Calculation of Quantum-Mechanical Descriptors for QSPR at the DFT Level: Is It Necessary?

Tomasz Puzyn,*,† Noriyuki Suzuki,† Maciej Haranczyk,‡ and Janusz Rak‡

Exposure Assessment Research Section, Research Center for Environmental Risk, National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, Ibaraki 305-8506 Japan, and Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland

Received January 18, 2008

Most of the recently published quantitative structure−property relationship (QSPR) models, which can be used to predict environmentally relevant physicochemical data for persistent organic pollutants (e.g., polychlorinated dibenzo-*p*-dioxins, dibenzofurans, and biphenyls), employ molecular descriptors obtained by means of relatively costly calculations at the density functional theory (DFT) level. However, new semiempirical methods, PM6 and RM1, have recently been developed by J. J. P. Stewart's group. In this study, we compared various QSPR models based on DFT (B3LYP functional) descriptors with the same models based on semiempirical (PM6 and RM1) descriptors. We recalibrated 10 previously published models (for different properties and groups of congeneric compounds) employing PM6 and RM1 descriptors instead of B3LYP ones. We demonstrated that by applying RM1 and PM6 descriptors, we could obtain QSPR models with quality similar to that of models based on B3LYP descriptors. This level of accuracy was out of reach for the models employing AM1- and PM3-based descriptors.

INTRODUCTION

Quantitative structure−property relationship (QSPR) methods are among the most practical tools in computational physical chemistry. These methods are based on the axiom that the variance in the physicochemical properties of chemical compounds is determined by the variance in their molecular structures. Thus, if experimental data are available for only some chemicals in a group, one can predict the missing data from molecular descriptors calculated for the whole group and a suitable mathematical model.[1–3]

The method has found many applications. One of the most prominent examples involves the use of calculated physicochemical data to assess the environmental risk of new, potentially hazardous substances. Many reports present QSPR models for predicting important properties such as water solubility, supercooled liquid−vapor pressure, partition coefficients, and environmental half-lives for several types of environmental pollutants.[4–8]

A common problem in QSPR modeling is choosing a proper description of the variance between the individual molecular structures within a set of compounds. The choice is especially important for groups of structurally similar congeners; 'congeners' are defined as compounds having the same carbon skeleton but differing substitution patterns (e.g., 1-chlorodibenzofuran, 1,4-dichlorodibenzofuran, 1,4,7-trichlorodibenzofuran).[9,10] Well-known examples of congeneric groups identified as persistent organic pollutants (POPs) are polychlorinated dibenzo-*p*-dioxins (PCDDs), polychlorinated dibenzofurans (PCDFs), polychlorinated naphthalenes (PCNs), and polychlorinated biphenyls (PCBs).[10]

Because the compounds in these groups are highly similar, the relative differences between the descriptor values for the congeners are usually small. Therefore, the descriptors must be determined as precisely as possible; the error in the calculated descriptor value must be significantly lower than the real variance of that descriptor between the congeners.[10]

Because the number of congeners in a typical congeneric family varies between about 70 and 4000, experimental measurement of the physicochemical properties for all the congeners is often practically impossible. However, precise and complete data are necessary for a comprehensive estimation of environmental transport routes and exposure levels and for the final risk assessment.[11,12] Experimental studies on large families of congeners are impractical, whereas computational studies can be automated with specialized software. Recently, we developed ConGENER,[13,14] a software package that combinatorially generates families of congeners and facilitates their characterization with quantum chemistry software packages. ConGENER is one of applications of the hybrid quantum mechanical-combinatorial approach developed by one of us.[15]

Several types of molecular descriptors of varying sophistication exist, including molecular weight, number of substituent groups, and graph-based connectivity indices.[16–20] However, many researchers have highlighted the usefulness of quantum-mechanical (QM) descriptors in the modeling of water solubility,[10,21–24] vapor pressure,[21,25–28] and partition coefficients for POPs.[21,29–34] QM descriptors, such as mean polarizability (α), dipole moment (μ), largest negative partial charge (q^-), and energies of the highest occupied and lowest unoccupied molecular orbitals (ϵ_{HOMO} and ϵ_{LUMO}, respectively), can be calculated at different levels of theory, including density functional theory (DFT) methods[10,22–28,32–34] and semiempirical methods.[21,29–31]

* Corresponding author phone: +81-298-50-2888; fax: +81-298-50-2920; e-mail: puzyn.tomasz@nies.go.jp.
† National Institute for Environmental Studies.
‡ University of Gdańsk.

10.1021/ci800021p CCC: $40.75
Published on Web 05/30/2008

The authors of several studies published from 2004 to 2007 concluded that the application of QM descriptors from DFT (B3LYP functional) calculations permits the description of even very small variances between individual congeners within a congeneric group, which is impossible when the "standard" semiempirical descriptors (based on such popular methods as PM3 and AM1) are used.[22–24,27,28,32–34] However, Stewart et al. published two new semiempirical methods (RM1 and PM6) in 2006 and 2007 suggesting that they should be much more precise than the previously used AM1, PM3, and even PM5 ones.[35,36] Because semiempirical methods are, in general, several orders of magnitude faster than DFT methods,[37] we decided to investigate the precision and applicability of these new RM1 and PM6 methods in QSPR-like studies for various congeners.

NEW SEMIEMPIRICAL METHODS (RM1, PM6)

Modified neglect of differential overlap (MNDO)-type semiempirical methods are much better suited for QSPR models than are any ab initio methods. Even for the Hartree−Fock−Roothan model, the least time-consuming ab initio method, the computational effort scales as N^4, where N is the number of basis functions.[37] In contrast, for MNDO-type methods, which employ the zero differential overlap approximation, the effort related to the construction of the Fock matrix decreases from N^4 to N^2. Moreover, semiempirical methods involve only a minimum basis set, whereas ab initio methods, especially those accounting for correlation effects, require extended basis sets (perhaps of triple-ζ quality with several sets of polarization functions and many diffuse functions) to ensure a reliable accuracy. As a result, the semiempirical wave function can be calculated for a molecular system comprising more than 100,000 atoms at the cost of only several hours on a personal computer,[38] a task that is impractical at any ab initio level.

The approximations and simplifications of semiempirical methods come at a price, however. In the past, the most serious drawbacks, drawbacks that sometimes made usage of the methods questionable, were relatively low accuracy and a number of artifacts related to the parametrization. However, with the advent of new modifications of neglect of diatomic differential overlap (NDDO) approximations, such as the RM1[35] and PM6 modifications,[36] this disadvantage seems to have been at least partially overcome. The fact that the RM1 method is a reparameterization of AM1 for six elements (C, H, N, O, P, and S) and the halogens enables the vast majority of biologically important molecules to be modeled. This method uses the same approximations as AM1; the main difference between the two methods lies in that fact that all RM1 parameters are optimized, unlike the case for AM1 and similar to the case for PM3. In consequence, the average error in the enthalpies of formation for a training set of 1736 organic molecules was reduced from 11.15 kcal/mol at the AM1 level to only 5.77 kcal/mol for the RM1 model.

The PM6 model introduced several modifications to the NDDO core−core interaction term and to the method of parameter optimization.[36] This new model is statistically even better than its predecessor, RM1. The PM6 average unsigned error (AUE) between the calculated and reference heats of formation for 1373 molecules consisting of biologically relevant elements is only 4.4 kcal/mol. The equivalent AUEs for other methods are as follows: RM1, 5.0; B3LYP 6−31G(d), 5.2; PM5, 5.7; PM3, 6.3; HF 6−31G(d), 7.4; and AM1, 10.0 kcal/mol. These comparisons indicate that both PM6 and RM1 perform better than the quite reliable but much more time-consuming B3LYP method.

These results prompted us to reevaluate various published QSPR models based on B3LYP descriptors, employing the same descriptors but estimating them at the RM1 and PM6 levels. To proceed with such an analysis, we carried out unconstrained RM1 and PM6 geometry optimizations of the molecules studied, using the eigenvector following (EF) optimization procedure. The final gradient norm of the energy gradient was always less than 0.1 kcal/mol. The nature of stationary points was revealed with calculations of matrices of second derivatives of energy with respect to the nuclear coordinates (Hessians). We verified that for all localized minima, exclusively positive eigenvalues appeared in the Hessian matrix. The calculated semiempirical wave functions enabled the appropriate descriptors to be evaluated. Five such descriptors were employed within our analysis: the energies of the highest occupied (ϵ_{HOMO}) and lowest unoccupied (ϵ_{LUMO}) molecular orbitals, the values of the dipole moment (μ) and the mean polarizability (α), and the largest Mulliken negative charge (q^-).

METHODOLOGY OF THE COMPARISON BETWEEN DFT (B3LYP) AND THE SEMIEMPIRICAL (RM1, PM6) METHODS ACCOMPLISHED IN THIS STUDY

In the first stage of the study, we searched for 10 QSPR reference models published in peer-reviewed journals, models that predicted different environmentally relevant physicochemical properties for POPs using DFT quantum-mechanical descriptors. We chose the B3LYP functional because it is the most commonly used functional in QSPR studies. Moreover, we decided to look for models employing the descriptors calculated with Pople-style basis sets of different sizes, including 6−31G(d), 6−311G(d,p), and 6−311++G(d,p). The models we selected also represented different ways to design the applicability domain: from a domain covering only three or four groups of molecules containing the same number of chlorine substituents (i.e., dichlorobiphenyls), through a domain consisting of a single congeneric family (i.e., PCDDs), to a domain consisting of even two similar families of compounds (PCDDs and PCDFs). In addition, to make our comparisons not influenced by the modeling technique, we selected only models originally obtained by means of one algorithm−multiple linear regression.

Next, we recalibrated each of the literature-derived DFT models with the same descriptors calculated at the semiempirical level (PM6 and RM1). We applied the standard four-step QSPR procedure: (i) collection of experimental data, (ii) molecular modeling, (iii) construction of a QSPR model, and (iv) validation of the QSPR model.

The comparison of the original (literature-derived) and recalibrated models was based on the following seven measures: the determination coefficient (R^2), the cross-validation coefficient (Q^2_{CV}), the external validation coefficient (Q^2_{Ext}), the standard deviation (SD), the root-mean-square error of the leave-one-out cross-validation ($RMSE_{CV}$), the root-mean-square error of prediction (based on an external

validation ($RMSE_{Ext}$)), and the individual regression coefficients in the model's equation.

R^2 and SD are the two measures of the goodness-of-fit, and they do not provide information on the predictive ability of a QSPR model. According to the accepted standards, a model should be appropriately validated to confirm its predictivity. The validation can be internal, wherein the same compounds are used for calibration and validation of the model (of course, not at the same time), or external, wherein the model is calibrated on one (training) set of compounds and then validated with a second (validation) set, not previously used for calibration. Internal validation also gives information on robustness, that is, it demonstrates how sensitive the model is to whether individual elements are included in or excluded from the training set.

One of the most commonly applied internal validation techniques is the leave-one-out cross-validation (LOO-CV). In each cycle of this procedure, one compound from the training set is excluded and the model is parametrized for $n - 1$ compounds. Then the residual value (the difference between the observed and predicted values) is calculated for the excluded compound. Subsequently, this compound is included in the training set, and a different compound is excluded. The procedure is repeated for all the compounds in the training set. The measures connected to internal validation by LOO-CV, Q^2_{CV}, and $RMSE_{CV}$ are defined as follows

$$Q^2_{cv} = 1 - \frac{\sum_{i=1}^{n} \left(y_i^{obs} - y_i^{predcv}\right)^2}{\sum_{i=1}^{n} \left(y_i^{obs} - \bar{y}^{obs}\right)^2} \quad (1)$$

$$RMSE_{CV} = \sqrt{\frac{\sum_{n=1}^{n} \left(y_i^{obs} - y_i^{predcv}\right)^2}{n}} \quad (2)$$

where y_i^{obs} is the experimental (observed) value of the property for the ith compound; y_i^{predcv} is the predicted value for the temporary excluded (cross-validated) ith compound; $\bar{y}^{obs}$ is the mean experimental value of the property in the training set; and n is the number of compounds in the training set.

Similarly, the measures of predictive ability based on external validation, Q^2_{Ext} and $RMSE_{Ext}$, are defined by eqs 3 and 4

$$Q^2_{Ext} = 1 - \frac{\sum_{j=1}^{k} \left(y_j^{obs} - y_j^{pred}\right)^2}{\sum_{j=1}^{k} \left(y_j^{obs} - \bar{y}^{obs}\right)^2} \quad (3)$$

$$RMSE_{Ext} = \sqrt{\frac{\sum_{j=1}^{k} \left(y_j^{obs} - y_j^{pred}\right)^2}{k}} \quad (4)$$

where y_j^{obs} is the experimental (observed) value of the property for the jth compound; y_j^{pred} is the predicted value for the jth compound and k is the number of compounds in the validation set.[39]

To make our study more directly comparable, we used the same experimental data as the original authors, and we used exactly the same method for splitting the data into training and validation sets in each case (for more details, refer to the Supporting Information). Moreover, we validated the models in the same way, even if the original authors made only an internal validation by the LOO-CV technique (no external validation).

RESULTS AND DISCUSSION

Review of the Original Models. Three of the selected 10 models predicted the logarithms of supercooled liquid−vapor pressure (log P_L),[26–28] three more predicted water solubility (–log S_W),[22–24] two predicted n-octanol/water partition coefficients (log K_{OW}),[22,33] and two predicted n-octanol/air partition coefficients (log K_{OA})[26,32] (Table 1). The original models used five quantum-mechanical descriptors (α, mean polarizability; μ, dipole moment; q^-, highest negative Mulliken partial charge on atoms; ϵ_{HOMO}, energy of the highest occupied molecular orbital; and ϵ_{LUMO}, energy of the lowest unoccupied molecular orbital) in various combinations.

Note that some of the quality measures were not given by the authors, so we had to calculate them additionally from the original data. For instance, Staikova et al.[26] reported that the models predicting log P_L and log K_{OA} (models 2 and 9; Table 2) were calculated separately for different groups, including chlorobiphenyls, and that for each of the individual compound classes, the values of R^2 were satisfactory. However, because the model equation and other statistics were not given, we repeated those calculations using the

Table 1. Characteristics of the 10 Reference DFT Models Collected from the Literature

no	predicted property	domain[a]	DFT method/ basis set	descriptors	no. of compds: training (n)	no. of compds: validation (k)	ref
1	log P_L	PCDEs	B3LYP/6−31G(d)	q^-, μ, α	72	35	28
2	log P_L	monoCNs - triCNs	B3LYP/6−311G(d,p)	α	25	-	26
3	log P_L	PCNs	B3LYP/6−311++G(d,p)	α	10	5	27
4	–log S_W	PCDDs	B3LYP/6−31G(d)	α	12	-	22
5	–log S_w	PCDDs & PCDFs	B3LYP/6−31G(d)	α, μ	21	-	23
6	–log S_w	PCBs	B3LYP/6−31G(d)	α, q^-	133	-	24
7	log K_{OW}	PCDDs	B3LYP/6−31G(d)	α	41	-	22
8	log K_{OW}	PCBs	B3LYP/6−31G(d)	ϵ_{HOMO}, ϵ_{LUMO}	135	-	33
9	log K_{OA}	monoCBs-tetraCBs	B3LYP/6−311G(d,p)	α	38	-	26
10	log K_{OA}	PCNs	B3LYP/6−311++G(d,p)	α	30	13	32

[a] For a detailed list of the compounds, please refer to the Supporting Information.

Table 2. Comparison of the Models Derived from the Literature Descriptors (DFT) and from the Recalibrated Descriptors (Semiempirical)

no.	method	equation	R^2	Q^2_{CV}	Q^2_{Ext}	SD	$RMSE_{CV}$	$RMSE_{Ext}$
1.	B3LYP/6−31G(d)[a]	$\log P_L = 12.29 + 8.90\ q^- - 0.12\ \mu - 0.06\ \alpha$	0.988	0.986	0.982[c]	0.13	0.14	0.16[d]
	PM6[b]	$\log P_L = 6.94 + 0.82\ q^- - 0.10\ \mu - 0.04\ \alpha$	0.985	0.983	0.986[c]	0.15	0.15	0.19[d]
	RM1[b]	$\log P_L = 9.46 + 2.86\ q^- - 0.12\ \mu - 0.08\ \alpha$	0.989	0.987	0.978[c]	0.13	0.14	0.17[d]
2.	B3LYP/6−311G(d,p)[a]	$\log P_L = 6.3945 - 0.3457\ \alpha$ [c]	0.870[c]	0.835[c]	-	0.18[e]	0.19[e]	-
	PM6[b]	$\log P_L = 6.1289 - 0.2890\ \alpha$	0.890	0.860	-	0.16	0.18	-
	RM1[b]	$\log P_L = 7.6026 - 0.5263\ \alpha$	0.858	0.820	-	0.18	0.20	-
3.	B3LYP/6−311++G(d,p)[a]	$\log P_L = 6.08 - 0.30\ \alpha$	0.989	0.980	0.979	0.17	0.20	0.19
	PM6[b]	$\log P_L = 5.63 - 0.28\ \alpha$	0.989	0.981	0.982	0.16	0.19	0.18
	RM1[b]	$\log P_L = 6.98 - 0.49\ \alpha$	0.986	0.976	0.977	0.18	0.21	0.20
4.	B3LYP/6−31G(d)[a]	$-\log S_W = -3.6425 + 0.0693\ \alpha$	0.978	0.978	-	0.30	0.31	-
	PM6[b]	$-\log S_W = -3.4016 + 0.0563\ \alpha$	0.977	0.962	-	0.33	0.39	-
	RM1[b]	$-\log S_W = -6.3637 + 0.1012\ \alpha$	0.976	0.960	-	0.34	0.31	-
5.	B3LYP/6−31G(d)[a]	$-\log S_W = -1.8013 - 0.1367\ \mu + 0.0602\ \alpha$	0.940	0.915	-	0.40	0.51[c]	-
	PM6[b]	$-\log S_W = -1.9889 - 0.1872\ \mu + 0.0514\ \alpha$	0.954	0.937	-	0.43	0.47	-
	RM1[b]	$-\log S_W = -4.6709 - 0.176\ \mu + 0.0921\ \alpha$	0.955	0.938	-	0.42	0.46	-
6.	B3LYP/6−31G(d)[a]	$-\log S_W = -0.047 + 0.049\ \alpha + 8.239\ q^-$	0.944	0.940	-	0.27	0.27[c]	-
	PM6[b]	$-\log S_W = -1.297 + 0.040\ \alpha - 1.161\ q^-$	0.943	0.941	-	0.28	0.28	-
	RM1[b]	$-\log S_W = -2.943 + 0.075\ \alpha - 1.056\ q^-$	0.951	0.949	-	0.26	0.26	-
7.	B3LYP/6−31G(d)[a]	$\log K_{OW} = 0.3909 + 0.0334\ \alpha$	0.866	0.853	-	0.27	0.28	-
	PM6[b]	$\log K_{OW} = 0.5175 + 0.0271\ \alpha$	0.865	0.851	-	0.28	0.28	-
	RM1[b]	$\log K_{OW} = -0.9626 + 0.0492\ \alpha$	0.868	0.854	-	0.27	0.28	-
8.	B3LYP/6−31G(d)[a]	$\log K_{OW} = -3.073 - 18.704\ \epsilon_{HOMO} - 34.076\ \epsilon_{LUMO} + 0.017\ \alpha$	0.948	0.946	-	0.18	0.18[c]	-
	PM6[b]	$\log K_{OW} = 1.128 - 2.011\ \epsilon_{HOMO} - 37.545\ \epsilon_{LUMO} + 0.014\ \alpha$	0.935	0.939	-	0.20	0.20	-
	RM1[b]	$\log K_{OW} = 3.018 + 5.872\ \epsilon_{HOMO} - 31.122\ \epsilon_{LUMO} + 0.031\ \alpha$	0.939	0.934	-	0.20	0.20	-
9.	B3LYP/6−311G(d,p)[a]	$\log K_{OW} = -1.3487 + 0.3995\ \alpha$[c]	0.908[e]	0.898[e]	-	0.22[e]	0.22[e]	-
	PM6[b]	$\log K_{OW} = -0.0990 + 0.3058\ \alpha$	0.839	0.820	-	0.29	0.30	-
	RM1[b]	$\log K_{OW} = -2.5286 + 0.5748\ \alpha$	0.880	0.867	-	0.25	0.26	-
10.	B3LYP/6−311++G(d,p)[a]	$\log K_{OA} = -1.17 + 0.37\ \alpha$	0.968	0.963	0.933	0.24	0.24	0.33
	PM6[b]	$\log K_{OA} = -0.59 + 0.34\ \alpha$	0.973	0.969	0.937	0.22	0.22	0.32
	RM1[b]	$\log K_{OA} = -2.24 + 0.59\ \alpha$	0.966	0.962	0.934	0.24	0.24	0.32

[a] The reference models collected from the literature. For more detail description, please refer to Table 1 and the Supporting Information. [b] Models calculated in this study. The training (and validation) sets used in the corresponding literature-derived models were used in each case. [c] The correlation coefficient of the validation model (a model for an external validation set). [d] The error of cross-validation for the external validation model. [e] The value was not given by the authors of the original paper. It was calculated within this study based on the original data.

originally reported experimental data and the values of mean polarizabilities.

In addition, the authors applied different model validation techniques. Models 2 and 9 (Table 2) were not validated at all, so we calculated Q^2 and $RMSE_{CV}$ for them. Models 4, 6, and 7 (Table 2) were only internally validated by the LOO-CV technique. Models 5 and 8 (Table 2) were validated internally (LOO-CV) first, and then additional "validation" models were calculated for some compounds. However, according to our understanding, the additional "validation" models were calculated on the basis of compounds taken from the previously used calibration sets, so there actually were no external validations, only internal validations. Therefore, we decided to use only Q^2 and $RMSE_{CV}$ for further comparisons.

Only models 1, 3, and 10 (Table 2) were validated with a fully external set of compounds, but the validation strategies differed. For models 3 and 10, the molecular descriptors (polarizabilities) were used for calculation of, respectively, $\log P_L$ and $\log K_{OW}$ for the validating compounds, from appropriate model equations, previously calibrated with the use of the training compounds. In model 1, however, the validation compounds were applied for calibration of a new validation model. In consequence, the original authors did not calculate the values of Q^2_{Ext} and $RMSE_{Ext}$. In that case, we decided to use Q^2_{CV} and $RMSE_{CV}$ of this new validation model as the quasi-external measures of the original model's predictivity. However, those values should be interpreted with care.

Comparison between the Methods. The most surprising and important observation from the comparison of the models (Table 2) is that there were practically no significant differences between the quality measures of the corresponding models. In most cases, the differences for the correlation coefficients (R^2, Q^2_{CV}, and Q^2_{Ext}) and the errors (SD, $RMSE_{CV}$, and $RMSE_{Ext}$) were in the second or third decimal place. The exceptions were model 9, where $0.880 < R^2 < 0.908$ and $0.22 < RMSE_{CV} < 0.30$, and model 5, where $0.46 < RMSE_{CV} < 0.51$. However, even for those cases, the ranges of quality measures were still acceptable. These results indicate that it is possible to construct PM6- or RM1-based QSPR models of quality similar to that of models obtained with B3LYP descriptors. It is also important that the same phenomena was observed for models with differently specified applicability domains (e.g., PCDDs only and PCDDs/Fs together) and for B3LYP-based models employing different basis sets (e.g., 6−31G(d), 6−311G(d,p), and 6−311++G(d,p)). This indicates that the results of the study do not depend on the design of the applicability domain or on the size of the basis set used in the DFT calculations.

The comparison of the corresponding coefficients in the equations describing individual models is interesting. We

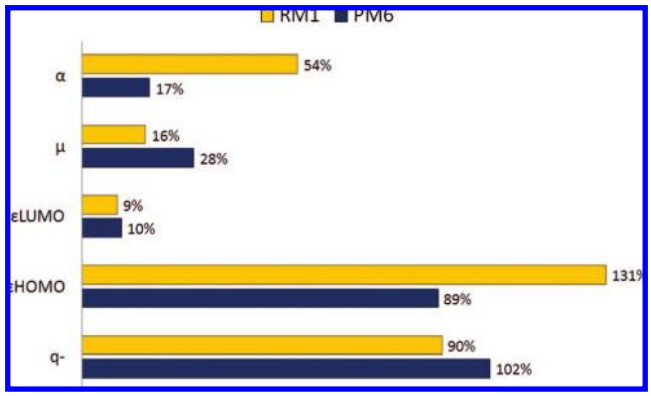

Figure 1. Mean values of the relative differences (%) between the coefficients from the models based on the B3LYP descriptors and the corresponding models using descriptors derived from the two semiempirical methods.

calculated the relative differences (in percentages) between the coefficients from the models based on the B3LYP descriptors and the corresponding models based on the descriptors derived from the two semiempirical methods. Then, because differences between the coefficients (scalars) are related to differences in descriptors (vectors), it was possible to calculate a mean relative "error" for each descriptor (Figure 1). Finally, we should point out that such descriptors as ϵ_{HOMO} and q^- are definitely less "robust" in the QM method of calculation than are α, μ, and ϵ_{LUMO} (the relative differences between the DFT and semiempirical methods are significantly higher in the first case). However, even if there is a higher systematic error in the ϵ_{HOMO} and q^- calculated by means of the PM6 and RM1 methods, one can still use these descriptors for QSPR modeling and get a model with satisfactory statistical characteristics of quality (R^2, Q^2_{CV} and Q^2_{Ext}, SD, $RMSE_{CV}$, and $RMSE_{Ext}$). But owing to this systematic error, especially for models including ϵ_{HOMO} and q^-, predictions should be made with care, using only descriptors calculated by the method that was applied for the original training set.

Judging whether the PM6 or RM1 method gives better results is difficult. For instance, the mean polarizabilities obtained with PM6 were closer to the B3LYP values than those obtained with RM1 (Figure 1), but the dipole moments from RM1 were closer to those from B3LYP than were those from PM6. Therefore, we strongly recommend that QSPR models based on both semiempirical methods always be tested and that the model that gives better statistics be used. Both methods (PM6 and RM1) notably seem to represent important milestones in the development of quantum-mechanical descriptors for QSPR.

There are still many published papers presenting QSPR models for POPs based on older semiempirical methods such as AM1 and PM3.[21,31] The main argument of adherents to the semiempirical methods is that such methods are relatively cheap in comparison to DFT and a purely ab initio scheme.[37,38] But, as was mentioned earlier, some authors using DFT descriptors demonstrated the high precision of their approach relative to that of semiempirical methods. Because experimental data are lacking, comprehensive and direct comparison of the quality of the descriptors from both DFT and semiempirical approaches for the compounds investigated within this study was impossible. We have found only data on experimentally measured dipole moments. In addition, those data were available only for mono- and diCNs.[40] However, we decided to use even those incomplete data to cast some light on the problem of precision. The dipole moment is related to the calculated wave function; therefore it seems to be an appropriate, relative measure of the precision. We combined the experimental data and results from additional calculations performed at five levels of the theory: B3LYP/6−31G(d), B3LYP/6−311++G(d,p), PM3, PM6, and RM1. The quality of the two new semiempirical methods was substantially improved over that of PM3 (Figure 2). The dipole moment profiles obtained with the two new methods were close to those obtained by DFT, although there were some problems for individual congeners. For example, the dipole moment of 1,8-dichloronaphthalene calculated by both PM6 and RM1 was slightly lower than that for 2,3-dichloronaphthalene,

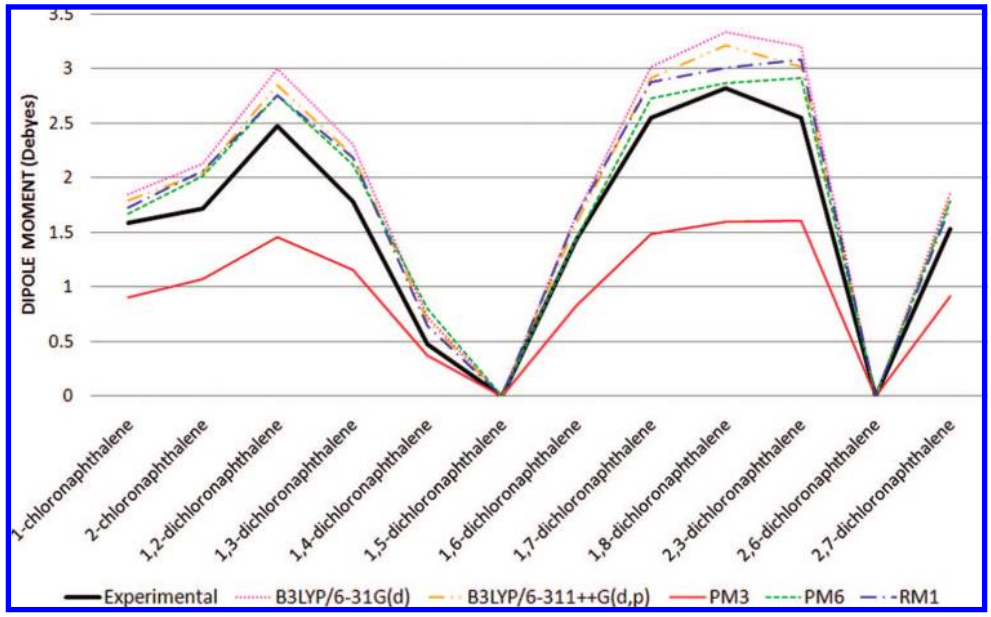

Figure 2. Dipole moments (in Debyes) for mono- and dichloronaphthalenes (experimentally measured and calculated by the five QM methods). Experimental data were taken from ref 40.

whereas the experiment and the DFT calculations gave opposite results.

The calculations for 1,8-dichloronaphthalene at the B3LYP/6−311++G(d,p) level took about 1.5 h on two processors (geometry optimization and frequency analysis), whereas the same tasks by PM6 and RM1 methods required only 2.5 and 1.2 min, respectively, on a single processor. Therefore, if one works with a very large set of congeners and if some minimal differences between them are negligible, one can expect a relatively good QSPR model when PM6 and RM1 descriptors are used.

In this contribution we have investigated only the applicability of the quantum-mechanical descriptors and the most common methods of their calculation. However, the next logical step would be to compare the efficiency between the QSPR models based on the semiempirical methods and the models using simpler, even two-dimensional (2D) or one-dimensional (1D) descriptors. For instance, in our previous studies[10,27,29,32] we have observed that some of the energetic parameters (i.e., ϵ_{HOMO} or the ionization potential) in a group of congeners were strongly correlated to the substitution pattern. Similar suggestions have been made also by the other authors,[24,28,33,41] who expressed the pattern by a set of simple topological descriptors. Therefore, if similar predictive ability characterizes both types of the QSPR models for congeners, those developed with the 1D or 2D descriptors and those based on the descriptors from the RM1 and PM6 calculations, then is it necessary to apply the quantum-mechanical descriptors at all?

CONCLUSION

We strongly recommend that QSPR studies are carried out with PM6 and RM1 descriptors instead of the much more expensive DFT descriptors. The use of semiempirical descriptors will allow the researchers to obtain QSPR models for thousands of new chemicals. These models are of similar quality to DFT-based models but can be built in a relatively short time. It must be noted that the comparisons presented here were made only for congeneric sets of selected POPs. Our future work will extend the comparison to other groups of substances as well as different types of descriptors.

ACKNOWLEDGMENT

T.P. thanks the Japan Society for the Promotion of Science (JSPS) for a Postdoctoral Fellowship for Foreign Researchers and the National Institute for Environmental Studies for hosting him as the JSPS fellow. M.H. holds an award from the Foundation for the Development of the University of Gdansk (FRUG). This work was partially supported by the Polish State Committee for Scientific Research (KBN) (grants DS/8221-4-0140-8 (J.R.) and N204 127 31/2963 (M.H.)).

Supporting Information Available: Detailed lists of the studied compounds, the ways of splitting data into the training and validation sets, the values of molecular descriptors, and the experimental and predicted values of the physicochemical properties. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) 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.

(2) Karcher, W.; Devillers, J. SAR and QSAR in the environmental chemistry and toxicology: scientific tool or wishful thinking? In *Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology*; Karcher, W., Devillers, J., Eds.; Kluwer Academic Publishers: Dordrecht, Boston, London, United Kingdom, 1990; pp 1−12.

(3) Katritzky, A. R.; Petrukhin, R.; Tatham, D.; Basak, S.; Benfenati, E. Interpretation of Quantitative Structure-Property and -Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 679–685.

(4) Katritzky, A. R.; Wang, Y. L.; Sild, S.; Tamm, T.; Karelson, M. QSPR studies on vapor pressure, aqueous solubility, and the prediction of water-air partition coefficients. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720–725.

(5) Gramatica, P.; Papa, E. Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure. *Environ. Sci. Technol.* **2007**, *41*, 2833–2839.

(6) Gramatica, P.; Pilutti, P.; Papa, E. QSAR prediction of ozone tropospheric degradation. *QSAR Comb. Sci.* **2003**, *22*, 364–373.

(7) Kahn, I.; Fara, D.; Karelson, M.; Maran, U.; Andersson, P. L. QSPR treatment of the soil sorption coefficients of organic pollutants. *J. Chem. Inf. Model.* **2005**, *45*, 94–105.

(8) Yan, A.; Gasteiger, J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429–434.

(9) Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSAR. *J. Mol. Struct. - THEOCHEM* **2003**, *622*, 39–51.

(10) Puzyn, T.; Mostrąg, A.; Falandysz, J.; Kholod, Y.; Leszczynski, J. Predicting water solubility for congeners: Chloronaphthalenes - a case study. *Chemom. Intell. Lab. Syst*. 2008, submitted for publication.

(11) Guidance document on the use of multimedia models for estimating overall environmental persistence and long-range transport. OECD series on testing and assessment No. 45; Organisation for Economic Co-operation and Development: Paris, France, 2004.

(12) Fenner, K.; Scheringer, M.; MacLeod, M.; Matthies, M.; McKone, T. E.; Stroebe, M.; Beyer, A.; Bonnell, M.; LeGall, A.-C.; Klasmeier, J.; Mackay, D.; van de Meent, D.; Pennington, D.; Scharenberg, B.; Suzuki, N.; Wania, F. Comparing estimates of persistence and long-range transport potential among multimedia models. *Environ. Sci. Technol.* **2005**, *39*, 1932–1942.

(13) Haranczyk, M.; Puzyn, T.; Sadowski, P. ConGENER - a tool for modeling of the congeneric sets of environmental pollutants. *QSAR Comb. Sci*. 2008, in press (DOI: 10.1002/qsar.200710149).

(14) Haranczyk, M. ConGENER, version 1.0. Available free of charge at http://congener.sourceforge.net (accessed Feb 28, 2008).

(15) Haranczyk, M.; Gutowski, M. Quantum mechanical energy-based screening of combinatorially generated library of tautomers. TauTGen: a tautomer generator program. *J. Chem. Inf. Model.* **2007**, *47*, 686–694.

(16) Gramatica, P.; Navas, N.; Todeschini, R. 3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs). *Chemomom. Intell. Lab. Syst.* **1998**, *40*, 53–63.

(17) Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; Wiley-VCH Verlag: Weinheim, 2000.

(18) Toropov, A. A.; Leszczynska, D.; Leszczynski, J. Predicting water solubility and octanol water partition coefficient for carbon nanotubes based on the chiral vector. *Comput. Biol. Chem.* **2007**, *31*, 127–128.

(19) Katritzky, A. R.; Kulshyn, O. V.; Stoyanova-Slavova, I.; Dobchev, D. A.; Kuanar, M.; Fara, D. C.; Karelson, M. Antimalarial activity: a QSAR modeling using CODESSA PRO software. *Bioorg. Med. Chem.* **2006**, *14*, 2333–2357.

(20) Kuz'min, V. E.; Artemenko, A. G.; Polischuk, P. G.; Muratov, E. N.; Hromov, A. I.; Liahovskiy, A. V.; Andronati, S. A.; Makan, S. Y. Hierarchic system of QSAR models (1D-4D) on the base of simplex representation of molecular structure. *J. Mol. Model.* **2005**, *11*, 457–467.

(21) Yang, P.; Chen, J.; Chen, S.; Yuan, X.; Schramm, K. W.; Kettrup, A. QSPR models for physicochemical properties of polychlorinated diphenyl ethers. *Sci. Total Environ.* **2003**, *305*, 65–76.

(22) Yang, G. Y.; Yu, J.; Wang, Z. Y.; Zeng, X. L.; Ju, X. H. QSPR study on the aqueous solubility ($-\lg S_W$) and n-Octanol/water partition coefficients ($\lg K_{OW}$) of polychlorinated dibenzo-p-dioxins (PCDDs). *QSAR Comb. Sci.* **2007**, *26*, 352–357.

(23) Yang, G.; Zhang, X.; Wang, Z.; Liu, H.; Ju, X. Estimation of aqueous solubility ($-\lg S_W$) of all polychlorinated fibenzo-furans (PCDF) and polychlorinated dibenzo-*p*-dioxins (PCDD) congeners by density functional theory. *J. Mol. Struct. - THEOCHEM* **2006**, *766*, 25–33.

(24) Wei, X.-Y.; Ge, Z.-G.; Wang, Z.-Y.; Xu, J. Estimation of aquaueous solubility (-lgS_W) of all polychlorinated biphenyl (PCB) congeners by density functional theory and position of Cl substitution (N_{PCS}) method. *Chin. J. Struct. Chem.* **2007**, *26*, 519–528.

(25) Staikova, M.; Messih, P.; Lei, Y. D.; Wania, F.; Donaldson, D. J. Prediction of subcooled vapor pressures of nonpolar organic compounds using a one-parameter QSPR. *J. Chem. Eng. Data* **2005**, *50*, 438–443.

(26) Staikova, M.; Wania, F.; Donaldson, D. J. Molecular polarizability as a single-parameter predictor of vapour pressures and octanol-air partitioning coefficients of non-polar compounds: a priori approach and results. *Atmos. Environ.* **2004**, *38*, 213–225.

(27) Puzyn, T.; Falandysz, J. Application and comparison of different chemometric approaches in QSPR modelling of supercooled liquid vapour pressures for chloronaphthalenes. *SAR QSAR Environ. Res.* **2007**, *18*, 299–313.

(28) Zeng, X. L.; Wang, Z. Y.; Ge, Z. G.; Liu, H. X. Quantitative structure-property relationships for predicting subcooled liquid vapor pressure (P_L) of 209 polychlorinated diphenyl ethers (PCDEs) by DFT and the position of Cl substitution (PCS) methods. *Atmos. Environ.* **2007**, *41*, 3590–3603.

(29) Puzyn, T.; Rostkowski, P.; Świeczkowski, A.; Jędrusiak, A.; Falandysz, J. Prediction of environmental partition coefficients and the Henry's law constants for 135 congeners of chlorodibenzotiophene. *Chemosphere* **2006**, *62*, 1817–1828.

(30) Chen, J.; Harner, T.; Ding, G.; Quan, X.; Schramm, K. W.; Kettrup, A. Universal predictive models on octanol-air partition coefficients at different temperatures for persistent organic pollutants. *Environ. Toxicol. Chem.* **2004**, *23*, 2309–2317.

(31) Padmanabhan, J.; Parthasarathi, R.; Subramanian, V.; Chattaraj, P. K. QSPR models for polychlorinated biphenyls: n-Octanol/water partition coefficient. *Bioorg. Med. Chem.* **2006**, *14*, 1021–1028.

(32) Puzyn, T.; Falandysz, J. QSPR modeling of partition coefficients and Henry's Law constants for 75 chloronaphthalene congeners by means of six chemometric approaches - a comparative study. *J. Phys. Chem. Ref. Data* **2007**, *36*, 203–214.

(33) Han, X. Y.; Wang, Z. Y.; Zhai, Z. C.; Wang, L. S. Estimation of n-octanol/water partition coefficients (K_{OW}) of all PCB congeners by ab initio and a Cl substitution position method. *QSAR Comb. Sci.* **2006**, *25*, 333–341.

(34) Zhou, W.; Zhai, Z.; Wang, Z.; Wang, L. Estimation of n-octanol/water partition coefficients (K_{OW}) of all PCB congeners by density functional theory. *J. Mol. Struc. - THEOCHEM* **2005**, *755*, 137–145.

(35) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. RM1: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.* **2006**, *27*, 1101–1111.

(36) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.

(37) Jensen, F. *Introduction to computational chemistry*; John Wiley & Sons: Chichester, United Kingdom, 1999.

(38) Anikin, N. A.; Anisimov, V. M.; Bugaenko, V. L.; Bobrikov, V. V.; Andreyev, A. M. Local SCF method for semiempirical quantum-chemical calculation of ultralarge biomolecules. *J. Chem. Phys.* **2004**, *121*, 1266–1270.

(39) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–670.

(40) Eucken, A.; Hellwege, K. H. *Landolt-Börnstein Zahlenwerte und Funktionen aus Physic, Chemie, Astronomie, Geophysik, Technik*; Springer-Verlag: Berlin, Germany, 1951 (in German).

(41) Isayev, O.; Rasulev, B.; Gorb, L.; Leszczynski, J. Structure-toxicity relationships of nitroaromatic compounds. *Mol. Diversity* **2006**, *10*, 233–245.

CI800021P