinated dibenzo-*p*-dioxins (PCDDs). We find that the active center is concentrated near the lateral C–Cl bonds.

1 Introduction

Relating the properties or activity of molecules to their structure is an area of scientific interest dating back to the second half of the 19th century.[1] The potential rewards of finding such relationships at quantitative level with relevance to the agrochemical or pharmaceutical industry cannot be underestimated. More recently studies of toxicity and biodegradability[2] have also benefited increasingly from quantitative structure–activity/property relationships (QSAR/QSPR) under growing environmental awareness.

As QSAR techniques established themselves[3] after the systematisation of Hansch and Fujita,[4] molecular similarity was considered as a source of descriptors.[5–9] The number and variety of theoretical molecular descriptors ever applied in QSAR is overwhelming. They have recently been classified by Karelson[10] into the following categories: constitutional and geometric, topological, electrostatic- or charge distribution-related, quantum chemical- or MO-related, solvational, thermodynamic or "combined". To the best of our knowledge the method we present here cannot be found in this extensive categorisation, but would reside in both the charge distribution and the quantum chemical categories.

Over the last few years we have been interested in injecting quantum mechanical data into QSAR/QSPRs, in particular by the topological approach[11] according to the theory of "atoms in molecules" (AIM).[12–16] AIM properties have also been used to model aromaticity[17] and hydrogen bond donor capacity.[18] The dramatic enhancement of computational power now makes it feasible to investigate the predictive capability of topological descriptors drawn from *ab initio* wavefunctions. The first successful use of AIM topological descriptors showed how they accurately predict Hammett σ values of *p*- and *m*-benzoic acids.[19] This led to the introduction of quantum topological molecular similarity (QTMS), which is a new method that we continue to explore and fine-tune.[20] The same paper[19] inspired the development of StruQT,[21] a 3D representation using quantum chemical topology. Quantum chemical topology generates graphs endowed with a physical basis, which is usually less well-defined in classical chemical graph theory, revisited by Wiener,[22] developed by Randic,[23,24] Balaban[25] and Hosoya[26] and extended by Kier and Hall.[27,28]

QTMS consists of three phases: generation of quantum data, extraction of topological descriptors, and model construction and interpretation. The essence of our method is reviewed in Section 3 and full details are explained elsewhere.[29] Since QTMS is a novel method its range of applicability is not clear at present, although ongoing work has produced a growing number of successful QSARs, subject to rigorous statistical treatment and with a minimum of chemometric manipulation. One of QTMS's main features is its ability to localise the "active center". More precisely QTMS is able to rank bonds according to their importance or influence in explaining the observed activity. It should be pointed out that the "active center" can be rather diffuse or contain unexpected bonds, which could turn out to be true "contaminations".

In the first part of this paper we demonstrate how the active center can be localised in an extensive set of carboxylic acids. As explained below we prove that QTMS can be used beyond a set of strictly congeneric molecules, with a restricted or minimal common skeleton. Then, having established the capacity of QTMS to point out the active center we predict it for a set of ecologically relevant molecules, namely polychlorinated dibenzo-*p*-dioxins (PCDD). As such, this study complements a previous one on medically relevant (*E*)-1-phenylbut-1-en-3-ones.[30] In this contribution we introduce for the first time (topological) *atomic* properties into QTMS.

2 Quantum chemical topology

This theory of AIM is the most elaborately researched and documented way of partitioning a quantum system (*e.g.* a molecule, van der Waals complex or crystal) into atomic constituents, based on the electron density ρ. AIM provides a consistent way of partitioning and hence localising chemical information, irrespective of the particular mathematical

† The IUPAC name for dibenzo-*p*-dioxin is dibenzo[*b*,*e*][1,4]dioxin.

DOI: 10.1039/b203412c

representation of ρ. In this paper we use *bond* properties and *atomic* properties to characterise a molecule, both formulated in the context of AIM.

It is a working hypothesis of QTMS that a bond can be described by quantum chemical functions evaluated at so-called *bond critical points* (BCPs). These are points, occurring roughly in between two bonded nuclei, where the gradient of the electron distribution vanishes (or $\nabla\rho = 0$). At the BCPs the Hessian of ρ has two negative eigenvalues ($\lambda_1 < \lambda_2 < 0$) and one positive one ($\lambda_3 > 0$). The sum of the eigenvalues is the Laplacian, denoted by $\nabla^2\rho$, which is a measure of how much ρ is concentrated ($\nabla^2\rho < 0$) or depleted ($\nabla^2\rho > 0$) in a point. This function, together with the electron density, are two components of the QTMS vector that describes a bond. A third component is given by the ellipticity,[13] denoted by ε, and defined as $(\lambda_1/\lambda_2) - 1$. It is always positive because $\lambda_1 < \lambda_2 < 0$ at the BCP, and is zero for a cylindrically symmetrical bond. A fourth component is a type of kinetic energy density,[31] denoted by $K(\boldsymbol{r})$.[13] It is defined as

$$K(\boldsymbol{r}) = -\tfrac{1}{4}N \int \mathrm{d}\tau' [\psi^*\nabla^2\psi + \psi\nabla^2\psi^*]$$

where $\int \mathrm{d}\tau'$ denotes an integration over the spin coordinates of all N electrons except one and ψ is the wavefunction. Finally we add the equilibrium bond length R_e as a fifth component. The justification for this addition is given Part 3,[29] partially based on work done in Part 2[32] on the relation between BCP properties and bond length. In summary, the QTMS bond descriptor vector in this paper consists of five components or $\boldsymbol{P}_{\mathrm{bond}} = (\rho, \nabla^2\rho, \varepsilon, K, R_e)$.

A second type of QTMS vector consists of atomic properties, which are computed as volume integrals over so-called atomic basins. The latter appear as bounded regions of real 3D space dominated by a so-called attractor, which is typically a nucleus. If one traces $\nabla\rho$ over a very short stretch, re-evaluates it, and traces it further, the traced path typically terminates in a nuclear attractor. The collection of paths terminating at an attractor constitutes an atomic basin, denoted by Ω. The atomic population $N(\Omega)$ is simply given as the volume integral of ρ over the atomic basin. The atomic dipole moment results from the integration of ρ times a position vector measured from the nucleus. Higher multipole moments are defined by multiplying the appropriate (spherical) tensors by ρ and integration over Ω.[13] We include the population, the magnitude of the dipole, quadrupole, octopole and hexadecapole moments, as well as the atomic volume and the kinetic energy into a seven-dimensional atomic descriptor vector $\boldsymbol{P}_{\mathrm{atom}}$. Although it is possible, in principle, to combine atomic and bond topological descriptors we have not yet carried out such an analysis.

3 QTMS

The full details of this method have been published before.[29] Basically there are three phases in a QTMS analysis. The first phase consists of the data generation based on quantum chemical methods ranging from semi-empirical to density functional theory (DFT) calculations. The second phase is the extraction of topological descriptors from the wavefunctions, including geometry-optimised bond lengths. The third phase encompasses the construction of a model, which is currently carried out by, but not confined to, partial least squares (PLS),[33] an advanced multi-linear regression technique.

Initially an estimated geometry is obtained from the interactive program MOLDEN,[34] which is passed on to the *ab initio* program GAUSSIAN98.[35] We use the semi-empirical method AM1,[36] which we call level A. Since this level of theory is unable to produce a sensible topology we only retrieve bond lengths from it. The next level of theory that we have been systematically using in our QTMS work[29,30,37] is HF/3-21G(d)//HF/3-21G(d), which means that according to the standard notation of Gaussian basis sets[38] the geometry has been optimised at the Hartree–Fock level using the 3-21G(d) basis set and that the wavefunction has also been obtained at this level. For convenience we designate this level as level B. In line with previous work level D corresponds to B3LYP/6-311+G(2d,p)//HF/6-31G(d), where the latter level was used only for the geometry optimization. Finally we report some results at B3LYP/6-311+G(2d,p)//B3LYP/6-311+G(2d,p), computationally the most expensive level, which we call level E.

Secondly the wavefunction is read by (a local version of) the program MORPHY98,[39] which locates the BCPs using an automatic and robust algorithm.[40] It also performs the atomic integrations.[41,42] At this stage the property vectors $\boldsymbol{P}_{\mathrm{atom}}$ and $\boldsymbol{P}_{\mathrm{bond}}$ are constructed yielding a discrete quantum fingerprint for each molecule in the QSAR set. The atomic integrations are much more CPU-intensive than the computation of BCP properties. Note that the addition of further BCP descriptors to $\boldsymbol{P}_{\mathrm{bond}}$ may introduce redundant variables in view of the presence of perfect colinearities. For example, since the Laplacian is the sum of the three Hessian eigenvalues, $\nabla^2\rho = \lambda_1 + \lambda_2 + \lambda_3$, the addition of the three eigenvalues to the descriptor set would not add any more information. In view of the increased number of descriptors the value of the cross-validated correlation coefficient q^2 (see below) will mostly decrease, which is undesirable.

Thirdly the program SIMCA-P program[43] performs the PLS analysis, which we have utilised with the recommended values.[44] A prime but potentially misleading measure of the quality of a regression is the correlation coefficient r^2. Hence it is customary that the r^2 is stated in conjunction with the cross-validated[45] r^2, denoted by q^2, which measures the internal predictivity, here based on SIMCA-P's default "leave one seventh of the data out" (rather than "leave one out"). To be protected against correlations found by chance we applied the rather stringent "randomisation validation test", measured by r^2_{int} and q^2_{int}, where "int" stands for "intercept". This test estimates the probability that a good fit is obtained after random permutation of the Y variables ("activities"). Each randomisation and subsequent PLS analysis generates a new r^2 and q^2 value, which is plotted against (the absolute value of) the correlation coefficient between the original set of Y variables and its permutation. A line is fitted through the r^2 values and another through the q^2 values, yielding the aforementioned intercepts r^2_{int} and q^2_{int}. Models are typically[33] validated if $r^2_{\mathrm{int}} < 0.4$ and $q^2_{\mathrm{int}} < 0.05$.

The program SIMCA-P prescribes a criterion for the significance of a *latent variable* (LV), *i.e.* if q^2 corresponding to a newly constructed LV is smaller than 0.097, then the LV is not significant to the model. In this case no more LVs are computed and the PLS regression is terminated. It is vital that we obtain a model that passes the randomisation test, no matter how high the r^2 and q^2 values turn out to be. For example, excellent regression statistics may be obtained with four LVs but without passing the randomisation test. In other words, there is no point in reporting a model with a high number of LVs and excellent r^2 and q^2 values if the model is obtained by pure chance in the first place (*i.e.* it does not pass the randomisation test).

Typically a model passes the randomisation test if the number of LVs is reduced. A gradual reduction of the number of LVs, until the randomisation test is passed, results in a steady omission of information. This manipulation is acceptable because it is automatically penalised by a concomitant decrease of the r^2 and q^2 values. In other words, the perhaps artificial reduction of information is automatically compensated by a deteriorating regression quality.

In order to interpret our model we compress the number of descriptors variables ("X") for each bond *via* principle component analysis (PCA),[45] performed by the program SPSS.[46] It is important to realise that PLS is carried out *again*, this time on the extracted PCs rather than on the "raw" variables, which are the components of $\boldsymbol{P}_{\mathrm{atom}}$ or $\boldsymbol{P}_{\mathrm{bond}}$. In summary, the first PLS

analysis yields the regression statistics ("quality and validity" of fit), whereas the second PLS analysis focuses on the interpretation of the regression. The so-called variable importance in the projection (VIP)[44] value is used to indicate the relative importance of each PC. We assume in QTMS that the "active center" of a molecule consists of the PCs with the highest VIP values.

It is convenient to cast the information of the VIP plot onto the molecular skeleton of the acids *via* a colour code. Variables with a VIP value less than one can be discarded as unimportant.[33] The bonds with *any* VIP value of more than one up to the highest value found in a given QSAR are assigned a colour. The bonds with VIP values smaller than one remain black. In other words, if at least one of the PCs that represent a bond has a VIP value larger than one it is assigned a colour (other than black). We propose to use five colours mirroring the visible spectrum by assigning red to the interval with the highest VIP values. The subsequent linearly spaced intervals of decreasing VIP values are assigned the colours yellow, green, blue and violet respectively. The colour code expresses the *relative* importance of the bonds inside a given QSAR and shows how well localised or diffuse the active center is. Since the colours have no significance as indicators of absolute VIP values they cannot be used to compare plots between different QSARs.

4 Carboxylic acids

In this section we focus on two general matters concerning QTMS: the common skeleton and the use of atomic properties *versus* bond properties. All wavefunctions were obtained at level E. In previous work[29] *p*-, *m*-benzoic and *p*-phenylacetic acids were *separately* regressed against their respective Hammett acidity constants, *i.e.* σ_p, σ_m and σ_p^0. The *p*-benzoic acids were described by 15 bonds including C_{10}–X_{15}, where X denotes the substituent. Fig. 1 shows the numerical labeling scheme of the

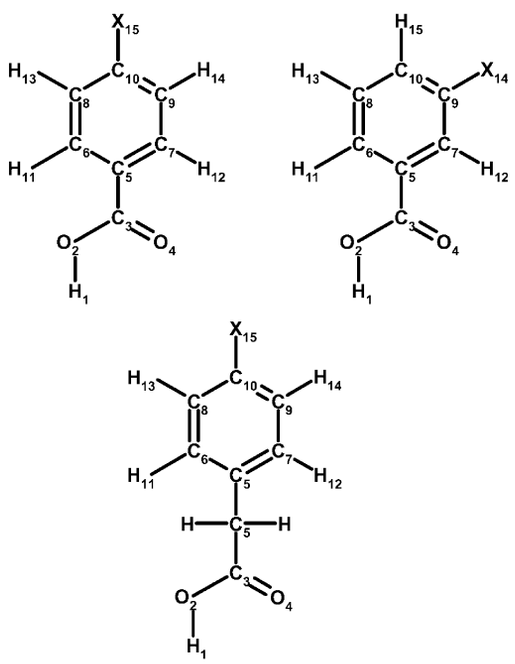

Fig. 1 Labeling scheme and common molecular skeletons for *p*-, *m*-benzoic and *p*-phenylacetic acids.

benzoic acids. Any QTMS analysis so far benefits from the fact that bonds between different molecules from a single congeneric set can be unambiguously mapped onto each other. This is because the only bonds allowed in the topological description are those that all molecules have in common, disregarding differences in the atomic number Z of the bonded nuclei. For example, in all *p*-benzoic acids the C_7–C_9 bonds can easily be identified and put in 1–1 correspondence with each other, and so can C_{10}–X_{15}, but bonds within very different substituents such as OCH_3 and NH_2 cannot. Hence these bonds do not belong to the common molecular skeleton. In summary there is no reason why *p*- and *m*-benzoic acids cannot be regressed in the same QSAR against their respective σ_p and σ_m values.

Moreover, with a small modification, the *p*-phenylacetic acids can be added to this set. These acids contain a methylene group that cannot be brought into correspondence with any fragment in the benzoic acids. This is why the methylene hydrogens in Fig. 1 do not have numerical labels. Furthermore, the methylene carbon is labeled as C_5 because it takes on the role of the atom to which the COOH group is bonded, which is the C_5 atom in the benzoic acids. As a result it is still possible to map the 15 bonds of the *p*-phenylacetic acids to the corresponding bonds in the benzoic acids. The currently preferred[47] values for the Hammett σ constants of all carboxylic acids, including two extra sets that are considered below, are shown in Table 1.

Fig. 2 shows the VIP plot of the benzoic and phenylacetic

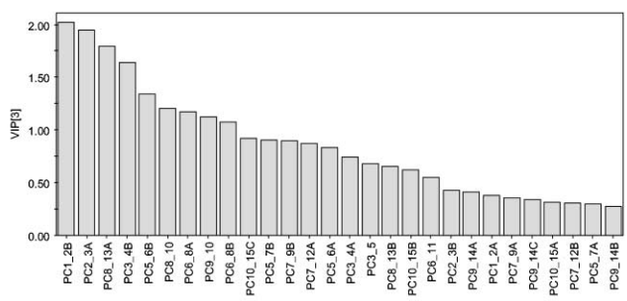

Fig. 2 Variable importance in the projection (VIP) plot for the PLS regression of *p*-, *m*-benzoic acids and *p*-phenylacetic acids.

acids based on 28 PCs, each representing topological information furnished by any of the 15 bonds of the common skeleton. Some bonds are characterised by more than one PC, for example, C_9–H_{14} appears as three different PCs (A, B and C).

It is clear from Fig. 3 that the active center (red) occurs where

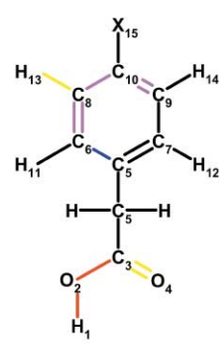

Fig. 3 Colour-coded plot expressing the influence (VIP) of the bonds in explaining the observed acidity of all benzoic and phenylacetic acids. The active site has the highest influence (red). Legend: red > yellow > green > blue > purple > black (see text).

one expects it since the O–H bond breaks and reforms in the establishment of the acids' equilibria with their ionic forms. The further the bonds are removed from the O–H bond, the more their importance diminishes with the exception of the C_8–H_{13}, which we could regard as a contamination. Whether this is an artifact of PLS is an open question.

A different Hammett parameter, denoted by σ^I, concentrates on the so-called field-inductive component of acidity relationships. An unsaturated system that lacks any resonance effects is 4-X-bicyclo[2.2.2]octane-1-carboxylic acid. We have added 11

Table 1 The Hammett acidity constants for all five sets of carboxylic acids

p-Benzoic	σ_p	*m*-Benzoic	σ_m	Phenyl	σ_p^0	Bicyclo	σ^I	Poly-subst. benzoic	$\Sigma\sigma$
H	0	$N(CH_3)_2$	−0.16	H	0	CH_3	0	3,4-Di-Cl	0.52
$N(CH_3)_2$	−0.83	$NHCH_3$	−0.21	$N(CH_3)_2$	−0.48	CH_2CH_3	−0.01	3-Cl, 4-OCH_3	0.27
$NHCH_3$	−0.7	NH_2	−0.16	$NHCH_3$	−0.43	F	0.43	3-Br, 4-CH_3	0.15
NH_2	−0.66	OCH_3	0.12	NH_2	−0.36	Br	0.45	3-CH_3, 4-OCH_3	−0.26
OCH_3	−0.27	CH_3	−0.07	OCH_3	−0.11	Cl	0.45	3-CH_3, 4-$N(CH_3)_2$	−0.30
CH_3	−0.17	CH_2CH_3	−0.07	CH_3	−0.12	CF_3	0.38	3-OCH_3, 4-OH	−0.33
CH_2CH_3	−0.15	$CHCH_2$	0.06	CH_2CH_3	−0.13	CN	0.58	3-NO_2, 4-NO_2	1.38
$CHCH_2$	−0.04	F	0.34	$CHCH_2$	0	NO_2	0.64	3-NO_2, 4-Br	0.83
F	0.06	CF_3	0.43	F	0.21	OH	0.28	3-NH_2, 4-CH_3	−0.21
SH	0.15	CN	0.56	SH	0.07	OCH_3	0.29	3-$N(CH_3)_2$, 4-CH_3	−0.18
Br	0.23	NO_2	0.71	Br	0.3			3-OCH_3, 5-OCH_3	0.05
Cl	0.23	SH	0.25	Cl	0.28			3-OH, 5-OH	0.16
CF_3	0.54	Br	0.39	CF_3	0.54			3-OH, 4-OCH_3, 5-NO_2	0.63
CN	0.66	Cl	0.37	CN	0.68				
NO_2	0.78			NO_2	0.8				

such substituted bicyclo acids (Table 1) to our set to further test the predictive power of QTMS for a variety of carboxylic acids. Since the COOH group and the varying substituent X are bonded to a bicyclooctane skeleton, rather than to a phenyl ring as in the benzoic acids, a 1–1 map between the bonds of the bicyclo and the benzoic acids cannot be set up in a straightforward manner. The simplest way to proceed is to focus on the bonds that the acids have in common and only include them in the topological description. Given that the common bonds should preferably contain the active center we only include the O–H, C=O, C–O and C–C bonds in the topological description. Similarly there is no difficulty in adding a further 13 poly-substituted benzoic acids to our set. Fig. 4 shows how the measured Hammett acidity constants correlate with the predicted ones for all 68 carboxylic acids including *p*-, *m*-benzoic acids, *p*-phenylacetic acids and trisubstituted benzoic acids. The range and spread of the data are clearly satisfactory and PLS yields a model of two latent variables with an r^2 value of 0.91, while q^2 is 0.90. The model is valid according to SIMCA's default randomisation criteria. In the corresponding VIP plot (which is not shown) the dominant PC describing the O–H protrudes well above the second representing the C–O bond, which, in turn, protrudes above the third PC with a VIP value less than one.

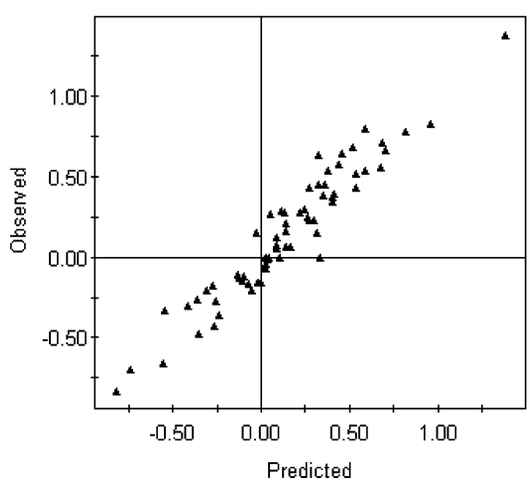

Fig. 4 Measured *versus* predicted Hammett acidity constants for all 68 carboxylic acids including *p*-, *m*-benzoic acids, *p*-phenylacetic acids and trisubstituted benzoic acids.

It should be noted that QTMS is able to model successfully carboxylic acids that have a different mode of action, as expressed by the different sets of σ constants.

All atoms in the *p*-benzoic acids were integrated. A standard PLS analysis with recommended values yields a model with two latent variables. This model does not pass the randomisation test but is validated when the model is restricted to only one LV. The corresponding r^2 and q^2 are 0.97 and 0.96, respectively. The PC with the highest VIP value is that describing H_1, the dissociated proton of the carboxy group, closely followed by O_4, the keto oxygen. Unfortunately the "active center" turns out be very diffuse involving ring carbons and ring hydrogens, as well as the expected O_2, in the fifth place.

It is remarkable that molecular information obtained in the gas phase can predict Hammett acidity constants, which are derived from acid–base equilibria and hence implicitly include solvation effects. The reason for this success is still obscure. The successful theoretical prediction of hydrogen bond donor capacity is an example of other work [18] that uses only gas-phase data to calculate solvation properties.

5 PCDDs

Given the recent prominence of QSARs in environmental toxicology we decided to apply QTMS to a well-known set of molecules known as polychlorinated dibenzo-*p*-dioxins (PCDDs). It is feasible to examine PCDDs by QTMS because of their modest size and molecular rigidity. Indeed, since virtually no work has been carried out on the influence of conformational flexibility on topological descriptors, lack of rigidity would currently pose a serious practical problem. PCDDs produce a wide span of toxic effects most of which involve binding to the aromatic hydrocarbon (Ah) receptor whose structure is unknown.

Fig. 5 shows the labeling scheme of the thirteen PCDDs under investigation marking the eight possible positions of chlorine substitution. Three types of toxicological data are provided in Table 2: the ability to bind to the cytosolic Ah receptor, to stimulate the induction of aryl hydrocarbon hydroxylase (AHH) and ethoxyresofurin *O*-deethylase (EROD). The

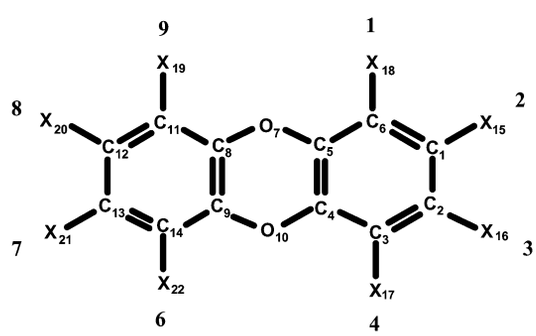

Fig. 5 Representation of polychlorinated dibenzo-*p*-dioxins shown with the numbered molecular skeleton where X = H or Cl. The eight bold numerical labels in the periphery mark the conventional numbering scheme for halogen substitution.

Table 2 Measured biological activities for a set of PCDDs

Name[a]	Compound[b]	pEC_{50}(bind)[c]	pEC_{50}(AHH)[c]	pEC_{50}(EROD)[c]
TCDD	2,3,7,8	8.000	9.721	10.143
PCDD2	1,2,3,7,8	7.102	7.770	7.959
PCDD3	2,3,6,7	6.796	7.959	7.215
PCDD4	2,3,6	6.658	–	–
PCDD5	1,2,3,4,7,8	6.553	8.387	8.678
PCDD6	1,3,7,8	6.102	6.495	6.229
PCDD7	1,2,4,7,8	5.959	7.959	7.678
PCDD8	1,2,3,4	5.886	5.620	5.432
PCDD9	2,3,7	7.149	6.854	6.444
PCDD10	2,8	5.495	4.000	4.000
PCDD11	1,2,3,4,7	5.194	6.086	6.180
PCDD12	1,2,4	4.886	5.658	4.319
PCDD14	1	4.000	4.000	4.000

[a] Following the nomenclature of Waller and McKinney.[48] [b] The numerical labels refer to the positions of chlorine substitution in the diagram of Fig. 5. [c] Measured values from ref. [51].

Table 3 Survey of PLS analysis obtained at various levels of theory and topological descriptors for pEC_{50}(bind) measured for a set of 13 PCDDs

Level	Descriptor	No. of LV orig.[a]	No. of LV valid.[b]	r^2	q^2
A	Bond length	3	2	0.84	0.68
B	BCP	4	1	0.74	0.57
	Atom	2	1	0.72	0.50
D	BCP	2	1	0.72	0.54
	Atom	3	1	0.72	0.46

[a] Original number of LVs obtained using the default LV significance criterion of SIMCA-P. [b] Maximum number of LVs that yield a model that passes the randomisation validation test.

Table 4 Survey of PLS analysis obtained at various levels of theory and topological descriptors for pEC_{50}(AHH) measured for a set of 13 PCDDs

Level	Descriptor	No. of LV orig.[a]	No. of LV valid.[b]	r^2	q^2
A	Bond length	1	1	0.40	0.20
B	BCP	1	1	0.73	0.45
	Atom		No model		
D	BCP	1	1	0.75	0.38
	Atom		No model		

[a] Original number of LVs obtained using the default LV significance criterion of SIMCA-P. [b] Maximum number of LVs that yield a model that passes the randomisation validation test.

Table 5 Survey of PLS analysis obtained at various levels of theory and topological descriptors for pEC_{50}(EROD) measured for a set of 13 PCDDs

Level	Descriptor	No. of LV orig.[a]	No. of LV valid.[b]	r^2	q^2
A	Bond length	1	1	0.42	0.25
B	BCP	1	1	0.74	0.52
	Atom	1	1	0.72	0.21
D	BCP	1	1	0.80	0.55
	Atom		No model		

[a] Original number of LVs obtained using the default LV significance criterion of SIMCA-P. [b] Maximum number of LVs that yield a model that passes the randomisation validation test.

activities are given as the negative of the base-ten logarithm of the concentration required to produce a response in a given time, respectively denoted by pEC_{50}(bind), pEC_{50}(AHH) and pEC_{50}(EROD). The geometries and wavefunctions were generated at levels A, B and D. The molecule PCDD13 (or OCDD), which has Cl substitution at all available sites, could not be included due to convergence problems.

Tables 3–5 summarise the PLS analyses of the three (measured) response variables *versus* the BCP and atomic properties generated at the three levels of theory. In each case we list the number of LVs obtained with the default cut-off criterion of SIMCA-P. If this model did not pass the randomisation test then the number of LVs was reduced until the model did. The r^2 and q^2 values of those models were then reported. Note that some entries in the tables do not state a model at all because no significant LV could be found (*i.e.* with an individual q^2 of at least 0.097). Overall the pEC_{50}(bind) response produced the most predictive models. This observation is in line with the results of a CoMFA study,[48] which reports q^2 values of 0.72, 0.42 and 0.23 for pEC_{50}(bind), pEC_{50}(AHH) and pEC_{50}(EROD), respectively. The worse predictivity of the latter two activities is expected since they are related to the much more complicated biological response of enzyme induction.

It is gratifying to see that level A, which is a semi-empirical method many orders of magnitude faster than *ab initio* calculations, delivers the best pEC_{50}(bind) model incorporating only

bond lengths. However, the other pEC_{50} models clearly benefit from the *ab initio* data in combination with topological descriptors.

In order to compare our work with the CoMFA study of Waller and McKinney[48] we predict the binding affinity of two heptachlorodibenzo-*p*-dioxins, with substitution patterns 1,2,3,4,6,7,8 and 1,2,3,4,6,7,9 computed at level A. Measured binding data are not available but Waller and McKinney predicted binding affinities of 5.70 and 4.12 respectively, whereas we predict 6.38 and 6.42.

The three measured pEC_{50} activities are thought to be mediated by a common (Ah or dioxin) receptor mechanism of action.[48,49] Since the activity refers to ability to bind to a receptor, the active center is not involved in bond breaking as in the case of carboxylic acids. The colour-coded molecular skeletons in Fig. 6 show the VIP of the PCs associated with

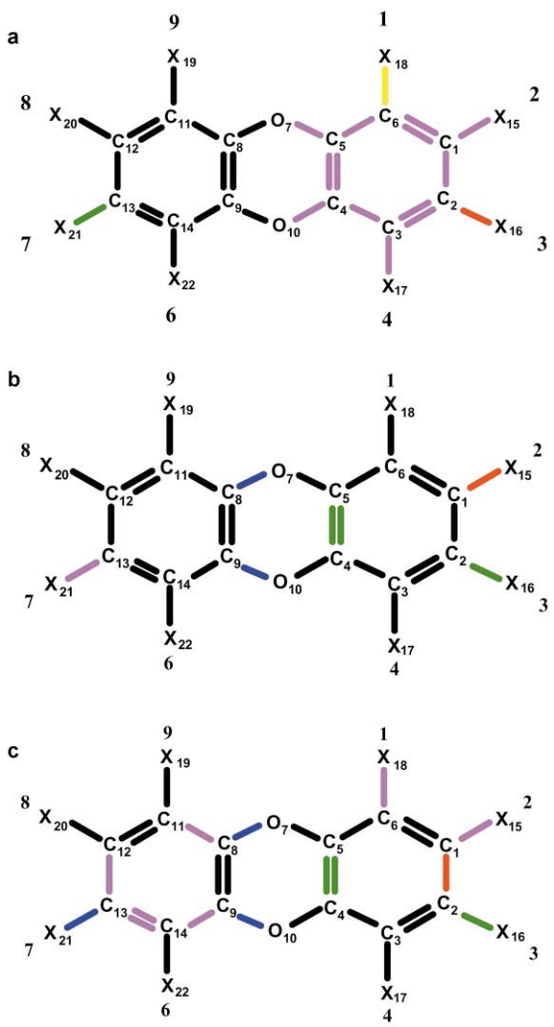

Fig. 6 Colour-coded plot expressing the influence (VIP) of the bonds in explaining the observed acidity pEC_{50}(bind) of the PCDDs at (a) level A, (b) level B and (c) level D. The active site has the highest influence (red). Legend: red > yellow > green > blue > purple > black (see text).

each bond for pEC_{50}(bind) at levels A, B and D. Before we interpret these plots we should be aware of the full consequence of the numbering convention adopted in work on PCDDs. This convention fixes the molecule in space and assigns absolute meaning to left and right, or up and down. For example, according to Table 2 PCDD10 is disubstituted at positions 2 and 8, which are "up". After rotation around the C_2 axis, lying in the molecular plane in the long direction of the molecule, the two Cls end up in positions 3 and 7. Of course, the rotated molecule is identical to the first one but appears to be different in view of the absolute character of the numbering convention. Only the *relative* configuration of a substitution pattern is relevant, *i.e.* how the Cls are positioned with respect to each other and the oxygens. It is meaningful to distinguish the positions 2,3,7 and 8 as *lateral*, because regardless of whether the position is left, right, up or down it is at the outside of the molecule (*i.e.* most remote from the oxygens). Equally, positions 1,4,6 and 9 are called *peri*,[50] which is another meaningful term referring to the relative configuration.

The way the set of PCDDs is expressed in terms of the absolute (or fixed space) convention has a substitution bias to the right. In other words, in total there are 33 Cls at the right and 16 Cls at the left. This means that, as a consequence of an arbitrary choice, the right-hand side is substituted more heavily. However, there is no bias between the up and down side (25 Cls *versus* 24 Cls, respectively).

It is clear from Fig. 6 that the lateral bonds on the right have the highest VIP values. There are variations depending on the level of calculation, but we invariably find that the red bonds are on the right, and in the lateral region. Since the right-hand side is most heavily substituted this means that the presence of Cls at the lateral side influences the activity most. Of the lesser contributors the yellow C_6–X_{18} bond in Fig. 6(a) could be regarded as a contamination. Secondly a change in computation level (A, B or D) alters the red bond between up and down. However, we cannot infer any information from a difference in bond highlighting between the up and down side, because they are equally substituted.

Fig. 7 shows a color-coded molecular skeleton for pEC_{50}-

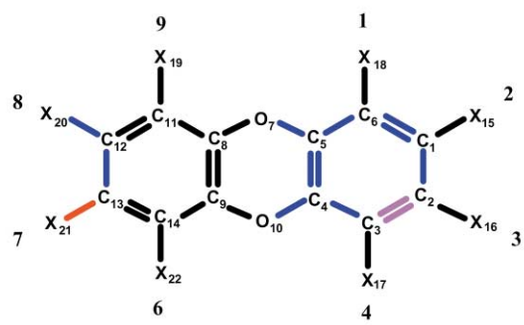

Fig. 7 Colour-coded plot expressing the influence (VIP) of the bonds in explaining the observed acidity pEC_{50}(AHH) of the PCDDs at level A. The active site has the highest influence (red). Legend: red > yellow > green > blue > purple > black (see text).

(AHH) computed at level A where the highlighted bond (red, C_{13}–X_{21}) is prominent (*i.e.* sharp decline in the VIP profile) and appears at the left. Interestingly, this means that under our current QTMS working hypothesis the lateral positions are again most relevant to explain the observed activity, but only provided the number of Cls substitutions is low.

Our main conclusion about the importance of the lateral side overlaps with the deductions made by Bonati *et al.* in their electrostatic (MEP) study[49] of the enzyme–substrate recognition step. They claim that electronic polarization along the principal axis leads to a strong charge concentration at both the lateral positions, where it is available for interaction with receptor electrophilic sites, and to a charge depletion over the oxygen regions, where an electron donor site can act favorably. Our results complement the CoMFA work of Waller and McKinney in which they point out the importance of lateral halogen substitution on overall stereo-electronic structure. They claim that the molecular polarisability is most affected by lateral halogens. We feel that further research into the convergence of results on these systems from different analyses including QTMS is desirable.

6 Conclusions

We have further developed QTMS along previously suggested lines.[29] We showed that excellent regression statistics are obtained for the acidity of carboxylic acids with a more relaxed definition of the common molecular skeleton. We have applied QTMS to an ecologically relevant set of molecules (PCDDs) and shown that it competes in predictive quality with CoMFA and a study based on the molecular electrostatic potential. Using convenient colour plots we recover the COOH group as the active center of the carboxylic acids and confirm that Cl-substitution at the lateral side in the PCDDs is highly relevant to explain variations in binding affinity. For the current systems (topological) atomic properties do not seem to add much to the predictive power or interpretation of a model solely based on computationally cheaper BCP properties. More work is needed on the sharpening of the active center. This could be accomplished by modified regression paths or by completely alternative techniques, such as neural networks or genetic algorithms.

Acknowledgements

We thank Dr S. E. O'Brien for generating wavefunctions for the carboxylic acids and PCDDs, and for his preliminary work on the PCDDs.[37]

References

1 A. Crum-Brown and T. R. Fraser, *Trans. R. Soc. Edinburgh*, 1868, **25**, 151.
2 J. Damborsky and T. W. Schultz, *Chemosphere*, 1997, **34**, 429.
3 C. Hansch and A. Leo, *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, ed. ACS, 1995.
4 C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 1964, **86**, 1616.
5 F. J. Luque, F. Sanz, F. Illas, R. Pouplana and Y. G. Smyers, *Eur. J. Med. Chem.*, 1988, **23**, 7.
6 R. Carbo, E. Besalu, L. Amat and X. Fradera, *J. Math. Chem.*, 1995, **18**, 237.
7 L. Amat, D. Robert, E. Besalu and R. Carbo-Dorca, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 624.
8 R. Carbo-Dorca, D. Robert, L. Amat, X. Girones and E. Besalu, *Molecular Similarity in QSAR and Drug Design*, Springer, Berlin, 2000.
9 R. Ponec, L. Amat and R. Carbo-Dorca, *J. Comput.-Aided Mol. Des.*, 1999, **13**, 259.
10 M. Karelson, *Molecular Descriptors in QSAR/QSPR*, Wiley-Interscience, New York, 2000.
11 P. L. A. Popelier, in *Molecular Similarity and complementarity based on the theory of atoms in molecules*, ed. P. M. Dean, Blackie Academic, London, 1995.
12 R. F. W. Bader, *Atoms in Molecules. A Quantum Theory*, Oxford University Press, Oxford, 1990.
13 P. L. A. Popelier, *Atoms in Molecules. An Introduction*, Pearson Education, London, 2000.
14 P. L. A. Popelier, F. M. Aicken and S. E. O'Brien, *Chem. Modell.*, 2000, **1**, 143.
15 P. L. A. Popelier and P. J. Smith, *Chem. Modell.*, 2002, **2**, 391.
16 R. J. Gillespie and P. L. A. Popelier, *Chemical Bonding and Molecular Geometry from Lewis to Electron Densities*, Oxford University Press, New York, 2001.
17 S. T. Howard and T. M. Krygowski, *Can. J. Chem.*, 1997, **75**, 1174.
18 J. A. Platts, *Phys. Chem. Chem. Phys.*, 2000, **2**, 973.
19 P. L. A. Popelier, *J. Phys. Chem. A*, 1999, **103**, 2883.
20 S. E. O'Brien and P. L. A. Popelier, *ECCOMAS conference proceedings*, Barcelona, Spain, 2000.
21 B. K. Alsberg, N. Marchand-Geneste and R. D. King, *Chemom. Intell. Lab. Syst.*, 2000, **54**, 75.
22 H. Wiener, *J. Chem. Phys.*, 1947, **69**, 17.
23 M. Randic, *J. Am. Chem. Soc.*, 1975, **97**, 6609.
24 M. Randic and J. Zupan, *J. Chem. Inf. Comput. Sci*, 2001, **41**, 550.
25 A. T. Balaban, *J. Chem. Inf. Comput. Sci*, 1997, **37**, 645.
26 I. Gutman, H. Hosoya and D. Babic, *J. Chem. Soc., Faraday Trans.*, 1996, **92**, 625.
27 L. B. Kier and L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976.
28 L. B. Kier, *Molecular structure description: the electrotopological state*, Academic, San Diego, CA, US, 1999.
29 S. E. O'Brien and P. L. A. Popelier, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 764.
30 S. E. O'Brien and P. L. A. Popelier, *J. Chem. Soc., Perkin Trans. 2*, 2002, 478.
31 Since we use modern DFT (*i.e.* B3LYP) wavefunctions are expressed *via* Kohn–Sham orbitals. Hence the quoted kinetic energy density refers only to the non-interacting reference system (which lacks Coulomb correlation). Part of the "true" kinetic energy is absorbed in the exchange-correlation functional. This part is not reported here since we restrict ourselves to the expression for $K(\boldsymbol{r})$ given in the text.
32 S. E. O'Brien and P. L. A. Popelier, *Can. J. Chem.*, 1999, **77**, 28.
33 S. Wold, M. Sjostrom and L. Eriksson, in *Partial Least Squares Projections to Latent Structures (PLS) in Chemistry*, ed. P. von Ragué Schleyer, New York, 1998.
34 G. Schaftenaar and J. H. Noordik, *J. Comput.-Aided Mol. Des.*, 2000, **14**, 123.
35 GAUSSIAN98, Gaussian 98, Revision A.7, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle and J. A. Pople, Gaussian, Inc., Pittsburgh PA, USA, 1998.
36 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 902.
37 S. E. O'Brien, Thesis, Quantum Molecular Similarity, An Atoms in Molecules Approach, UMIST, Manchester, 2000.
38 J. B. Foresman and A. Frisch, *Exploring Chemistry with Electronic Structure Methods*, Gaussian Inc., 1996.
39 MORPHY98, a program written by P. L. A. Popelier with a contribution from R. G. A. Bone, UMIST, Manchester, England, 1998.
40 P. L. A. Popelier, *Chem. Phys. Lett.*, 1994, **228**, 160.
41 P. L. A. Popelier, *Mol. Phys.*, 1996, **87**, 1169.
42 P. L. A. Popelier, *Comput. Phys. Commun.*, 1998, **108**, 180.
43 UMETRICS, SIMCA-P 8.0, http://www.umetrics.com, 1998 Umea.
44 UMETRICS, SIMCA-P 8.0 User Guide and Tutorial, Umea, 1999.
45 L. Livingstone, *Data Analysis for Chemists*, 1st edn., Oxford University Press, Oxford, 1995.
46 SPSS Inc., version 10.0.7, http://www.spss.com, Chicago, US, 2000.
47 C. Hansch, A. Leo and D. Hoekman, *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*, ed. S. R. Heller, American Chemical Society, Washington, 1995.
48 C. L. Waller and J. D. McKinney, *J. Med. Chem.*, 1992, **35**, 3660.
49 L. Bonati, E. Fraschini, M. Lasagni and D. Pitea, *J. Mol. Struct. (THEOCHEM)*, 1994, **303**, 43.
50 J. Damborsky, M. Lynam and M. Kuty, in *Structure-Biodegradability Relationships for Chlorinated Dibenzo-p-Dioxins and Dibenzofurans*, ed. R.-M. Wittich, R. G. Landes, Austin, TX, USA, 1998.
51 S. Safe, *Annu. Rev. Pharmacol. Toxicol.*, 1986, **26**, 371.