Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry Letters 18 (2008) 2604–2609
Bioorganic & Medicinal Chemistry Letters

Exploring predictive QSAR models for hepatocyte toxicity of phenols using QTMS descriptors

Kunal Roy*,† and Paul L. A. Popelier*

Manchester Interdisciplinary Biocenter (MIB), 131 Princess Street, Manchester, M1 7DN, United Kingdom

Received 5 February 2008; revised 12 March 2008; accepted 12 March 2008
Available online 16 March 2008

Abstract—We construct predictive QSAR models for hepatocyte toxicity data of phenols using Quantum Topological Molecular Similarity (QTMS) descriptors along with hydrophobicity ($\log P$) as predictor variables. The QTMS descriptors were calculated at different levels of theory including AM1, HF/3-21G(d), HF/6-31G(d), B3LYP/6-31+G(d,p), B3LYP/6-311+G(2d,p) and MP2/6-311+G(2d,p). The external predictability of the best models at the higher levels of theory is higher than that at the lower levels. Moreover, the best QTMS models are better in external predictability than the PLS models using pK_a and Hammett σ^+ along with $\log P$. The current study implies the advantage of quantum chemically derived descriptors over physicochemical (experimentally derived or tabular) electronic descriptors in QSAR studies.

Phenols are widely distributed in edible plants and found in tea, fruits and vegetables.[1,2] Phenols are also used in many industries as intermediates or as biocides themselves.[3] They occur as industrial wastes and being relatively soluble in water and detectable in rivers, ponds, and soil, they are direct pollutants in the environment. They can also be produced by environmental degradation of more complex molecules.[4] Phenols are supposed to be the most toxic water pollutant as they are carcinogenic in nature.[5,6] Because of their prevalence in the environment, human exposure to phenols is ubiquitous, and thus there is much interest in determining their potential hazard. Although the ability of polyphenols to protect cell from the oxidative stress has been demonstrated, there is increasing evidence of their prooxidant cytotoxicity.[7] The toxicity of phenols is confusing as the same polyphenol compounds can behave both as antioxidants and prooxidants, depending on concentration and free radical source.[8–11] Exposure of mammalian cells to polyphenols is accompanied by an increase in intracellular reactive oxygen species levels.[12] The mechanism underlying the toxicity of phenols is mainly related to lipophilicity and electrophilic effects.[13–15]

Phenols are important in nutrition and medicine given their cytotoxic potential. Second, they are prevalent in the environment and are likely to elicit often unknown ecotoxic effects. Hence, there is much interest recently in applying quantitative structure–activity relationships (QSARs) to predict the toxic potential of phenolic compounds. Liu et al. have explored a QSAR for toxicity of chlorophenols on L929 cells.[16] Padmanavan et al. have used group philicity and electrophilicity as possible descriptors for modelling ecotoxicity applied to chlorophenols.[3] Rule-based ensemble modelling was applied to develop a model with predictive capability for distinguishing between four different modes of toxic action for a set of 220 phenols by Norinder et al.[17] A three-dimensional QSAR study has been reported for in-vitro toxicity of chlorophenols to HepG2 cells.[18] Aptula et al. have modelled toxicity of di- and trihydroxybenzenes to *Tetrahymena pyriformis* using hydrophobilicity and electrophilicity indices.[15] Non-linear QSAR modelling of the toxicity of phenol derivatives to *Tetrahymena pyriformis* has been reported by Devillers.[19] Schüürmann et al. reported stepwise discrimination between four modes of toxic action of phenols in the *Tetrahymena pyriformis* assay.[20] Determination of mechanisms of toxic action of phenols to *Tetrahymena pyriformis* by QSAR

Keywords: QTMS; Toxicity; Ab initio; Phenols; QSAR; External validation; Electron density; Atoms in molecules; Quantum chemical topology.
* Corresponding authors. Tel.: +91 98315 94140; fax: +91 33 2837 1078 (K.R.); tel.: +44 161 3064511; fax: +44 161 3065201 (P.L.A.P.); e-mail addresses: kroy@pharma.jdvu.ac.in; pla@manchester.ac.uk
† Currently on leave from Drug Theoretics and Cheminformatics Lab, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India.

0960-894X/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2008.03.035

methods has also been reported by Ren.[21] Wang et al. have reported mechanism-based QSARs for the inhibition of substituted phenols on germination rate of *Cucumis sativus*.[22]

Moridani et al.[23] have reported quantitative structure–toxicity relationships of phenols in isolated rat hepatocytes using $\log P$, pK_a and Brown's variation of Hammett electronic constant (σ^+). Physicochemical electronic descriptors (pK_a and σ^+) are important in this data set, which prompts the use of quantum topological molecular similarity (QTMS) indices since they are known to be successful.[24–29] The data set of Moridani et al.[23] serves as a data source for the present work. The objective is to obtain models with improved predictive ability and better linked to modern ab initio wavefunctions. The details of QTMS descriptors can be found in the previous publications.[24–29]

To start with, an estimated geometry was obtained using the program GaussView,[30] which was then passed on to the ab initio program GAUSSIAN03.[31] We used, in succession, AM1, HF/3-21G(d), HF/6-31G(d), B3LYP/6-31+G(d,p), B3LYP/6-311+G(2d,p) and MP2/6-311+G(2d,p), passing on the optimised geometry of each level as a starting geometry for the next. Since the AM1 level is unable to produce a sensible topology, only bond lengths were retrieved from it. Second, the wavefunction was read by a local version of the program MORPHY98,[32] which locates the BCPs using an automatic and robust algorithm.[33] The BCP descriptors of eight common bonds of the phenolic compounds (six C–C aromatic bonds, one C–O bond and one O–H bond) were considered as variables for the statistical model development. Thirdly, the program SIMCA[34] was used for partial least squares (PLS) analysis of the data set. PLS is a generalization of regression, which can handle data with strongly correlated and/or noisy or numerous independent variables.[35] It gives a reduced solution, which is statistically more robust than Multiple Linear Regression (MLR). The linear PLS model finds 'new variables' (latent variables (LV) or independent scores) that are linear combinations of the original variables. To avoid overfitting, a strict test for the significance of each consecutive LV is necessary and no new LVs are added when they become non-significant. Cross validation is a practical and reliable method of testing this significance.[36] For the development of the PLS models, a hierarchical method was adopted. Initially, for each level (except AM1), PLS models were developed for each category of descriptors, that is, ρ, $\nabla^2\rho$, ε, λ, K, G, and equilibrium bond lengths. Note that λ_1, λ_2 and λ_3 are lumped together into the class of λ. There are $3 \times 8 = 24$ descriptors in the class of λ while in other classes there are only eight. At the outset, models were tried with all available descriptors, but subsequently, descriptors with smaller VIP (variable importance for the projection) values were gradually deleted until a model with the best Q^2 (leave-one-seventh-out cross validation) was obtained. Then, using important descriptors appearing in the PLS equations of different descriptor classes, the PLS model for the combined set of descriptors was developed.

The main target of any QSAR modelling is that the developed model should be robust enough to be capable of making accurate and reliable predictions of biological activities of new compounds.[37–39] So, QSAR models that are developed from a training set should be validated using new chemical entities for checking the predictive capacity of the developed models. The validation strategies check the reliability of the developed models for their possible application on a new set of data, and confidence of prediction can thus be judged.[40] In many cases, enough new chemicals being unavailable for prediction purpose, the original data set is divided into a training set and a test set. For the present work, the compounds were ranked according to the toxicity values and every fourth compound was selected as a test set compound. Preliminary modelling showed three compounds as outliers. They were reported to have a different mode of action in the original paper.[23] These three compounds are catechol, hydroquinone and 2-nitrophenol, which are also reported to be outliers in other studies.[19,41] Thus, these compounds were not considered in the present study.

For external validation, a predictive coefficient R^2_{pred} was calculated via Eq. 1,

$$R^2_{\text{pred}} = 1 - \frac{\sum(Y_{\text{obs}} - Y_{\text{pred}})^2}{\sum(Y_{\text{obs}} - \overline{Y}_{\text{Training}})^2} \tag{1}$$

where Y_{obs} and Y_{pred}, respectively, represent the observed and predicted toxicity values of the test set compounds, while $\overline{Y}_{\text{Training}}$ represents the mean observed value of the training set. The R^2_{pred} value is in part controlled by the magnitude of $(Y_{\text{test}} - \overline{Y}_{\text{training}})^2$. The squared regression coefficient (r^2) between observed and predicted values of the test set compounds does not necessarily indicate that the predicted values are very near to observed activity values (there may be considerable numerical differences between the values in spite of maintaining a good overall intercorrelation). To better indicate the external predictive capacity of a model, a modified r^2 term (r^2_{m}), defined before,[42] is given in Eq. 2,

$$r^2_m = r^2\left(1 - \sqrt{r^2 - r^2_0}\right) \tag{2}$$

The magnitude r^2_0 in Eq. 2 represents the squared correlation coefficient between the observed and predicted values of the test set compounds setting the intercept to zero.[43] Note that r^2 is always larger than r^2_0. In case of good external prediction, predicted values will be very close to observed activity values. So, the r^2 value will be very near to the r^2_o value. In the best case, r^2_{m} will be equal to r^2 whereas in the worst case r^2_{m} value will be zero.

For all the developed models we have reported the coefficient of variation (R^2), leave-one-seventh-out cross-validation R^2 (Q^2) for the training set and the R^2_{pred}, r^2, r^2_0 and r^2_{m} values for the test set. The final models were also subjected to a randomisation test. In this test, the toxicity data (Y) are randomly permuted keeping the descriptor matrix intact, followed by a PLS run. Each

randomisation and subsequent PLS analysis generates a new set of R^2 and Q^2 values, which are plotted against the correlation coefficient between the original Y values and the permuted Y values. The intercepts for the R^2 and Q^2 lines in this plot are a measure of the overfit. A model is considered[44] valid if $R^2_{int} < 0.4$ and $Q^2_{int} < 0.05$.

Moridani et al.[23] employed MLR to model the data set (Table S1 in Supplementary Data) using $\log P$, and pK_a and σ^+ as the electronic descriptors. They tried bond dissociation energy and a redox potential as additional electronic descriptors, but the data for these descriptors are not available for all the compounds presented in their paper. Furthermore, the QSAR relations with these two descriptors were inferior[23] to those containing pK_a and σ^+. Hence, we will compare our results only to the QSAR models involving pK_a and σ^+ in addition to $\log P$.

Here we compare our models to those of Moridani et al.[23] As QTMS does not encode for hydrophobicity we have retained $\log P$ in all our models. Since hydrophobicity is an important descriptor for hepatocyte toxicity,[23] $\log P$ appears as a significant contributor in all the models. Our objective is increasing the predictive potential of the models using QTMS descriptors instead of σ^+ and pK_a. Since we used PLS rather than MLR and since we divided the data set into a training and a test set, we cannot directly compare our results to those of Moridani et al.[23]

A preliminary PLS model of the data matrix composed of $\log P$, pK_a and σ^+ led to a model with one LV showing a R^2 value of 0.565 and a Q^2 value of 0.330. The PLS score plot in Figure S1 (in Supplementary Data) clearly shows three outliers: 2-nitrophenol, catechol and hydroquinone. These compounds were not considered in the rest of the analysis.

Table 1 summarizes all PLS models. We first discuss the three models with only physicochemical electronic descriptors. Model 1 was developed from $\log P$ and pK_a using one LV. This gave a Q^2 value of 0.696 while the R^2 value for the training set was 0.759. When this model was used to predict the toxicity values of the test set compounds, the R^2_{pred} value was found to be −0.345 and the r^2_m value was 0.018. Clearly, although model 1 passes the internal cross-validation test (since $Q^2 > 0.5$), the predictive capacity of the model based on the test set is practically nil. Model 2 involves $\log P$ and σ^+, leading to a Q^2 value of 0.823 and a R^2 value of 0.864. The predictive capacity of the model was found to be acceptable ($R^2_{pred} = 0.616$, $r^2_m = 0.544$). The third model used all three descriptors ($\log P$, pK_a and σ^+). This one-LV model showed results ($Q^2 = 0.782$, $R^2 = 0.850$, $R^2_{pred} = 0.409$, $r^2_m = 0.391$) inferior to model 2. Clearly, using the physicochemical descriptors, model 2 is the best one showing acceptable internal and external validation parameters.

Now we discuss the models containing QTMS descriptors, systematically increasing the level of theory. At AM1 level, only bond distances were considered as descriptors. At the start all eight bond lengths along with $\log P$ were considered as descriptors, followed by deletion of less significant descriptors based on VIP values, eventually resulting in model 4. A two-LV model (model 4) with three descriptors (one of them being $\log P$) showed a R^2 value of 0.832 and Q^2 value of 0.787 for the training set. When the model was used to predict the toxicity values of the test set compounds, the R^2_{pred} and r^2_m values were found to be 0.511 and 0.479, respectively. Clearly, model 4 is inferior to model 2. Hence we proceed to the higher levels of theory.

At HF/3-21G(d) level we obtain eight models, one of which was developed from the pool of all descriptors. Based on both internal validation and external validation characteristics, the best model was derived from bond lengths (model 5). Model 5 with six descriptors and three LVs showed R^2 and Q^2 values of 0.890 and 0.808, respectively, for the training set, and R^2_{pred} and r^2_m values of 0.667 and 0.570, respectively, for the test set. Clearly, this model is better in both internal and external validation characteristics than model 2, which is based on σ^+. The next best model was based on the ellipticity descriptor (model 8). This model shows R^2 and Q^2 values of 0.790 and 0.780, respectively, for the training set and R^2_{pred} and r^2_m values of 0.710 and 0.568, respectively, for the test set and thus this model is inferior to both model 2 and model 5 (however, r^2_m value of model 8 is better than those of model 2). While considering all descriptor types, model 12 containing six descriptors and three components gave acceptable internal validation characteristics ($Q^2 = 0.756$), but the external validation parameters are miserably poor.

At the next level, HF/6-31G(d), we again derived eight models (models 13–20), as is the case for all higher levels. The best model at this level came from the pool of all descriptors (model 20). Model 20 with eleven descriptors and four LVs gave excellent R^2 (0.911) and Q^2 (0.815) values while the external validation parameters were very good ($R^2_{pred} = 0.730$, $r^2_m = 0.796$). This model was better than models 2 and 5 in both internal and external validation characteristics. Only considering models built from descriptors from a single class, model 18 with seven K descriptors ($R^2 = 0.879$, $Q^2 = 0.791$) closely follows model 20. The R^2_{pred} value of model 18 is slightly higher than that of model 20; however, the r^2_m value (0.734) of model 18 is lower than that of model 20. In search of models for higher predictive capacity (external predictability), BCP properties were calculated at still higher levels.

Models 21–28 were obtained at B3LYP/6-31+G(d,p) level. Based on the external validation characteristics, model 22 with four descriptors and two LVs showed a predictive R^2 value of 0.774 and r^2_m value of 0.786. These values are marginally lower than the corresponding values of model 20 obtained at the HF/6-31G(d) level. The R^2 and Q^2 values for training set for model 22 are 0.867 and 0.790, respectively. Another model (model 26) with four K descriptor shows R^2_{pred} and r^2_m values lower than those of model 22, but the internal validation character-

Table 1. Comparative analysis of QSAR models based on hierarchical PLS for hepatocyte toxicity data

Level of theory	Model no.	Type of descriptors in addition to log P	No. of descriptors	LV[a]	R^2	Q^2	R^2_{pred}	r^2	r^2_0	r^2_m
MP2/6-311+G(2d,p)	44	All	4	2	0.847	0.780	0.525	0.561	0.531	0.464
	43	**G**	**8**	**2**	**0.869**	**0.797**	**0.889**	**0.891**	**0.890**	**0.863**
	42	**K**	**3**	**2**	**0.874**	**0.839**	**0.733**	**0.734**	**0.732**	**0.701**
	41	λ[b]	3	2	0.814	0.612	−0.736	0.059	−0.543	0.013
	40	ε	6	2	0.831	0.738	0.287	0.369	0.348	0.316
	39	∇^2_ρ	4	2	0.838	0.752	0.593	0.596	0.592	0.558
	38	$\boldsymbol{\rho}$	3	2	0.860	0.816	0.761	0.780	0.754	0.654
	37	Distance	3	2	0.843	0.803	0.474	0.497	0.481	0.434
B3LYP/6-311+G(2d,p)	36	All	4	2	0.788	0.740	−0.252	0.085	−0.176	0.042
	35	G	7	2	0.886	0.857	0.810	0.878	0.842	0.711
	34	**K**	**7**	**3**	**0.886**	**0.768**	**0.814**	**0.821**	**0.820**	**0.795**
	33	λ	13	3	0.856	0.634	0.687	0.679	0.677	0.649
	32	ε	5	2	0.841	0.749	0.009	0.156	0.084	0.114
	31	∇^2_ρ	4	2	0.833	0.721	−0.704	0.012	−0.495	0.003
	30	$\boldsymbol{\rho}$	**6**	**3**	**0.889**	**0.798**	**0.744**	**0.735**	**0.734**	**0.712**
	29	Distance	4	2	0.871	0.829	0.250	0.831	0.245	0.195
B3LYP/6-31+G(d,p)	28	All	7	4	0.908	0.800	0.776	0.795	0.774	0.680
	27	G	4	2	0.859	0.817	0.583	0.576	0.564	0.513
	26	**K**	**4**	**3**	**0.897**	**0.809**	**0.706**	**0.698**	**0.698**	**0.698**
	25	λ	4	2	0.883	0.824	0.547	0.542	0.528	0.478
	24	ε	2	1	0.733	0.694	0.065	0.124	−0.128	0.062
	23	∇^2_ρ	2	1	0.728	0.701	−0.749	0.000	−0.662	0.000
	22	$\boldsymbol{\rho}$	**4**	**2**	**0.867**	**0.790**	**0.774**	**0.786**	**0.786**	**0.786**
	21	Distance	5	2	0.908	0.834	0.376	0.733	0.414	0.319
HF/6-31G(d)	**20**	**All**	**11**	**4**	**0.911**	**0.815**	**0.730**	**0.796**	**0.796**	**0.796**
	19	G	5	2	0.811	0.747	0.313	0.292	0.291	0.283
	18	**K**	**7**	**3**	**0.879**	**0.791**	**0.733**	**0.734**	**0.734**	**0.734**
	17	λ	9	3	0.911	0.830	0.670	0.764	0.753	0.674
	16	ε	2	1	0.805	0.800	0.665	0.697	0.697	0.697
	15	∇^2_ρ	4	3	0.842	0.765	0.570	0.570	0.568	0.545
	14	ρ	5	3	0.898	0.816	0.515	0.501	0.499	0.479
	13	Distance	4	2	0.834	0.769	0.675	0.780	0.737	0.618
HF/3-21G(d)	12	All	6	3	0.828	0.756	0.128	0.162	0.106	0.124
	11	G	3	2	0.816	0.767	0.191	0.204	0.106	0.164
	10	K	7	3	0.855	0.771	0.386	0.357	0.357	0.357
	9	λ	3	2	0.795	0.735	−0.330	0.041	−0.241	0.019
	8	$\boldsymbol{\varepsilon}$	**2**	**1**	**0.790**	**0.780**	**0.644**	**0.710**	**0.670**	**0.568**
	7	∇^2_ρ	5	1	0.787	0.733	−10.42	0.023	−0.264	0.011
	6	ρ	7	3	0.908	0.809	0.434	0.702	0.481	0.372
	5	**Distance**	**6**	**3**	**0.890**	**0.808**	**0.667**	**0.726**	**0.680**	**0.570**
AM1	**4**	**Distance**	**3**	**2**	**0.832**	**0.787**	**0.511**	**0.495**	**0.494**	**0.479**
Physicochemical	3	$\sigma^+ + pK_a$	3	1	0.850	0.782	0.410	0.409	0.407	0.391
	2	$\boldsymbol{\sigma^+}$	**2**	**1**	**0.864**	**0.823**	**0.616**	**0.625**	**0.608**	**0.544**
	1	pK_a	2	1	0.759	0.696	−0.345	0.039	−0.263	0.018

Bold models indicate the best two models (for each level of theory) according to r^2_m values.
[a] Number of latent variables.
[b] λ_1, λ_2 and λ_3.

istics ($Q^2 = 0.809$) of model 26 are marginally higher than model 22. It is important to note that the best model at this level, based on internal validation, is model 21 with a Q^2 value of 0.834; however, the external predictive capacity of this model is miserable ($R^2_{pred} = 0.376$, $r^2_m = 0.319$). However, the squared correlation coefficient (r^2) between the observed and predicted values of the test set compounds shows a high value (0.733) signifying the redundancy of this (r^2) parameter. Model 28, which is based on the pool of all descriptors, shows the highest R^2 for the training set ($R^2 = 0.908$) and the Q^2 value is very near to that of models 22 and 26. The predictive R^2 value (0.776) of this model is also higher than that of models 22 and 26, but the r^2_m (0.680) value is lower than the corresponding values of models 22 and 26.

Models 29–36 were obtained at B3LYP/6-311+G(2d,p) level. Based on external validation, model 34 with seven K descriptors and three LVs is the best one at this level. The R^2_{pred} and r^2_m values of this model are 0.814 and 0.795, respectively. Obviously, this model is far better

in external predictability than model 2 using σ^+ as electronic descriptor. The internal validation parameter Q^2 of model 34 is also very encouraging (0.768). However, another model (model 35) based on G descriptors shows a higher Q^2 value (0.857) while showing comparatively lower R^2_{pred} and r^2_m values (0.810 and 0.711, respectively). Model 36 based on the pool of all descriptors shows an acceptable Q^2 value (0.740), but external predictability of the model is practically nil. The root mean square errors of prediction (RMSEP) of models 34 and 35 are 0.229 and 0.231 whereas root mean square errors of estimate (RMSEE) for the training set in both cases are 0.187.

At the MP2/6-311+G(2d,p) level, models 37–44 were obtained. For the models at this level, compound 6 had to be omitted from the training set because of RAM shortage to calculate the corresponding wavefunction. However, all seven test set compounds were used to validate the external predictability. Based on external validation, model 43 with eight G descriptors and two LVs shows R^2_{pred} and r^2_m values of 0.889 and 0.863, respectively, which are higher than those of any other model. The internal validation parameter (Q^2) of model 43 is also very good. However, another model (model 42) involving K descriptors shows a higher Q^2 value while showing lower R^2_{pred} and r^2_m values. The model 44 based on all descriptors shows an acceptable Q^2 value while the R^2_{pred} and r^2_m values are poor. The RMSEE values of models 42 and 43 are 0.198 and 0.202, respectively, while RMSEP values are 0.272 and 0.175, respectively.

Selected models were subjected to randomisation test with 100 permutations (default[34] is 20) in each case. For all the models tested, R^2_{int} values are less than 0.4 and Q^2_{int} values are less than 0.05 (Table S2 in Supplementary Data). This indicates that the models are not obtained by chance.

Discussing the data as a whole we can note the following. Hydrophobicity ($\log P$) being very important for hepatotoxicity of phenols,[23] the $\log P$ term appears in all the models developed. Among the electronic descriptors, pK_a and σ^+ have been used. Using physicochemical electronic descriptors, the best model was model 2 (with σ^+) showing 61.6% predicted variance (external) and 82.3% predicted variance (internal). As our objective was to replace physicochemical descriptors with computationally derived (i.e. quantum chemical) electronic descriptors to develop models with better external predictability, we developed models with QTMS descriptors calculated at different levels of theory starting from AM1 to MP2. Based on R^2_{pred}, r^2 and r^2_m as metrics to denote external predictability, the best model was model 43 with G descriptors at the highest level [MP2]. This model showed R^2_{pred}, r^2 and r^2_m values of 0.889, 0.891 and 0.863, respectively, which are considerably higher than the corresponding values of model 2 (0.616, 0.625 and 0.544, respectively). When internal validation characteristics were considered, the best model was model 35 obtained at B3LYP/6-311+G(2d,p) level. This model shows a Q^2 value of 0.857, which is slightly higher than the corresponding value of model 2 (0.823). It should be remembered that only external validation can indicate the true predictability of a model.

We conclude that replacing physicochemical electronic descriptors with computationally derived quantum chemical (QTMS) descriptors increases external predictability of QSAR models for hepatocyte toxicity data of phenols. Moreover, experimental values of pK_a or tabular value of σ^+ may not be available for all new query compounds. In such cases, computationally derived QTMS descriptors give an extra advantage. The predictability increases specially at the higher levels of calculations. This reconfirms the advantage of using quantum chemically derived descriptors instead of physicochemical (experimentally derived or tabular) electronic descriptors in QSAR studies.

Acknowledgments

We thank the Commonwealth Scholarship Commission in the UK for their financial support towards an Academic Staff Fellowship for one of us (K.R.). We thank Mr. Alex Harding for carrying out the MP2 calculations.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmcl.2008.03.035.

References and notes

1. Huang, M.-T.; Osawa, T.; Ho, C.-T.; Rosen, R. T.. In *Food Phytochemicals for Cancer Prevention*; American Chemical Society: Washington, DC, USA, 1994; vol. I.
2. Ho, C.-T.; Osawa, T.; Huang, M.-T.; Rosen, R. T.. In *Food Phytochemicals for Cancer Prevention*; American Chemical Society: Washington, DC, USA, 1994; vol. II.
3. Padmanabhan, J.; Parthasarathi, R.; Subramanian, V.; Chattaraj, P. K. *Chem. Res. Toxicol.* **2006**, *19*, 356.
4. Verschueren, K. *Handbook of Environmental Data on Organic Chemicals*, 3rd ed.; Van Nostrand Reinhold Company: New York, USA, 1996.
5. Della, G. M.; Monaco, P.; Pinto, G.; Pollio, A.; Previtera, L.; Temussi, F. *Bull. Environ. Contam. Toxicol.* **2001**, *67*, 352.
6. Garg, R.; Kapur, S.; Hansch, C. *Med. Res. Rev.* **2001**, *21*, 73.
7. Galati, G.; O'Brien, P. J. *Free Radical Biol. Med.* **2004**, *37*, 287.
8. Sergediene, E.; Jonsson, K.; Szymusiak, H.; Tyrakowska, B.; Rietjens, I. M. C. M.; Narimantas, C. *FEBS Lett.* **1999**, *462*, 392.
9. Cao, G.; Sofic, E.; Prior, R. L. *Free Radical Biol. Med.* **1997**, *22*, 749.
10. Day, B. W.; Tyurin, V. A.; Tyurina, Y. Y.; Liu, M.; Facey, J. A.; Carta, G.; Kisin, E. R.; Dubey, R. K.; Kagan, V. E. *Chem. Res. Toxicol.* **1999**, *12*, 28.
11. Filipe, P.; Haigle, J.; Silva, J. N.; Freitas, J.; Fernandes, A.; Maziere, J.-C.; Maziere, C.; Santus, R.; Morliere, P. *Eur. J. Biochem.* **2004**, *271*, 1991.

12. Nemeikaite-Ceniene, A.; Imbrasaite, A.; Sergediene, E.; Cenas, N.. *Arch. Biochem. Biophys.* **2005**, *441*, 182.
13. Hansch, C.; McKarns, S. C.; Smith, C. J.; Doolittle, D. J. *Chem. Biol. Interact.* **2000**, *127*, 61.
14. Mekapati, S. B.; Hansch, C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 956.
15. Aptula, A. O.; Roberts, D. W.; Cronin, M. T. D.; Schultz, T. W. *Chem. Res. Toxicol.* **2005**, *18*, 844.
16. Liu, X.; Chen, J.; Yu, H.; Zhao, J.; Giesy, J. P.; Wang, X. *Chemosphere* **2006**, *64*, 1619.
17. Norinder, U.; Lidén, P.; Boström, H. *Mol. Divers.* **2006**, *10*, 207.
18. Liu, Y.; Chen, J. N.; Zhao, J. S.; Yu, H. X.; Wang, X. D.; Jiang, J.; Jin, H. J.; Zhang, J. F.; Wang, L. S. *Chemosphere* **2005**, *60*, 791.
19. Devillers, J. *SAR QSAR Environ. Res.* **2004**, *15*, 237.
20. Schüürmann, G.; Aptula, A. O.; Kühne, R.; Ebert, R. U. *Chem. Res. Toxicol.* **2003**, *16*, 974.
21. Ren, S. *Environ. Toxicol.* **2002**, *17*, 119.
22. Wang, X.; Yu, J.; Wang, Y.; Wang, L. *Chemosphere* **2002**, *46*, 241.
23. Moridani, M. Y.; Siraki, A.; O'Brien, P. J. *Chem. Biol. Interact.* **2003**, *145*, 213.
24. Popelier, P. L. A. *J. Phys. Chem. A* **1999**, *103*, 2883.
25. O'Brien, S. E.; Popelier, P. L. A. *J. Chem. Inf. Comput. Sci.* **2001**, *41*.
26. Smith, P. J.; Popelier, P. L. A. *Org. Biomol. Chem.* **2005**, *3*, 3399.
27. Loader, R. J.; Singh, N. K.; O'Malley, P. J.; Popelier, P. L. A. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1249.
28. Singh, N.; Loader, R.; O'Malley, P. J.; Popelier, P. L. A. *J. Phys. Chem. A* **2006**, *110*, 6498.
29. Popelier, P. L. A.; Smith, P. J. *Eur. J. Med. Chem.* **2006**, *41*, 862.
30. GaussView3.0 Semichem Inc., Gaussian Inc., Pittsburgh, PA, USA, 2003.
31. GAUSSIAN03, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A. J.; Vreven, J., T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. In Gaussian, Inc., Pittsburgh PA, 2003.
32. MORPHY98, a program written by P.L.A. Popelier with a contribution from R.G.A. Bone, UMIST, Manchester, England, EU (1998).
33. Popelier, P. L. A. *Chem. Phys. Lett.* **1994**, *228*, 160.
34. UMETRICS *SIMCA-P 10.0*, info@umetrics.com: www.umetrics.com, Umeå, Sweden, 2002.
35. Wold, S.; Sjostrom, M.; Eriksson, L. *Chemo. Intell. Lab. Sys.* **2001**, *58*, 109.
36. Wold, S. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H. E., Ed.; VCH: Weinheim, Germany, 1995; p 195.
37. Roy, K. *Expert Opin. Drug Discov.* **2007**, *2*, 1567.
38. Leonard, J. T.; Roy, K., *QSAR Comb. Sci. 25*, 235.
39. Roy, P. P.; Leonard, J. T.; Roy, K. *Chemo. Intell. Lab. Sys.* **2008**, *90*, 31.
40. Roy, K.; Mandal, A. S., *J. Enzyme Inhib. Med. Chem.* DOI: doi:10.1080/14756360701811379 **2008**.
41. Cronin, M. T. D.; Aptula, A. O.; Duffy, J. C.; Netzeva, T. I.; Rowe, P. H.; Valkova, I. V.; Schultz, T. W. *Chemosphere* **2002**, *49*, 1201.
42. Roy, P. P.; Roy, K. *QSAR, Comb. Sci.* **2008**, *27*, 302.
43. Golbraikh, A.; Tropsha, A. *J. Molec. Graph Model.* **2002**, *20*, 269.
44. Wold, S.; Sjostrom, M.; Eriksson, L. Partial least squares projections to latent structures (PLS) in chemistry. In *Encyclopedia of Comparative Chemistry*; Schleyer, P. v. R., Ed.; Wiley: Chichester, GB, 1998; Vol. 3, p 2006.