

## Description of the Electronic Structure of Organic Chemicals Using Semiempirical and Ab Initio Methods for Development of Toxicological QSARs

Tatiana I. Netzeva,<sup>†</sup> Aynur O. Aptula,<sup>‡</sup> Emilio Benfenati,<sup>§</sup> Mark T. D. Cronin,<sup>\*,†</sup> Giuseppina Gini,<sup>||</sup> Iglia Lessigiarska,<sup>⊥</sup> Uko Maran,<sup>&</sup> Marjan Vračko,<sup>#</sup> and Gerrit Schüürmann<sup>‡</sup>

School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, England, Department of Chemical Ecotoxicology, UFZ Centre for Environmental Research, Leipzig-Halle, Permoserstrasse 15, D-04318, Germany, Department of Environmental Health Sciences, Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri, via Eritrea 62, 20157 Milan, Italy, Department of Electronics and Information, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy, European Centre for the Validation of Alternative Methods (ECVAM), Institute for Health and Consumer Protection, Joint Research Centre, European Commission, 21020 Ispra (VA), Italy, Department of Chemistry, Tartu University, 2 Jakobi Street, Tartu EE 2400, Estonia, and National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

Received August 16, 2004

The quality of quantitative structure–activity relationship (QSAR) models depends on the quality of their constitutive elements including the biological activity, statistical procedure applied, and the physicochemical and structural descriptors. The aim of this study was to assess the comparative use of ab initio and semiempirical quantum chemical calculations for the development of toxicological QSARs applied to a large and chemically diverse data set. A heterogeneous collection of 568 organic compounds with 96 h acute toxicity measured to the fish fathead minnow (*Pimephales promelas*) was utilized. A total of 162 descriptors were calculated using the semiempirical AM1 Hamiltonian, and 121 descriptors were compiled using an ab initio (B3LYP/6-31G\*\*) method. The QSARs were derived using multiple linear regression (MLR) and partial least squares (PLS) analyses. Statistically similar models were obtained using AM1 and B3LYP calculated descriptors supported by the use of the logarithm of the octanol–water partition coefficient ( $\log K_{ow}$ ). The main difference between the models derived by both MLR and PLS with the two sets of quantum chemical descriptors was concentrated on the type of descriptors selected. It was concluded that for large-scale predictions, irrespective of the mechanism of toxic action, the use of precise but time-consuming ab initio methods does not offer considerable advantage compared to the semiempirical calculations and could be avoided.

### INTRODUCTION

Under the European Commission's proposal for new legislation relating to the registration of new and existing chemical substances (<http://europa.eu.int/comm/enterprise/chemicals/chempol/whitepaper/reach.htm>) it is foreseen that quantitative structure–activity relationships (QSARs) for toxicity and fate will be used more extensively. Potential applications of the QSARs include priority setting and triggering of in vitro and/or in vivo testing, waiving of testing for hazard classification and for providing dose–response, environmental fate, and mechanistic information.<sup>1,2</sup> The development of reliable QSARs that can respond to the increased regulatory need requires critical scientific approach to all the three constitutive elements of the modeling process, i.e., the toxicological data, the chemical descriptors, and the statistical procedure for QSAR formulation.<sup>3</sup> In addition

approaches are required for the assessment, evaluation, and validation of QSARs.<sup>4,5</sup>

It has been recently recognized that there are acceptably good resources for the development of QSARs for acute toxicity to fish.<sup>6</sup> Among the available databases, the acute toxicity database to fathead minnow (*Pimephales promelas*) is considered a “gold standard” for quality of toxicity measurement.<sup>7</sup> The chemicals were selected for testing from the Toxic Substances Control Act (TSCA) inventory of chemicals to represent a cross-section of industrial organic chemicals.<sup>8</sup> Thus, the fathead minnow database is chemically heterogeneous, and the substances in it represent a large spectrum of mechanisms of toxic action.

A fundamental assumption in the current study was that any nonspecific acute toxicity to aquatic organisms is a result of penetration of the molecules into the cells and their reactivity toward vital cellular components.<sup>9</sup> The transport process in QSAR analysis is most frequently modeled by the logarithm of the octanol–water partition coefficient ( $\log K_{ow}$ ), and the quality of its calculation has been the subject of numerous discussions.<sup>10,11</sup> An assessment of chemical reactivity (i.e. to the readiness of the molecule to react covalently with the biomacromolecules) is also essential for

\* Corresponding author phone: + 44 151 231 2066; fax: + 44 151 231 2170; e-mail: m.t.cronin@livjm.ac.uk.

<sup>†</sup> Liverpool John Moores University.

<sup>‡</sup> UFZ Centre for Environmental Research.

<sup>§</sup> Istituto di Ricerche Farmacologiche Mario Negri.

<sup>||</sup> Politecnico di Milano.

<sup>⊥</sup> European Centre for the Validation of Alternative Methods (ECVAM).

<sup>&</sup> Tartu University.

<sup>#</sup> National Institute of Chemistry.

the prediction of toxicological hazard. Reactivity is related to the electronic structure and can be modeled with quantum chemical descriptors.<sup>12</sup>

The emphasis in the present study was placed on the quality of the electronic indices for the development of toxicological QSARs and, more specifically, on the level of theory required for general (i.e. across many classes and mechanisms) prediction of acute aquatic toxicity from hydrophobicity and quantum chemical descriptors. Previous studies have analyzed the sources of variability in the toxicological QSARs,<sup>13</sup> the variability of the electronic indices (particularly the energy of the lowest unoccupied molecular orbital ( $E_{\text{LUMO}}$ )) both between selected Hamiltonians and selected software package,<sup>14</sup> and the possible ways to parametrize electrophilicity in terms of descriptors within the same Hamiltonian and software package.<sup>15</sup>

The aim of this study, therefore, was to assess the use of ab initio and semiempirical quantum chemical calculations for the development of toxicological QSARs applied to a large and chemically diverse data set, irrespective of the mechanism of toxic action. To achieve this aim, the B3LYP/6-31G\*\* ab initio method was used to calculate descriptors for 568 compounds with acute toxicity data to fathead minnow. B3LYP was preferred as a hybrid functional of the density functional theory (DFT), which has become a standard method to include the dynamic electron correlation.<sup>16</sup> B3LYP has proved superior to several other DFT functions and is often considered competitive to high level ab initio schemes as regards the computation of molecular geometries, energies, and ionization potentials.<sup>17</sup> It was compared to the semiempirical method AM1,<sup>18</sup> which is much faster than the ab initio schemes and reproduces experimental data (heat of formation, dipole moment, and ionization potential) with reasonable accuracy. It also performs reasonably well in the calculation of the hydrogen-bond energies and charge distribution associated with electrostatic potentials,<sup>19</sup> which is important predominantly for charge-controlled (hard-hard) interactions.<sup>16</sup> However, the semiempirical quantum chemical methods have been developed to reproduce only a few (mainly gas-phase) properties and geometries, and the performance of these methods for calculation of descriptors related to charge distribution and site-specific molecular reactivity is not clear a priori. The suitability of the descriptors calculated by B3LYP/6-31G\*\* and AM1 was investigated by the development of QSARs for acute fish toxicity utilizing these descriptors following the application of multiple linear regression (MLR) and partial least squares (PLS) analyses. These two techniques are recognized as the most established methods for the development of QSARs.<sup>20</sup> Both MLR and PLS used robust procedures for the selection of variables to ensure the unprejudiced achievement of the goal of this study.

## METHODS

**Biological Data.** A total of 568 organic chemicals representing several chemical classes and mechanisms of toxic action were considered in this study. Their acute toxicity to the fish *Pimephales promelas* (fathead minnow) was published elsewhere<sup>21</sup> and is provided electronically as Supporting Information.

The data set is chemically heterogeneous and includes compounds acting by different modes and mechanisms such

as narcosis (type I, II and III), uncouplers of the oxidative phosphorylation, reactive electrophiles/ proelectrophiles, acetylcholinesterase inhibitors, and central nervous system (seizure) agents.

As noted above, all toxicity data were retrieved from Russom et al.<sup>21</sup> Tests were conducted using Lake Superior water at  $25 \pm 1$  °C. Aqueous toxicant concentrations were measured in all tests with quality assurance criteria requiring 80% agreement between duplicate samples and 90–100% spike recovery. Flow-through exposures were conducted using cycling proportional, modified Benoit or electronic diluters. Median lethal concentrations ( $\text{LC}_{50}$ , in mg/L) were calculated using the Trimmed Spearman-Kärber method, with 95% confidence intervals being calculated when possible. For the purposes of this study the  $\text{LC}_{50}$  values were converted to mmol/L, and the logarithm of their reciprocal values ( $\log(1/\text{LC}_{50})$ ) was used for QSAR modeling.

**Molecular Descriptors.** The molecular descriptor set contained the logarithm of the octanol–water partition coefficient ( $\log K_{\text{ow}}$ ), and two sets of electronic descriptors were obtained using the semiempirical AM1 Hamiltonian and the ab initio method B3LYP/6-31G\*\*. The two sets of electronic descriptors were considered separately from each other, both in combination with  $\log K_{\text{ow}}$ .  $\log K_{\text{ow}}$  was calculated by PALLAS software (PALLAS version 3.0, CompuDrug International Inc., South San Francisco, CA).

Two methods were employed for the calculation of quantum chemical descriptors. These were the semiempirical method AM1<sup>18</sup> as implemented in MOPAC93 (Software release 2002, Stewart, J. J. P.) and the ab initio method B3LYP/6-31G\*\* as provided by the Gaussian 98 (Gaussian 98 revision A.7, Gaussian Inc., Pittsburgh, PA) package. All calculations included geometry optimization, employing CORINA 2001<sup>22,23</sup> to generate initial 3D structures from Simplified Molecular Input Line Entry Specification (SMILES) strings, followed by SYBYL force field geometry optimizations (SYBYL version 6.8, Tripos Associates, St. Louis, MO) before starting quantum chemical calculations.<sup>16</sup> Because no conformational search routines were employed it may well be that, in individual cases, the finally achieved geometry does not represent the energy minimum of the conformational space. Thus, the geometries can be considered as optimized on a screening level which, however, is considered sufficient for the purposes of the present analyses.

A total of 162 chemical descriptors were calculated using the AM1 method and 121 descriptors using the B3LYP/6-31G\*\* method. These included the energies of the highest occupied ( $E_{\text{HOMO}}$ ) and the lowest unoccupied ( $E_{\text{LUMO}}$ ) molecular orbitals, their derivatives electronegativity (EN) and molecular hardness (HD), dipole moment ( $\mu$ ), series of atomic partial charges (Q), charged partial surface area (CPSA) descriptors, a set of delocalizability indices (D) available only from the semiempirical AM1 method, and geometrical descriptors. A summary of the calculated descriptors is provided in Table 1. A more detailed discussion of their calculation and mechanistic meaning is available in the literature.<sup>16</sup>

**Statistical Analyses. Multiple Linear Regression Analysis.** An in-house algorithm written in C code for DOS (developed at the European Centre for the Validation of Alternative Methods, Institute for Health and Consumer Protection, Joint Research Centre, Ispra, Italy) was utilized for variable

**Table 1.** Descriptors Calculated for QSAR Analysis and Their Definition<sup>d</sup>

symbol	no. of descriptors	definition
Descriptors Based on Molecular Orbital Energies		
E <sub>HOMO</sub>	1	energy of the highest occupied molecular orbital
E <sub>LUMO</sub>	1	energy of the lowest unoccupied molecular orbital
EN	1	molecular electronegativity
HD	1	molecular hardness
Descriptors Based on Charge Distribution		
$\mu$	1	molecular dipole moment
QYp_mx	4	maximum positive atomic charge considering particular atom of type Y (Y = C, H, N, X: halogen)
QYp_avp	4	average positive partial charge considering particular atom of type Y and normalized with respect to the number of atoms of this type charged positively (Y = C, H, N, X: halogen)
QYp_av	4	average positive partial charge considering particular atom of type Y and normalized with respect to the total number of atoms of this type in the molecule (Y = C, H, N, X: halogen)
QYn_mx	4	maximum negative atomic charge considering particular atom of type Y (Y = C, N, O, X: halogen)
QYn_avn	4	average negative partial charge considering particular atom of type Y and normalized with respect to the number of atoms of this type charged negatively (Y = C, N, O, X: halogen)
QYn_av	4	average negative partial charge considering particular atom of type Y and normalized with respect to the total number of atoms of this type in the molecule (Y = C, N, O, X: halogen)
Qp_mx	1	maximum positive atomic charge in a molecule considering all non-hydrogen atoms
Qp_avp	1	average positive partial charge considering all non-hydrogen atoms and normalized with respect to the number of non-hydrogen atoms charged positively
Qp_av	1	average positive partial charge considering all non-hydrogen atoms and normalized with respect to the total number of non-hydrogen atoms in the molecule
Qn_mx	1	maximum negative atomic charge in a molecule considering all non-hydrogen atoms
Qn_avn	1	average negative partial charge considering all non-hydrogen atoms and normalized with respect to the number of non-hydrogen atoms charged negatively
Qn_av	1	average negative partial charge considering all non-hydrogen atoms and normalized with respect to the total number of non-hydrogen atoms in the molecule
Descriptors Based on Molecular Surface Area and Net Atomic Charges <sup>a</sup>		
PPSA_YY	14	partial positive surface area = $\Sigma (+SA_i)$ (YY = 1, 1H, 1Z, 2, 2H, 2Z, 3, 3H, 3Z, 4, 4Z, 5, 5H, 5Z)
PNSA_YY	10	partial negative surface area = $\Sigma (-SA_i)$ (YY = 1, 1Z, 2, 2Z, 3, 3Z, 4, 4Z, 5, 5Z)
DPSA_YY	6	difference in charged partial surface area = PPSA_YY - PNSA_YY (YY = 1, 1Z, 2, 2Z, 3, 3Z)
FPSA_YY	9	fractional positive surface area = PPSA/SA (YY = 1, 1H, 1Z, 2, 2H, 2Z, 3, 3H, 3Z)
FNSA_YY	6	fractional negative surface area = PNSA/SA (YY = 1, 1Z, 2, 2Z, 3, 3Z)
WPSA_YY	9	weighted positive surface area = (PPSA * SA)/1000 (YY = 1, 1H, 1Z, 2, 2H, 2Z, 3, 3H, 3Z)
WNSA_YY	6	weighted negative surface area = (PNSA * SA)/1000 (YY = 1, 1Z, 2, 2Z, 3, 3Z)
RPCG	3	relative positive charge = Qp_mx/(sum total positive charge) (RPCG, RPCG_H, and RPCG_Z)
RNCG	2	relative positive charge = Qn_mx/(sum total negative charge) (RNCG and RNCG_Z)
SPMX	3	maximum positive charge weighted by associated surface area = Qp_mx * (its respective surface area) (SPMX, SPMX_H, and SPMX_Z)
SNMX	2	maximum negative charge weighted by associated surface area = Qn_mx * (its respective surface area) (SNMX and SNMX_Z)
RPCS	3	relative maximum positive charge weighted by associated surface area = SA <sub>MPOS</sub> (SA of most positively charged atom) * RPCG (RPCS, RPCS_H, and RPCS_Z)
RNCS	2	relative maximum negative charge weighted by associated surface area = SA <sub>MNEG</sub> * RNCG (RNCS and RNCS_Z)
Descriptors Based on Molecular Orbital Wave Functions and Energies <sup>b</sup> (AM1 only)		
D <sup>N</sup> Y	4	sum of the acceptor delocalizabilities of all atomic sites of type Y in a molecule (Y = C, H, N, O)
D <sup>N</sup> Yav	4	average acceptor delocalizability of all atomic sites of type Y in a molecule (Y = C, H, N, O)
D <sup>N</sup> Yaomx	4	maximum acceptor delocalizability of all atomic orbitals (AOs) of all atoms of type Y (Y = C, H, N, O)
D <sup>N</sup> Ymx	4	maximum acceptor delocalizability of all atomic sites of type Y in a molecule (Y = C, H, N, O)
D <sup>E</sup> Y	4	sum of the donor delocalizabilities of all atomic sites of type Y in a molecule (Y = C, H, N, O)
D <sup>E</sup> Yav	4	average donor delocalizability of all atomic sites of type Y in a molecule (Y = C, H, N, O)
D <sup>E</sup> Yaomx	4	maximum donor delocalizability of all atomic orbitals (AOs) of all atoms of type Y (Y = C, H, N, O)
D <sup>E</sup> Ymx	4	maximum donor delocalizability considering particular atom type (Y = C, H, N, O)
D <sup>N</sup>	1	sum of the acceptor delocalizabilities of all atomic sites in a molecule
D <sup>N</sup> av	1	average acceptor delocalizability of all atomic sites in a molecule
D <sup>N</sup> mx	1	maximum acceptor delocalizability of all atomic sites in a molecule
D <sup>E</sup>	1	sum of the donor delocalizabilities of all atomic sites in a molecule
D <sup>E</sup> av	1	average donor delocalizability of all atomic sites in a molecule
D <sup>E</sup> mx	1	maximum donor delocalizability of all atomic sites in a molecule
PIrr	1	sum of self-polarizabilities of all atomic sites r in a molecule (cf. Schüürmann 1998)
PIrrav	1	average self-polarizability of all atomic sites r in a molecule
PIrrmx	1	maximum self-polarizability considering all atomic sites r in a molecule
Geometrical Descriptors <sup>c</sup>		
SA	1	molecular surface area
V	1	molecular volume
SA_Y	7	molecular surface area encoded by particular atoms (Y = C, H, O, N, X: halogen, S: sulfur, Z: heavy atom)
Rad	1	radius of sphere with volume identical to the van der Waals volume of the molecule
Oval	1	ovality = ratio of molecular surface area over surface area of sphere with identical volume (= measure of deviation from spherical shape)
Log K <sub>ow</sub>	1	octanol–water partition coefficient (PALLAS software)

<sup>a</sup> In the group of descriptors based on molecular surface area and net atomic charges: 1 = CPSA itself (i.e. PPSA and PNSA); 2 = total charged weighted CPSA. Equals to CPSA, multiplied by the sum of the positive or negative atomic charges; 3 = atomic charge weighted CPSA. Sum of the products of positively (negatively) charged atomic surface areas and associated atomic charges; 4 = average net atomic positive (negative) charge, weighted by associated surface area. Equals to CPSA, multiplied by the average positive or negative charge (Qp\_av or Qn\_av); 5 = average net atomic positive (negative) charge considering only positively (negatively) charged atoms, weighted by associated surface area. Equals to CPSA, multiplied by the average positive or negative charge when referring to positively or negatively charged atoms only (Qp\_avp or Qn\_avn); H = hydrogen atom; Z = heavy atom. <sup>b</sup> In the group of descriptors based on molecular orbital wave functions and energies: index N (nucleophilic) indicates acceptor delocalizability toward attack from nucleophile; index E (electrophilic) indicates donor delocalizability toward attack from electrophile. <sup>c</sup> For all geometric descriptors, MST radii were employed as listed in ref 42. <sup>d</sup> For more details cf. refs 16 and 34.



selection with MLR as the fitting function. The program allows the user to derive  $k$ -variable regression equations combining all possible combinations of  $k$  descriptors with a specified level of intercorrelation between them ( $r = 0.7$  in this study). This procedure is referred to as "all subsets analysis" further in the text. For best subsets identification, the maximum  $r^2$  criterion was used. To reduce the computational time for the development of all possible subsets, an initial reduction in the number of the variables was performed. It was based on the generation of all one-parameter equations between the descriptors and the dependent variable ( $\log(1/LC_{50})$ ). The statistical significance ( $p$ ) of the descriptors in the one-parameter equations was recorded, and descriptors with  $p > 0.05$  were discarded. As a result of this procedure,  $\log K_{ow}$ , together with 126 descriptors calculated by AM1, and 101 descriptors calculated by B3LYP/6-31G\*\* entered the all subset analyses.

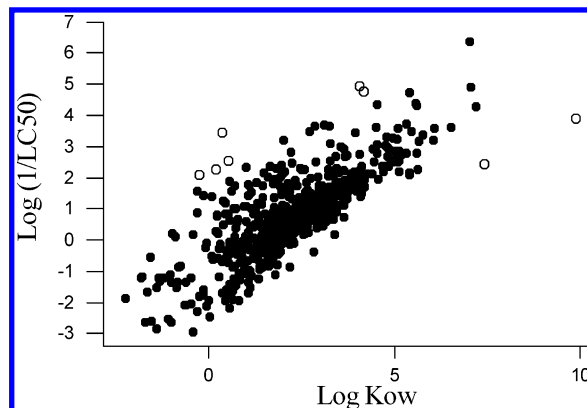
Further model evaluation and leave-one-out cross-validation were performed using the MINITAB ver. 13.1 (Minitab Inc., State College, PA) software. The first 10  $k$ -descriptor models (scored by  $r^2$ ) were considered. They were checked in MINITAB for the significance of the descriptors ( $p$  values  $> 0.05$  were not allowed), and their statistics (regression coefficient ( $r^2$ ), cross-validated regression coefficient ( $r^2_{CV}$ ), standard error of the estimate ( $s$ ), and the Fisher's criterion ( $F$ )) were recorded. The error of the coefficients is shown in parentheses in the QSARs. Outliers were not excluded from the MLR models.

**Partial Least-Squares Analysis.** PLS was applied using the SIMCA-P ver. 9.0 (Umetrics AB, Umeå, Sweden) software. Variable importance in the projection (VIP) was used as the criterion for initial variable reduction with PLS. VIP is defined as the influence of every descriptor in the model ( $D_k$ ) on the dependent variable  $Y$  (SIMCA-P User Guide).<sup>24</sup> Descriptors with high VIP values are more relevant to explain  $Y$  than descriptors with low VIP values. Initially, descriptors with  $VIP > 1$  were excluded. The remaining variables were screened in an iterative procedure in parallel with outlier exclusion until only important (with relatively high coefficients after centring and scaling of the variables) and significant (the value of the coefficient is higher than its error) descriptors remain in the model.

The statistical performance of PLS models was assessed by the cumulative sum of squares (SS) of all descriptors participating in the model explained by all components ( $r^2(X)$ ), cumulative SS of the toxicity explained by the model ( $r^2(Y)$ ), cumulative  $q^2(Y)$  for all components, and the root-mean-square error of the fit (RMSEE). The number of significant principal components was determined by the cross-validated  $q^2(Y)$ . A negative contribution of a component to the cumulative  $q^2(Y)$  was used as an indication that the component is not significant.

## RESULTS

The fathead minnow acute toxicity database, which is considered of the highest quality for the development of aquatic toxicology QSARs,<sup>25</sup> was evaluated by octanol–water partition coefficient and two sets of electronic descriptors. The methods utilized for calculation of the electronic indices reflected different levels of theory in the description of the electronic structure. To compare the relevance of the



**Figure 1.** A plot of toxicity ( $\log(1/LC_{50})$ ) versus hydrophobicity ( $\log K_{ow}$ ) for the fathead minnow database. The outliers are indicated with an empty circle.

semiempirical and the ab initio methods for the development of acute aquatic toxicity QSARs, two statistical methods for analysis were employed.

The octanol–water partition coefficient was inevitably included in the QSAR analysis because this is the only single descriptor which is able to model a large portion of the variability in any acute aquatic toxicity database.<sup>26</sup> This fact was confirmed once again in the present study by development of a statistically significant, hydrophobicity-dependent model:

$$\log(1/LC_{50}) = 0.670 (0.022) \log K_{ow} - 0.631 (0.061) \quad (1)$$

$$n = 568, r^2 = 0.610, r^2_{CV} = 0.606, s = 0.865, F = 885$$

The relationship between the  $\log(1/LC_{50})$  and  $\log K_{ow}$  is illustrated in Figure 1. The inclusion of a quadratic term in eq 1 did not improve the statistics of the model ( $n = 568$ ,  $r^2 = 0.610$ ,  $r^2_{CV} = 0.601$ ,  $s = 0.866$ ,  $F = 442$ ,  $T_{(\log K_{ow})^2} = -0.05$ , where  $T$  is the coefficient-to-error ratio).

Six outliers to eq 1, all with positive standardized residuals greater than 3, were identified. These were strychnine, rotenone, acrolein, allyl alcohol, malononitrile, and N-vinylcarbazole. Two outliers with considerable negative residuals (4,4'-isopropylidenebis(2,6-dichlorophenol) and tetrabutyltin) were visually identified from Figure 1. All eight outliers were excluded, and the model was redeveloped:

$$\log(1/LC_{50}) = 0.700 (0.022) \log K_{ow} - 0.720 (0.058) \quad (2)$$

$$n = 560, r^2 = 0.650, r^2_{CV} = 0.647, s = 0.803, F = 1034$$

Further outliers with significant standardized residuals appeared, but they were not removed because the appearance of outliers is iterative in nature and no-efficient cutoff could be applied to stop this process. Moreover, from Figure 1 it can be seen that further exclusion of outliers will be a statistical artifact, and there are no single compounds that can influence significantly the slope and the intercept in eq 2.

The statistical performance of eq 2 demonstrated that although considerable, the contribution of  $\log K_{ow}$  to model the acute toxicity to fathead minnow in this large and heterogeneous data set is not enough for development of a

**Table 2.** Descriptors Used in, and Statistics of, the MLR Models Obtained from log  $K_{ow}$  and AM1 Calculated Quantum Chemical Descriptors ( $n = 568$ )

	$r^2$	$r^2_{CV}$	$s$	$F$
Two-Parameter Models				
Log $K_{ow}$ , $E_{LUMO}^a$	0.663	0.658	0.805	555
Log $K_{ow}$ , HD	0.656	0.652	0.813	539
Log $K_{ow}$ , $D^{N_{av}}$	0.654	0.650	0.815	534
Log $K_{ow}$ , $PIrr_{av}$	0.650	0.645	0.820	524
Log $K_{ow}$ , $PIrr_{mx}$	0.650	0.645	0.821	524
Log $K_{ow}$ , $D^{N_{Cav}}$	0.645	0.641	0.826	514
Log $K_{ow}$ , $FPSA_{1H}$	0.644	0.640	0.827	511
Log $K_{ow}$ , $D^{N_{Cmx}}$	0.642	0.638	0.829	507
Log $K_{ow}$ , $PNSA_{1Z}$	0.642	0.637	0.830	506
Log $K_{ow}$ , $PNSA_1$	0.642	0.637	0.830	506
Three-Parameter Models				
Log $K_{ow}$ , $FPSA_{1H}$ , $PIrr^b$	0.676	0.671	0.790	392
Log $K_{ow}$ , $FPSA_{1H}$ , $D^N$	0.676	0.670	0.790	392
Log $K_{ow}$ , $E_{LUMO}$ , $D^E$	0.675	0.669	0.792	390
Log $K_{ow}$ , $PIrr_{mx}$ , $D^{NO}$	0.675	0.669	0.792	390
Log $K_{ow}$ , $E_{LUMO}$ , $V$	0.674	0.668	0.792	389
Log $K_{ow}$ , $PIrr_{mx}$ , $SA_{-O}$	0.673	0.668	0.793	387
Log $K_{ow}$ , $E_{LUMO}$ , $SA$	0.673	0.667	0.793	387
Log $K_{ow}$ , $FPSA_{1H}$ , $D^E$	0.673	0.667	0.793	387
Log $K_{ow}$ , $D^{N_{av}}$ , $SA$	0.673	0.667	0.794	387
Log $K_{ow}$ , $D^{N_{av}}$ , $V$	0.673	0.667	0.794	386
Four-Parameter Models				
Log $K_{ow}$ , $PIrr_{mx}$ , $FPSA_{3H}$ , $D^{Nc}$	0.698	0.692	0.763	325
Log $K_{ow}$ , $PIrr_{mx}$ , $FPSA_{3H}$ , $PIrr$	0.697	0.691	0.765	324
Log $K_{ow}$ , $PIrr_{mx}$ , $SE$ , $FPSA_{3H}$	0.696	0.690	0.766	322
Log $K_{ow}$ , $PIrr_{mx}$ , $D^{N_{Cav}}$ , $SA$	0.693	0.686	0.769	318
Log $K_{ow}$ , $PIrr_{mx}$ , $FPSA_{3H}$ , $SA$	0.692	0.686	0.770	317
Log $K_{ow}$ , $PIrr_{mx}$ , $FPSA_{1H}$ , $Oval$	0.692	0.686	0.770	317
Log $K_{ow}$ , $PIrr_{mx}$ , $D^{N_{Cav}}$ , $D^E$	0.692	0.686	0.770	317
Log $K_{ow}$ , $PIrr_{mx}$ , $FPSA_{3H}$ , $D^{NC}$	0.692	0.685	0.771	316
Log $K_{ow}$ , $PIrr_{mx}$ , $D^{N_{Cav}}$ , $V$	0.692	0.685	0.771	316
Log $K_{ow}$ , $PIrr_{mx}$ , $D^{N_{Cav}}$ , $Oval$	0.692	0.685	0.771	316

<sup>a</sup>  $\log(1/LC_{50}) = 0.614 (0.022) \log K_{ow} - 0.240 (0.026) E_{LUMO} - 0.392 (0.062)$ . <sup>b</sup>  $\log(1/LC_{50}) = 0.533 (0.026) \log K_{ow} - 1.746 (0.177) FPSA_{1H} - 0.482 (0.065) PIrr - 0.131 (0.115)$ . <sup>c</sup>  $\log(1/LC_{50}) = 0.452 (0.028) \log K_{ow} - 95.29 (12.19) PIrr_{mx} - 19.89 (2.49) FPSA_{3H} + 0.228 (0.029) D^N - 12.20 (1.59)$ .

robust, predictive model. An improvement was sought by the respective inclusion of electronic descriptors from the two calculated pools (using AM1 and B3LYP/6-31G\*\* methods) without mixing of descriptors from the two sources.

**Multiple Linear Regression Analysis.** The best 10 two-, three-, and four-descriptor models obtained with log  $K_{ow}$  and descriptors calculated by the AM1 and B3LYP/6-31G\*\* methods are listed in Tables 2 and 3, respectively. Models with higher dimensionality were not developed due to lack of the further improvement of the statistical performance of the models when more descriptors were added. It is evident from Table 2 that the most preferred second (in addition to log  $K_{ow}$ ) descriptors when calculated by the semiempirical AM1 Hamiltonian were the energy of the lowest unoccupied molecular orbital ( $E_{LUMO}$ ) and the molecular hardness (HD). Indiscriminately better models with respect to the  $r^2$  were obtained with the same two descriptors, calculated using the ab initio B3LYP/6-31G\*\* function (Table 3). It is interesting to note that the correlation between  $E_{LUMO}$  and HD calculated by the two selected methods ( $E_{LUMO}$  (AM1)/ $E_{LUMO}$  (B3LYP),  $r^2 = 0.85$ ; HD (AM1)/HD (B3LYP),  $r^2 = 0.84$ ) is better than the correlation between the two descriptors calculated by the either of methods ( $E_{LUMO}$  (AM1)/HD (AM1),  $r^2 = 0.74$ ;  $E_{LUMO}$  (B3LYP)/HD (B3LYP),  $r^2 = 0.72$ ), which

**Table 3.** Descriptors Used in, and Statistics of, the MLR Models Obtained from log  $K_{ow}$  and B3LYP Calculated Quantum Chemical Descriptors ( $n = 568$ )

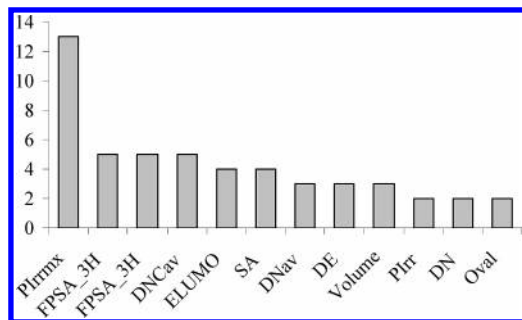
	$r^2$	$r^2_{CV}$	$s$	$F$
Two-Parameter Models				
Log $K_{ow}$ , $E_{LUMO}^a$	0.667	0.663	0.800	565
Log $K_{ow}$ , HD	0.664	0.660	0.803	559
Log $K_{ow}$ , $PPSA_{2Z}$	0.658	0.653	0.810	545
Log $K_{ow}$ , $FPSA_{2Z}$	0.657	0.653	0.812	541
Log $K_{ow}$ , $PPSA_{4Z}$	0.651	0.646	0.820	526
Log $K_{ow}$ , $WPSA_{2Z}$	0.648	0.643	0.822	521
Log $K_{ow}$ , $PPSA_{5Z}$	0.647	0.642	0.823	519
Log $K_{ow}$ , $DPSA_{2Z}$	0.647	0.643	0.823	518
Log $K_{ow}$ , $PNSA_4$	0.646	0.642	0.824	516
Log $K_{ow}$ , $DPSA_{3Z}$	0.646	0.641	0.825	515
Three-Parameter Models				
Log $K_{ow}$ , $E_{LUMO}$ , $WPSA_{2Z}^b$	0.685	0.679	0.778	409
Log $K_{ow}$ , HD, $PPSA_{2Z}$	0.685	0.679	0.779	408
Log $K_{ow}$ , HD, $WPSA_{2Z}$	0.684	0.678	0.780	407
Log $K_{ow}$ , $E_{LUMO}$ , $PPSA_{2Z}$	0.683	0.677	0.781	405
Log $K_{ow}$ , $E_{LUMO}$ , $DPSA_{2Z}$	0.682	0.677	0.782	404
Log $K_{ow}$ , $E_{LUMO}$ , $V$	0.682	0.677	0.782	404
Log $K_{ow}$ , $E_{LUMO}$ , $WNSA_{2Z}$	0.681	0.676	0.783	402
Log $K_{ow}$ , $E_{LUMO}$ , $WNSA_2$	0.681	0.676	0.783	402
Log $K_{ow}$ , $E_{LUMO}$ , $SA$	0.681	0.676	0.784	402
Log $K_{ow}$ , $E_{LUMO}$ , $WNSA_{1Z}$	0.680	0.675	0.785	400
Four-Parameter Models				
Log $K_{ow}$ , $E_{LUMO}$ , $DPSA_{2Z}$ , $Qp_{mx}^c$	0.693	0.686	0.770	317
Log $K_{ow}$ , $E_{LUMO}$ , $WPSA_{2Z}$ , $RPCG$	0.691	0.684	0.773	314
Log $K_{ow}$ , $E_{LUMO}$ , $WNSA_{3Z}$ , $Qp_{mx}$	0.689	0.682	0.775	312
Log $K_{ow}$ , $E_{LUMO}$ , $WNSA_3$ , $Qp_{mx}$	0.689	0.682	0.775	312
Log $K_{ow}$ , $E_{LUMO}$ , $WPSA_{2Z}$ , $Qn_{avn}$	0.689	0.682	0.775	312
Log $K_{ow}$ , $E_{LUMO}$ , $PPSA_{2Z}$ , $RPCG$	0.689	0.682	0.775	311
Log $K_{ow}$ , $E_{LUMO}$ , $WPSA_{2Z}$ , $Qp_{mx}$	0.689	0.682	0.775	311
Log $K_{ow}$ , $E_{LUMO}$ , $WPSA_{2Z}$ , $V$	0.689	0.681	0.775	311
Log $K_{ow}$ , HD, $WNSA_{1Z}$ , $FPSA_{2Z}$	0.689	0.682	0.775	311
Log $K_{ow}$ , HD, $FPSA_{2Z}$ , $WNSA_1$	0.689	0.682	0.775	311

<sup>a</sup>  $\log(1/LC_{50}) = 0.630 (0.021) \log K_{ow} - 0.242 (0.025) E_{LUMO} - 0.603 (0.057)$ . <sup>b</sup>  $\log(1/LC_{50}) = 0.589 (0.022) \log K_{ow} - 0.203 (0.025) E_{LUMO} + 0.0237 (0.0041) WPSA_{2Z} - 0.614 (0.055)$ . <sup>c</sup>  $\log(1/LC_{50}) = 0.563 (0.023) \log K_{ow} - 0.227 (0.026) E_{LUMO} + 0.00223 (0.0003) DPSA_{2Z} - 0.778 (0.181) Qp_{mx} - 0.535 (0.072)$ .

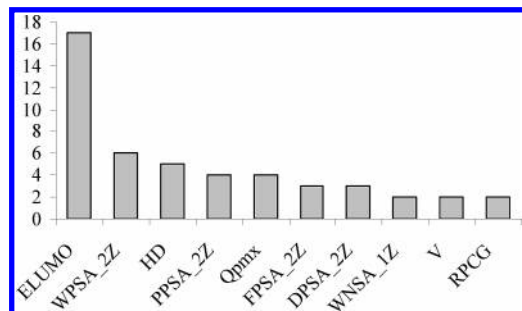
reflects the difference in information between  $E_{LUMO}$  and HD. Further common trends in the best two-parameter models could not be observed. Among the semiempirically calculated descriptors the choice went further to the delocalizability indices (which are not available from the ab initio methods) and different (between Tables 2 and 3, two-parameter models) partial charged surface area descriptors.

The comparison of the statistical criteria in Tables 2 and 3 indicates that there is not a substantial difference between  $r^2$ ,  $r^2_{CV}$ ,  $s$ , and  $F$  when comparing the best two-, three-, and four parameter models. Thus, for the 10 best two-parameter models in Table 2, the  $r^2$  changes from 0.663 to 0.642, while in Table 3 it changes from 0.667 to 0.646. Conversely, the  $r^2$  decreases from 0.698 to 0.692 in the best 10 four-parameter models in Table 2, while in Table 3 it is slightly worse and decreases from 0.693 to 0.689. Bearing in mind the imminent uncertainty associated with the measurement of the biological endpoint and the calculation of the log  $K_{ow}$ , the observed differences in the statistical criteria between Tables 2 and 3 can be considered insignificant.

Although the use of electronic indices calculated at different levels of theory resulted in equations with similar statistical quality, the type of the parameters that entered the analyses significantly affected the choice of the descriptors



**Figure 2.** Frequency of occurrence (number of models) of the quantum chemical descriptors obtained in AM1 calculation which were selected in the best 10 two-, three-, and four-descriptor models. Only descriptors that appear more than once are shown on the plot.



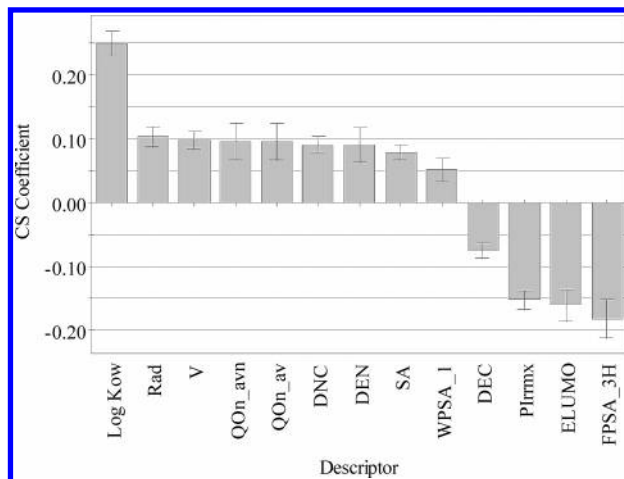
**Figure 3.** Frequency of occurrence (number of models) of the quantum chemical descriptors obtained from the B3LYP calculation which were selected in the best 10 two-, three-, and four-descriptor models. Only descriptors that appear more than once are shown on the plot.

**Table 4.** Descriptors, Their Raw Coefficients, and the Associated Statistics in the PLS Models Obtained from  $\log K_{ow}$  and Electronic Parameters

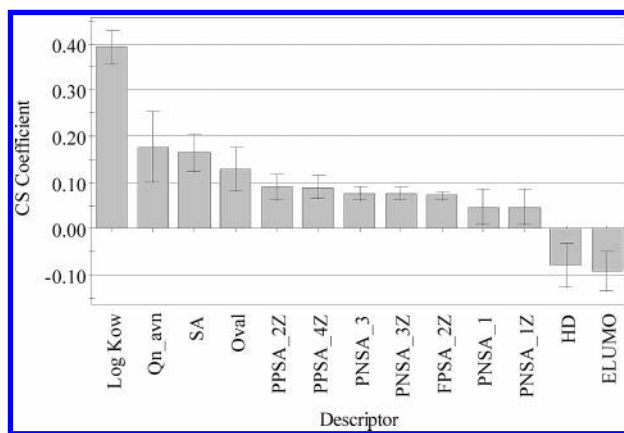
model with AM1 descriptors		model with B3LYP/6-31G** descriptors	
descriptor	coefficient	descriptor	coefficient
constant	-10.5	constant	-1.01
$\log K_{ow}$	0.214	$\log K_{ow}$	0.338
Rad	0.378	Qn_avn	3.13
V	0.00258	SA	0.00312
QOn_avn	0.766	Oval	1.02
QOn_av	0.766	PPSA_2Z	0.00452
D <sup>NC</sup>	0.115	PPSA_4Z	0.0673
D <sup>EN</sup>	0.528	PNSA_3	0.0116
SA	0.00151	PNSA_3Z	0.0116
WPSA_1	0.00264	FPSA_2Z	0.876
D <sup>EC</sup>	-0.107	PNSA_1	0.00203
Plrmx	-76.2	PNSA_1Z	0.00203
ELUMO	-0.159	HD	-0.227
FPSA_3H	-17.2	ELUMO	-0.0905
$n = 562$ , $PC = 2$ , $r^2(X) = 0.625$ , $r^2(Y) = 0.726$ , $q^2(Y) = 0.719$ , RMSEE = 0.721		$n = 562$ , $PC = 3$ , $r^2(X) = 0.754$ , $r^2(Y) = 0.724$ , $q^2(Y) = 0.711$ , RMSEE = 0.723	

in the best models. To illustrate this point, a count of the descriptors that appeared more than once in Tables 2 and 3 was performed. The frequency of the appearance of the descriptors (with exception of  $\log K_{ow}$  which appeared in all models) is shown in Figures 2 (from Table 2) and 3 (from Table 3).

**Partial Least-Squares Analysis.** Following the procedure for the selection of variables described in the materials and methods, two final PLS models were developed. They are summarized in Table 4. The outliers in the PLS model with AM1 descriptors included the following: acrolein, allyl



**Figure 4.** PLS coefficients, obtained after centering and scaling of  $\log K_{ow}$  and quantum chemical descriptors calculated using the AM1 method.



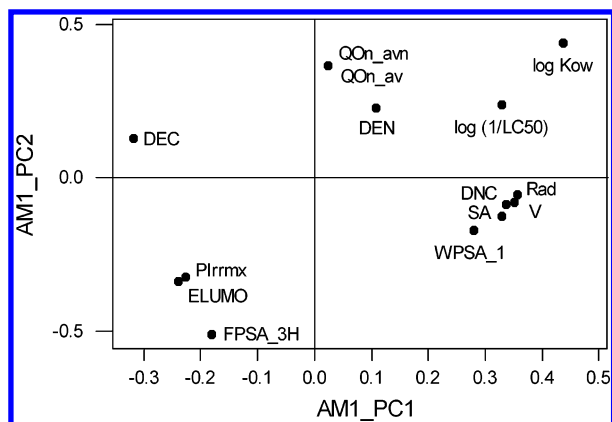
**Figure 5.** PLS coefficients, obtained after centering and scaling of  $\log K_{ow}$  and quantum chemical descriptors calculated using the B3LYP method.

alcohol, 2-propyn-1-ol, malononitrile, N-vinylcarbazole, 1,1-dimethylhydrazine, all of them with positive residuals. The final PLS model with B3LYP/6-31G\*\* descriptors was derived without acrolein, allyl alcohol, 2-propyn-1-ol, malononitrile, N-vinylcarbazole (with positive residuals), and 4,4'-isopropylidenebis(2,6-dichlorophenol) (with negative residuals).

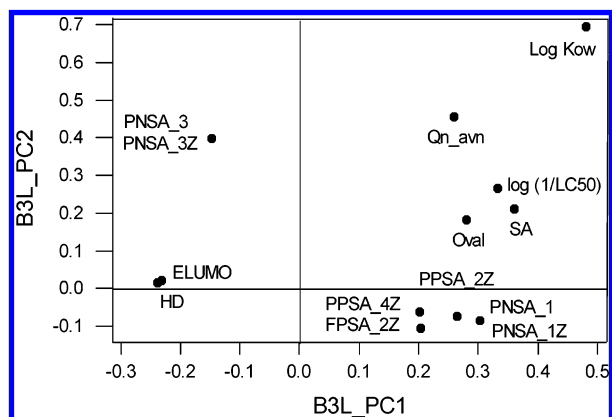
The two PLS models listed in Table 4 were derived using the same number of descriptors (thirteen) and without the same number of outliers (six). The outliers are almost the same (with one exception) between the two PLS models and agree with the outliers from eq 1. The statistical performance of the PLS models with AM1 and B3LYP/6-31G\*\* descriptors is almost equivalent with respect to all statistical criteria that have been assessed.

To visualize the relative importance of the descriptors, their standardized and centered coefficients were plotted in Figures 4 (from the PLS model with AM1 descriptors) and 5 (from the PLS model with B3LYP/6-31G\*\* descriptors). Note, that the PLS model with AM1 descriptors included four indices (D<sup>NC</sup>, D<sup>EC</sup>, D<sup>EN</sup>, and Plrmx) which were not available from the B3LYP/6-31G\*\* calculations. The score plots of the descriptors in the two PLS models (Table 4) are shown in Figures 6 (from the PLS model with AM1 descriptors) and 7 (from the PLS model with B3LYP/6-31G\*\* descriptors).





**Figure 6.** A score plot of the two principal components derived from  $\log K_{ow}$  and quantum chemical descriptors calculated using the AM1 method.



**Figure 7.** A score plot of the first two principal components derived from  $\log K_{ow}$  and quantum chemical descriptors calculated using the B3LYP method.

## DISCUSSION

This study describes the comparative development of QSARs for the toxicity to the fathead minnow of a heterogeneous group of organic chemicals utilizing the octanol–water partition coefficient and two sets of quantum chemical descriptors. The latter were generated by the AM1 and B3LYP/6-31G\*\* methods that reflect different levels of theory in the calculation of the electronic structure. Similar sized data sets with acute toxicity data to fathead minnow have been analyzed previously by other workers.<sup>27–30</sup> The descriptors in this study were calculated deliberately to enable the comparison and assessment of the need of ab initio calculations for the development of toxicological QSARs in a large and chemically heterogeneous data sets. The choice of the statistical tools (MLR and PLS) was motivated by the requirement for development of transparent and easily reproducible QSARs. The methods utilized for the selection of variables allowed objective judgment on the use of high-level but computationally demanding quantum chemical calculations for the development of acute aquatic toxicity QSARs.

In the QSAR analysis, the first descriptor ( $\log K_{ow}$ ) was selected empirically as a result of extensive experience in modeling of acute aquatic toxicity.<sup>26,31–33</sup> A strong trend of increasing toxicity with increasing hydrophobicity underpins the data set. Indeed,  $\log K_{ow}$  alone accounts for approximately 65% of the variance in the toxicity data (eq 2). This result implicitly makes the improving of the model difficult. The

observed general trend in the modeling of acute aquatic toxicity data is a consequence of the fact that to exhibit a toxic effect any chemical must be absorbed from the environment and that uptake is hydrophobicity dependent. It should be noted, however, that eqs 1 and 2 in this study are not baseline QSAR models for narcosis. Instead eqs 1 and 2 were developed to illustrate the portion of the variability in the fathead minnow toxicity values that can be explained by hydrophobicity.

The generation of all possible  $k$ -descriptor models is considered the most robust selection procedure that can be combined with any statistical method for analysis. It enables the best QSAR model to be found according to a predefined criterion. This procedure was chosen to ensure that no appropriate descriptor was omitted due to limitations of the statistical approach. As a result, populations of statistically similar  $k$ -descriptor MLR models were generated and subsequently analyzed.

The best two-parameter equations using  $\log K_{ow}$  and electronic indices calculated by the AM1 and B3LYP/6-31G\*\* methods were developed with  $E_{LUMO}$  and HD. Following the definition of frontier orbital energies, these parameters quantify—within the limits of Hartree–Fock theory—the energy associated with the acceptance ( $E_{LUMO}$ ) or donation ( $E_{HOMO}$ ) of an electron, corresponding to a full one-electron reduction and oxidation, respectively (cf. ref 16). As such,  $E_{LUMO}$  and  $E_{HOMO}$  are expected to correlate with the global readiness of the molecule to accept or donate electron charge, the latter of which appears to make them useful as global descriptors for electrophilicity and nucleophilicity in the context of computational toxicology. Note that the molecular electronegativity,  $EN = \frac{1}{2}(E_{HOMO} + E_{LUMO})$ , is an alternative global measure for the tendency of molecules to attract electrons.<sup>16,34</sup> It was originally defined considering the electron transfer between two reaction partners, with EN being equal for two chemical species A and B when the energies associated with the one-electron transfer from A to B and from B to A are equal.

Molecular hardness (HD) is another descriptor to characterize the readiness of a molecule to gain or lose electrons and can be defined as the resistance of the electronic structure to undergo changes.<sup>16</sup> Low HD indicates a greater readiness of the molecule to accept or donate electron charge and the high HD signifies considerable resistance to changes.

Comparison of Tables 2 and 3 (MLR results) as well as of Figures 4 and 5 (PLS results) shows that the charged partial surface area (CPSA) descriptors are generally more significant with B3LYP than with AM1. Keeping in mind that CPSA parameters characterize the potential of molecular surfaces to undergo Coulomb interactions, the differences between AM1 and B3LYP are likely to reflect respective differences in the quantification of net atomic charges. Indeed, a recent study revealed significant and systematic differences in the calculation of atomic charges between semiempirical and ab initio quantum chemical schemes.<sup>16</sup>

Another difference between Tables 2 and 3 concerns the use of wave function-based descriptors such as acceptor and donor delocalizabilities and the molecular self-polarizability, Plrr. It should be noted that these parameters were not available from the B3LYP level of quantum chemistry. As can be seen from Table 2, as well as from Figure 2, the maximum value of the self-polarizability within a given

molecule (Pirrmx) is frequently selected (and often the major) electronic parameter besides hydrophobicity when employing the semiempirical AM1 calculations. PIrr was originally introduced in a perturbational treatment of intermolecular interactions as a measure for the readiness of individual atomic sites to accommodate local changes in the energy of site-specific electrons<sup>34</sup> and can be understood as a quantity related to the molecular softness. As such, PIrr probably encodes information related to HD, the latter of which has proven particularly useful as a B3LYP descriptor for modeling fish toxicity (cf. Table 3 and Figure 3). Similar differences also hold true for the PLS results as can be seen from Figures 4 and 5.

Finally, Figures 6 and 7 reveal the grouping of the descriptors used in the development of PLS models with AM1 and B3LYP descriptors (Table 4). It is evident from the plots that the log  $K_{ow}$  serves as a single descriptor highly loaded in the first two principal components of the models as shown in Table 4. Figure 6 demonstrates a logical grouping of molecular size descriptors such as molecular volume (V), surface area (SA), and the radius of a sphere with a volume identical to the van der Waals volume of the molecule (Rad). However, the appearance of the sum of the acceptor delocalizabilities over all carbon atoms in the molecule ( $D^N C$ ) in the same cluster of descriptors is more difficult to explain. At a similar position in Figure 7, a large cluster formed by charge partial surface area descriptors (PNSA\_1, PNSA\_1Z, FPSA\_2Z, PPSA\_2Z, and PPSA\_4Z) can be seen. It is interesting to highlight the close relationship between  $E_{LUMO}$  and PIrrmx in Figure 6 and between  $E_{LUMO}$  and HD in Figure 7. This observation concurs well with the theoretical considerations regarding about the relationship between  $E_{LUMO}$  and HD,  $E_{LUMO}$  and PIrrmx, and, indirectly, between HD and PIrrmx.

## CONCLUSIONS

The need of ab initio quantum chemical calculations for the development of toxicological structure–activity relationships was assessed. The results demonstrated that for a large and chemically diverse data set the choice of the precise but time-consuming calculations at high level of theory does not contribute noticeably to the quality of the derived QSARs. Statistically similar models were obtained using semiempirical (AM1) and ab initio (B3LYP) calculated descriptors supported by the use of the octanol–water partition coefficient (log  $K_{ow}$ ). The considerable difference between the models concerned the choice of descriptor for QSAR, independent of the statistical technique applied (MLR or PLS). Thus, with B3LYP the charged partial surface area (CPSA) descriptors were generally significant, while with AM1 the preference was given to size descriptors and delocalizability indices (the latter not being available from the ab initio calculations). However, the specific interpretation of the multivariate models for large and structurally diverse data sets is clouded by the apparent or hidden relationships between the quantum chemical descriptors and lack of mechanistic uniformity within the chemical series.

## ACKNOWLEDGMENT

The authors acknowledge the European Union IMAG-ETOX Research Training Network (HPRN-CT- 1999-00015)

for the opportunity to work together and for the financial support of Drs. Netzeva and Aptula.

**Supporting Information Available:** Chemical Abstract Number (CAS), name, and acute toxicity to the fathead minnow (log (1/LC<sub>50</sub>)) of the compounds studied. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## REFERENCES AND NOTES

- (1) Hartung, T.; Bremer, S.; Casati, S.; Coecke, S.; Covri, R.; Fortaner, S.; Gribaldo, L.; Halder, M.; Roi, A. J.; Prieto, P.; Sabbioni, E.; Worth, A.; Zuang, V. ECVAM's response to the changing political environment for alternatives: consequences of the European Union chemicals and cosmetics policies. *ATLA* **2003**, *35*, 473–481.
- (2) Worth, A. P.; van Leeuwen, C. J.; Hartung, T. The prospects for using (Q)SARs in a changing political environment – high expectations and a key role for the European Commission's Joint Research Centre. *SAR QSAR Environ. Res.* **2004**, *15*, 331–343.
- (3) Schultz, T. W.; Cronin, M. T. D. Essential and desirable characteristics of ecotoxicity quantitative structure–activity relationships. *Environ. Toxicol. Chem.* **2003**, *22*, 599–607.
- (4) Worth, A. P.; Hartung, T.; van Leeuwen, C. J. The role of the European Centre for the Validation of Alternative Methods (ECVAM) in the validation of (Q)SARs. *SAR QSAR Environ. Res.* **2004**, *15*, 345–358.
- (5) Worth, A. P.; Cronin, M. T. D.; van Leeuwen, C. J. A framework for promoting the acceptance and regulatory use of (quantitative) structure–activity relationships. In *Predicting Chemical Toxicity and Fate*; Cronin, M. T. D., Livingstone, D. J., Eds.; CRC Press: Boca Raton, 2004; pp 427–438.
- (6) ECETOC, 2003. *QSARs: Evaluation of the Commercially Available Software for Human Health and Environmental Endpoints with Respect to Chemical Management Applications*; ECETOC Technical Report No. 89; Brussels, Belgium, pp 19–25.
- (7) Schultz, T. W.; Cronin, M. T. D.; Walker, J. D.; Aptula, A. O. Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective, *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 1–22.
- (8) Veith, G. D.; Greenwood, B.; Hunter, R. S.; Niemi, G. J.; Regal, R. R. On the intrinsic dimensionality of chemical structure space. *Chemosphere* **1988**, *17*, 1617–1630.
- (9) McFarland, J. W. Parabolic relation between drug potency and hydrophobicity. *J. Med. Chem.* **1970**, *13*, 1192–1196.
- (10) Dearden, J. C.; Netzeva, T. I.; Bibby, R. Comparison of a number of commercial software programs for the prediction of octanol–water partition coefficient. *J. Pharm. Pharmacol.* **2002**, *54* (Suppl.), S65–66.
- (11) Benfenati, E.; Gini, G.; Piclin, N.; Roncaglioni, A.; Vari, M. R. Predicting logP of pesticides using different software. *Chemosphere* **2003**, *53*, 1155–1164.
- (12) Mekenyan, O. G.; Veith, G. D. Relationships between descriptors for hydrophobicity and soft electrophilicity in predicting toxicity. *SAR QSAR Environ. Res.* **1993**, *1*, 335–344.
- (13) Benfenati, E.; Piclin, N.; Roncaglioni, A.; Vari, M. R. Factors influencing predictive models for toxicology. *SAR QSAR Environ. Res.* **2001**, *12*, 593–603.
- (14) Seward, J. R.; Cronin, M. T. D.; Schultz, T. W. The effect of precision of molecular orbital descriptors on toxicity modelling of selected pyridines. *SAR QSAR Environ. Res.* **2002**, *13*, 325–340.
- (15) Cronin, M. T. D.; Manga, N.; Seward, J. R.; Sinks, G. D.; Schultz, T. W. Parametrization of electrophilicity for the prediction of the toxicity of aromatic compounds. *Chem. Res. Toxicol.* **2001**, *14*, 1498–1505.
- (16) Schüürmann, G. Quantum chemical descriptors in structure–activity relationships – calculation, interpretation and comparison of methods. In *Predicting Chemical Toxicity and Fate*; Cronin, M. T. D., Livingstone, D. J., Eds.; Taylor and Francis: London, 2004; pp 85–149.
- (17) Curtiss, L. A.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and density function theories for the computation of ionisation potential and electron affinities. *J. Chem. Phys.* **1998**, *109*, 42–55.
- (18) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *99*, 4899–4907.
- (19) Alemán, C.; Luque, F. J.; Orozco, M. Suitability of the PM3-derived molecular electrostatic potentials. *J. Comput. Chem.* **1993**, *14*, 799–808.
- (20) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for reliability, uncertainty assessment,



- and applicability evaluations of classification and regression based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361–1375.
- (21) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ. Toxicol. Chem.* **1997**, *16*, 948–967.
  - (22) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537–547.
  - (23) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: atomic model builders. *Chem. Rev.* **1993**, *93*, 2567–2581.
  - (24) SIMCA-P 9. User Guide and Tutorial. Umetrics AB, Umea, Sweden, Appendix A, p 275.
  - (25) Cronin, M. T. D. Toxicological information for use in predictive modelling: quality, sources, and databases. In *Predictive Toxicology*; Helma, C., Ed.; Marcel Dekker Inc.: New York, 2005; in press.
  - (26) Cronin, M. T. D.; Netzeva, T. I.; Dearden, J. C.; Edwards, R.; Worgan, A. D. P. Assessment and modelling of the toxicity of organic chemicals to *Chlorella vulgaris*: development of a novel database. *Chem. Res. Toxicol.* **2004**, *17*, 545–554.
  - (27) Nendza, M.; Russom, C. L. QSAR modeling of the fathead minnow acute toxicity database. *Xenobiotica* **1991**, *21*, 147–170.
  - (28) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Use of graph-theoretical parameters in risk assessment of chemicals. *Toxicol. Lett.* **1995**, *79*, 239–250.
  - (29) Eldred, D. V.; Weikel, C. L.; Jurs, P. C.; Kaiser, K. L. E. Prediction of fathead minnow acute toxicity of organic compounds from molecular structure. *Chem. Res. Toxicol.* **1999**, *12*, 670–678.
  - (30) Klopman, G.; Saiakhov, R.; Rosenkranz, H. S. Multiple computer-automated structure evaluation study of aquatic toxicity II. Fathead minnow. *Environ. Toxicol. Chem.* **2000**, *19*, 441–447.
  - (31) Schultz, T. W.; Cronin, M. T. D.; Netzeva, T. I.; Aptula, A. O. Structure-toxicity relationships for aliphatic chemicals evaluated with *Tetrahymena pyriformis*. *Chem. Res. Toxicol.* **2002**, *15*, 1602–1609.
  - (32) Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. D. Selection of data sets for QSARs: analyses of *Tetrahymena* toxicity from aromatic compounds. *SAR QSAR Environ. Res.* **2003**, *14*, 59–81.
  - (33) Netzeva, T. I.; Schultz, T. W.; Aptula, A. O.; Cronin, M. T. D. Partial least squares modelling of the acute toxicity of aliphatic compounds to *Tetrahymena pyriformis*. *SAR QSAR Environ. Res.* **2003**, *14*, 265–283.
  - (34) Schüürmann, G. Ecotoxic modes of action of chemical substances. In *Ecotoxicology*; Schüürmann, G., Markert, B., Eds.; John Wiley and Spektrum Akademischer Verlag: New York, 1998; pp 665–749.

CI049747P