General and Class Specific Models for Prediction of Soil Sorption Using Various Physicochemical Descriptors

Patrik L. Andersson,*,† Uko Maran,‡ Dan Fara,‡ Mati Karelson,‡ and Joop L. M. Hermens†

Institute for Risk Assessment Sciences, Utrecht University, P.O. Box 80176, 3508 TD Utrecht,
The Netherlands, and Department of Chemistry, University of Tartu, Jakobi Street 2, Tartu 51014, Estonia

Received May 16, 2002

Diverse chemical descriptors were explored for use in QSAR models aimed to screen the soil sorption potential of organic compounds. The descriptors included logP, HyperChem QSARProperties descriptors, a combination of connectivity indices, geometrical, and quantum chemical measures, and two sets from the DRAGON and CODESSA program packages, respectively. Generally, the univariate logP models were capable of capturing most of the variation and give an indication of the sorption potential. The multivariate models required refined variable selection procedures but were shown to include crucial descriptors for modeling compound classes with specific chemical characteristics.

INTRODUCTION

The soil sorption potential of chemicals is an important parameter in environmental risk assessment procedures to estimate persistence and mobility of chemicals.[1−3] This refers to assessments of the bioavailability of chemicals for both soil- and water-living organisms. Soil sorption is most often expressed as a coefficient defined as the concentration of the chemical in soil divided by the concentration in the aqueous phase. To make comparisons possible of the soil sorption potential of chemicals compiled from various soils and studies, the coefficient is often expressed as K_{oc}, i.e., normalized to organic carbon content. The K_{oc} value will thus be largely independent of the soil, excluding variation from clay content, pH, surface area, and cation exchange capacity of the soil and the nature of the organic matter. Although experimental data are always preferred, soil sorption is difficult and time-consuming to measure. Thus, models of predictive capacity are warranted to assess the persistence and fate of the numerous pollutants and industrial chemicals in use.[1] A massive number of new compounds are registered each year, which means that models should be general and fast to give a first approximate value of their properties.

A common approach to predict soil sorption is to use the octanol−water partition coefficient or the water solubility as physicochemical descriptor.[1,4−7] However, these parameters are highly correlated and for certain chemicals these may not capture the physicochemical properties involved in the sorption process. Recent studies indicate that the K_{ow}-K_{oc} relationship is nonlinear for compounds with logK_{ow} above 6−7.[8,9] Alternative modeling schemes have been published where e.g. connectivity indices and quantum chemical descriptors were applied to describe the structural characteristics of the studied compounds.[4−7,9−12] Another significant concern in soil sorption modeling is the validity of the models. Several published models are trained for a certain class of chemicals, as others are more general. The chemical domain covered in the training of the model determines and limits the area of interpolation versus extrapolation for predicting untested compounds.[13] Thus, validation is a crucial step to reach accurate and reliable models.

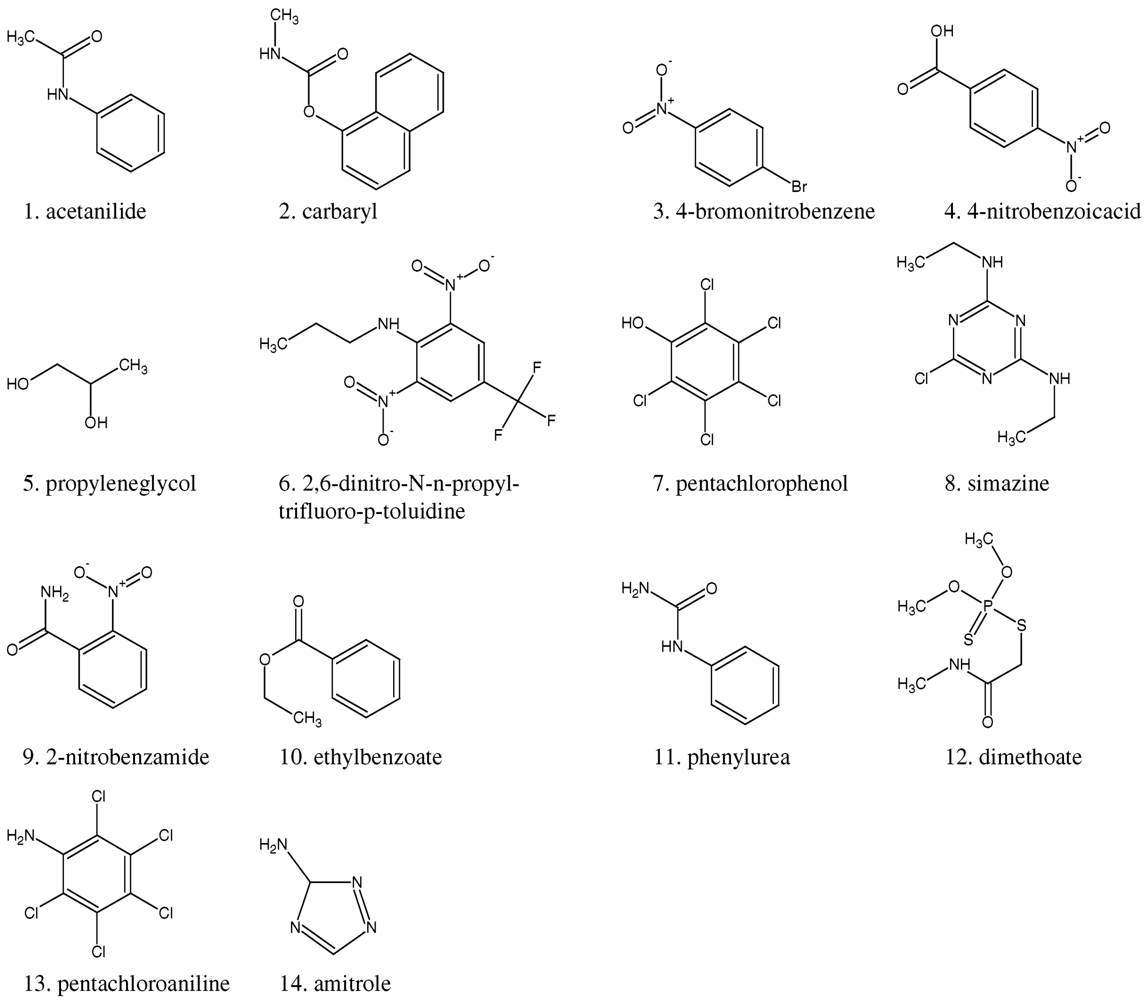
The primary focus of this study was to test different QSAR approaches that are applied in the development of models used for practical application in risk assessment. Various sets of chemical and structural descriptors were applied and assessed to model soil sorption of in total 342 diverse organic compounds. In principle, a method was searched with the capability of serving as an initial screen to identify highly sorptive compounds. Complex models including a multitude of descriptors or descriptors that need comprehensive calculations to be properly derived were compared with more fundamental models including one or a few parameters only. Both general models as well as models for specific chemical classes were considered. In addition, the importance of external validation of QSAR models was stressed in order to determine the reliability and capability of the models. The outcome was also interpreted in mechanistic terms, although we believe that for such type of mechanistic studies, data sets originating from a single study are more suitable because such data are more consistent. Such studies also lend themselves more for a detailed analysis of the physical-chemical interactions between chemicals and soil, see for example ref 14. The basis of this study was that only the structure of the compounds was known and that both the chemical descriptors and the models were to be calculated. Thus, the first phase in the modeling procedure was to encode and express the structures in various measures. In this study, five different sets of chemical descriptors were examined including the logarithm of the octanol−water partition coefficient (logP), descriptors from the QSARProperties module in HyperChem,[15] a diverse set of descriptors with topological indices and quantum chemical descriptors,[16] a large set of descriptors derived from DRAGON,[17] and

* Corresponding author phone: +4690-7865266; fax: +4690-128133; e-mail: patrik.andersson@chem.umu.se. Current address: Environmental Chemistry, Umeå University, SE-901 87 Umeå, Sweden.

† Utrecht University.

‡ University of Tartu.

10.1021/ci025540p CCC: $22.00
Published on Web 09/24/2002



1. acetanilide
2. carbaryl
3. 4-bromonitrobenzene
4. 4-nitrobenzoicacid
5. propyleneglycol
6. 2,6-dinitro-N-n-propyl-trifluoro-p-toluidine
7. pentachlorophenol
8. simazine
9. 2-nitrobenzamide
10. ethylbenzoate
11. phenylurea
12. dimethoate
13. pentachloroaniline
14. amitrole

Figure 1. Molecular structures of representatives for each class of compounds including 1 acetanilides, 2 carbamates, 3 nitrobenzenes, 4 organic acids, 5 alcohols, 6 dinitroanilines, 7 phenols and benzonitriles, 8 triazines, 9 amides, 10 esters, 11 phenylureas, 12 phosphates, 13 anilines, and 14 triazoles.

another complex set of descriptors calculated in CODESSA.[18,19] The soil sorption models were calculated using partial least squares (PLS) regression and analyzed in terms of internal and external validation statistics.[20] The statistical significance of the models was considered in relation to model complexity, ease of use, and physicochemical relevance.

METHODS

Soil Sorption Data. The K_{oc} data set applied in the present study was in part compiled and published by Sabljic.[6] The data set was later extended by Gerstl[4] and then presented in its present form by Sabljic et al.[7] Sabljic and co-workers[7] made the complete data set available in their publication and reported median values in most cases if multiple values were given. In the modeling process logarithmic K_{oc} values were used as soil sorption has been found to be log-normally distributed.[4] The K_{oc} of all chemicals was measured in their nonionic form, and Sabljic et al.[7] grouped them in 14 chemical classes, viz. acetanilides, alcohols, amides, anilines, carbamates, dinitroanilines, esters, nitrobenzenes, organic acids, phenols and benzonitriles, phenylureas, phosphates, triazines, and triazoles. The defined classes include between 10 and 52 compounds and sum up to totally 351 compounds. The chemicals span a relatively large variation in chemical properties and include various functional groups (Figure 1). For example the logK_{oc} value varies between 0 and 4.9, the calculated octanol−water partition coefficient from −1.1 to 8.4, and the molecular weight from 32 to 420. All chemicals are identified with their Chemical Abstract Service (CAS) registry number in parentheses.

The data set was recently used by Eriksson et al.[16] to illustrate the use of statistical multivariate design for selecting representative training sets from a structurally diverse set of chemicals. In this study, a set of 68 compounds was used as selected by the proposed cluster-based factorial design to represent the entire set.[16] In this procedure each class of chemicals was treated separately, and factorial design was applied to span the chemical variation within each class. This means that the training set should be representative of the complete set and thus includes chemicals of all classes and covers the total physicochemical variation. Among the original 351 compounds, 9 compounds were excluded due to errors in the input data, such as SMILES[21] string or trivial name. If the SMILES and the trivial name did not correspond or the CAS number could not be found, we decided to exclude the compound, as the original semiempirical descriptors, which was used in the comparison could be erroneous. The following compounds were excluded: 2-chloroacetanilide (587−65−5), maleic hydrazide (123−33−1), oxamyl

Table 1. Descriptors Applied in Models I to V[a]

model	descriptors
I	logP
II	logP
	solvent accessible surface area
	solvent accessible volume
	refractivity
	polarizability
III	logP
	molecular weight
	polarizability
	geometrical descriptors (6)
	global semiempirical descriptors (7)
	quadrupole moments (6)
	electrostatic potentials (12)
	partial charges (10)
	electrophilic and nucleophilic superdelocalizabilities (12)
	connectivity indices (6)
IV	logP
	constitutional descriptors (56)
	molecular walk counts (20)
	Galvez topological charge indices (21)
	charge descriptors (7)
	Randic molecular profiles (40)
	3D-MoRSE descriptors (160)
	GETAWAY descriptors (196)
	topological descriptors (69)
	BCUT descriptors (64)
	2D-autocorrelations (96)
	aromaticity descriptors (4)
	geometrical descriptors (18)
	WHIM descriptors (99)
	empirical descriptors (3)
V	logP
	constitutional descriptors (38)
	topological descriptors (37)
	geometrical descriptors (12)
	electrostatic descriptors (85)
	quantum-chemical descriptors (554)

[a] The number in parentheses indicates number of descriptors of that category.

(23135−22−0), SD12400 (unknown CAS number), sulfometuron methyl (74222−97−2), carbophenothion (786−19−6), imazalil (35554−44−0), tricyclazol (41814−78−2), and 4-trifluoromethylbenzyltriazole (unknown CAS number). Thus, the training and validation sets included 68 and 274 compounds, respectively.

Physicochemical Parameters. To describe the structural variation and physicochemical properties of the chemicals, various types of descriptors were tested. The initial idea was to put together principally two sets of descriptors for comparison of a complex versus a more simple approach. The complex set of descriptors, here referred to as model III, was used by Eriksson et al.[16] and comprises 62 diverse descriptors, as shown in Table 1. The set includes descriptors, such as quantum chemical, connectivity indices (including valence corrected), molecular volume, molecular surface area, polarizability, and logP. LogP was calculated in ClogP,[22] and the quantum chemical calculations were performed in MOPAC 6.0 using AM1 and PM3 parametrization as described by Müller.[5] Examples of such descriptors are the energies of the highest occupied and lowest unoccupied molecular orbital, heat of formation, dipole moment, electrophilic potentials, partial atomic charges, and electrophilic and nucleophilic superdelocalizabilities. A large number of these parameters are collinear, and the interpretation of parameters of importance is demanding. Due to the number of parameters and the conformational optimization procedure required in the semiempirical calculations, this set was regarded as the complex and sophisticated set. As alternative approach descriptors were generated by the use of the QSARProperties add-on to HyperChem.[15] This set of descriptors includes logP, solvent accessible surface area and solvent accessible volume, refractivity, and polarizability (Table 1).

Two additional sets of descriptors applied in models IV and V were calculated to expand the comparison. Model IV includes descriptors from DRAGON,[17] which accepts HyperChem files and yields automatically a total of 853 descriptors. The descriptors vary largely in origin and include e.g. constitutional, geometrical, connectivity indices, GETAWAY, and WHIM descriptors, see Table 1.[10,23,24] To model IV also logP from ClogP was added. The CODESSA program package was used to calculate descriptors to model V based on the geometric, electronic, and energetic characteristics from MOPAC[25] single point calculations in gas phase with AM1 parametrization.[26] The single point calculations were performed in order to have identical conformation as for the other descriptor sets. As seen in Table 1, the descriptors of model V include logP (ClogP) and various constitutional, geometrical, topological, electrostatic, and quantum chemical descriptors.[18,27,28] Examples of electrostatic descriptors are minimum and maximum atomic partial charges, polarity parameters, topological electronic indices, and charged partial surface area descriptors. Additional quantum chemical descriptors are related to molecular charge distributions, to valency such as bond orders and to bonding contributions, and quantum mechanically calculated energies, e.g. molecular orbital energies and heats of formation. The largest share of descriptors are quantum chemical, and these have proved their applicability in QSARs of various physicochemical and biological properties.[29] In CODESSA, the number of descriptors depends on the atomic constitution of the molecules, and for this set of compounds a total of 726 descriptors was calculated. As seen in Table 1, totally five sets of descriptors were evaluated including the model when logP was used as a single parameter.

Multivariate Tools. In the present study principal component analysis (PCA) was used to study the variation among the chemicals and the descriptors, respectively.[30] In PCA, the original data matrix is decomposed into new latent variables, and the variation among the objects, here compounds, is illustrated in score plots and among the descriptors in corresponding loading plots. To relate the chemical information as described by the physicochemical descriptors to the soil sorption coefficient, partial least squares modeling was used.[20] In PLS, the latent information in the independent parameters is related to the dependent variable, in this case the logK_{oc} values, through a weight vector. The number of significant descriptors is determined for both PCA and PLS by a cross-validation procedure.[31] For comparison of the calculated QSAR models following statistical measures are given: the variation explained in the X-matrix (here the chemical descriptors) (R^2X), the variation explained in Y (K_{oc}) (R^2Y), the cross-validated explained variance (Q^2), and the root-mean-square error of estimation (RMSEE) and prediction (RMSEP).[32] Further, to compare the significance of the descriptors for the model, the variable influence on projection (VIPs) was calculated.[33] The VIP value gives a summary of the influence of the descriptors over many

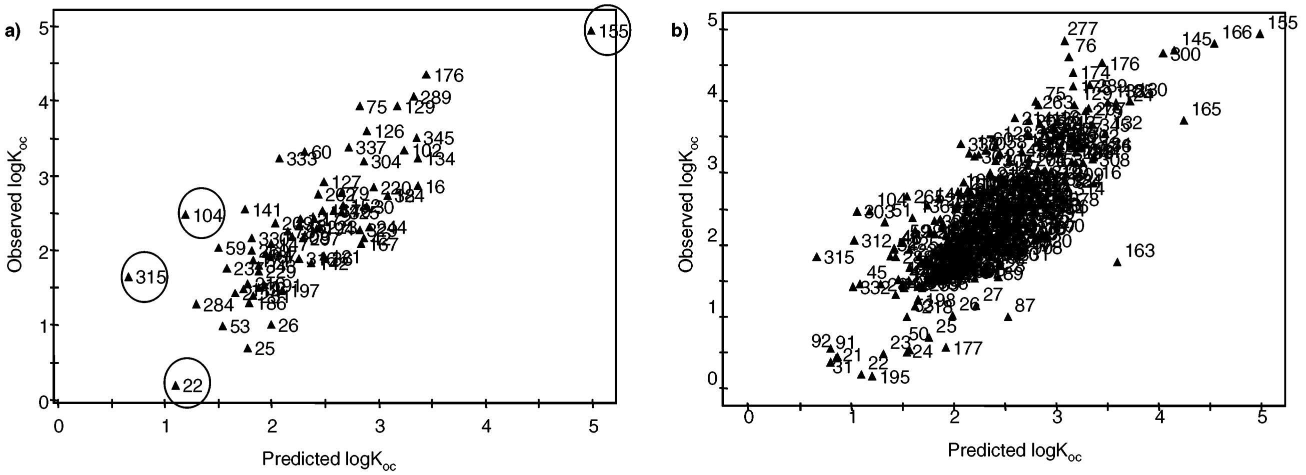


Figure 2. The soil sorption model I showing (a) observed versus by the model predicted logK_{oc} for the 68 compounds of the training set and (b) observed versus predicted logK_{oc} for both training and validation sets.

Table 2. Modeled Variation in the Descriptor Set (R^2X), Explained Variation in K_{oc} (R^2Y), Cross-Validated Explained Variation (Q^2), Root Mean Square Error of Estimate (RMSEE) and for Predictions (RMSEP) from the Soil Sorption Models[a]

model	R^2X	R^2Y	Q^2	RMSEE	RMSEP
I	1	0.62	0.61	0.54	0.51
II	0.99	0.70	0.65	0.49	0.58
III	0.49	0.69	0.49	0.49	0.56
IV	0.48	0.71	0.59	0.48	0.56
V	0.34	0.75	0.53	0.56	0.60

[a] Models II to V include 2 partial least squares components.

components and the higher VIP the more significant is the descriptor. To identify outliers in the models, the distance to the model in both X (DmodX) and Y (DmodY) was inspected. These measures correspond to the residual standard deviation in X and Y, respectively. The PCA and PLS calculations were performed using SIMCA-P 8.0.[32]

RESULTS AND DISCUSSION

In total, five general soil sorption models were developed based on the training set of 68 compounds and including various sets of physicochemical descriptors. The models were validated using the remaining 274 compounds as validation set, and all results are summarized in Table 2. In addition, class-specific models were calculated to study the quality of the descriptors in such restricted modeling applications. Details on both approaches are summarized below.

General Models. Model I. The first model and the most simple included the logarithm of the octanol−water coefficients as calculated in ClogP. This univariate model showed, as seen in Table 2, that a large share of the variation in soil sorption could be captured in such a basic model. The internal validation displayed a cross-validated explained variance of 0.61, and the external validation resulted in a RMSEP of 0.51 in logK_{oc}. Notably the error for the validation set (RMSEP) showed to be lower than the error for the training set (RMSEE). One explanation to this result could be that the training set includes relatively more special compounds than the validation set. These compounds could not be captured by the model or have high respectively low K_{oc} values, see in Figure 2 e.g. #22 ethanol (64−17−5), #104 asulam (3337−71−1), #155 di-2-ethylhexyl phthalate (117−81−7), and #315 dicrotophos (141−66−2). Further, a single parameter model will be less influenced and tailored for specific compounds, which means that these compounds will show large residuals in univariate models. Generally a model with higher degree of complexity should be better to handle compounds with specific and different physicochemical or soil sorption characteristics.

Model II. To achieve complementary descriptors to logP and to increase the complexity in the set, new descriptors were calculated using the QSARProperties module in HyperChem.[15] A method was searched where SMILES notations or CAS registry numbers could be used directly from a database and converted to, e.g. mol files. In this study, two specific approaches were compared to create 3D-structures from which descriptors could be calculated. In the first method the CAS number of each compound was submitted to the ChemIDplus[34] search system where a MDL Mol File of the structure was saved. This file was then imported to HyperChem to calculate descriptors. In the second approach, the structures were drawn manually in HyperChem followed by adjusting bond lengths and bond angles by standard procedures. These structures were then used to calculate the descriptors. As seen in Figure 3, the results of the two approaches match well for molecular surface area and volume (not shown), polarizability, and refractivity. However, the use of mol files in HyperChem may well give flawed log P results for certain type of chemicals (Figure 3). Thus, all structures were drawn by hand before calculating descriptors in QSARProperties.

The PLS model, which included 5 parameters, explained 70% of the variation in K_{oc} and showed a cross-validated explained variance of 0.65, see Table 2. Thus, in terms of Q^2 value this relatively simple model is more accurate than the univariate model, and the new descriptors seem to add vital information to the model. However, the external validation showed that the predictive capacity of model II did not improve as compared to model I. The most important parameter in model II, according to VIP value, was logP, whereas the others were of equal importance. As an attempt to include parameters that account for more hydrophilic characteristics of the compounds, the model was complemented with constitutional descriptors, such as the number of hydrogen bond donors and acceptors, and the number of

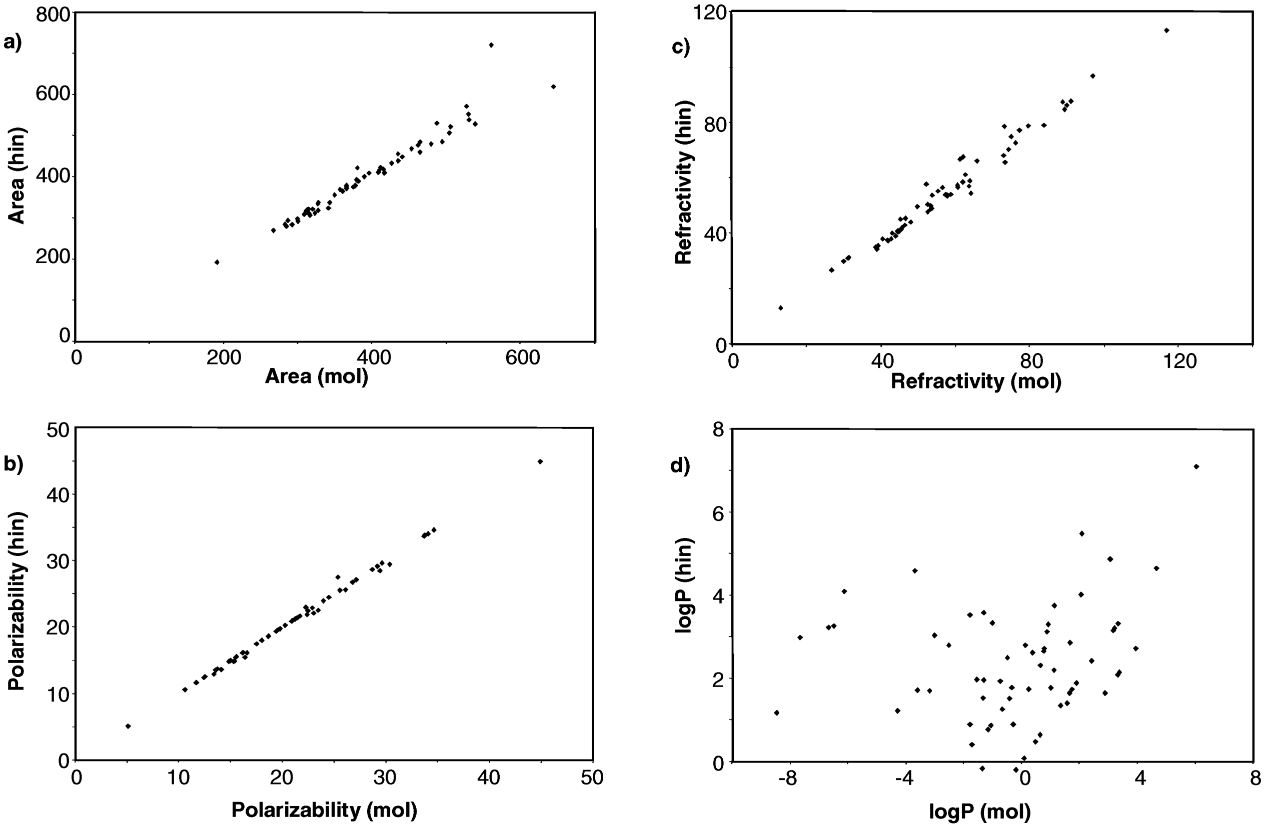


Figure 3. Physicochemical parameters calculated in HyperChem QSARProperties based on manually drawn structures (hin) versus structures taken from the ChemIDplus search system (mol). The plots show (a) solvent accessible surface area, (b) polarizability, (c) refractivity, and (d) logP.

following functional groups: $-OH$, $-NH$, $-NH_2$, $-CO$, and $-NO_2$. The hydrogen bond donor and acceptor counts had the highest influence, but in total these parameters only influenced the model negatively, i.e., Q^2 decreased and RMSEP increased. Therefore these descriptors were considered of low importance for a class independent model. An alternative model II was calculated using the logP from model I. This model showed a similar RMSEP as model I, which indicates that logP from HyperChem is probably less informative than logP from ClogP. Besides, this shows that the four parameters from QSARProperties do not supply significant information to improve the performance of the soil sorption model. The differences in logP may be due to the two different approaches to calculate logP of the applied programs. QSARProperties uses atomic parameters based on a classification scheme,[35] while ClogP applies a molecular fragment method.

Model III. The PLS model including 62 descriptors showed a high internal correlation with a R^2Y of 0.69 but a low cross-validated explained variance (0.49). The external validation of the model scored a RMSEP similar to model II (Table 2). The interpretation of descriptors of importance is relatively complicated and the information principally hidden in the large number of descriptors. As seen in the VIP plot in Figure 4, logP was the absolute most significant parameter followed by molecular weight, polarizability, electrophilic superdelocalizabilicity, connectivity indices and molecular volume and surface measures. A large number of parameters were of minor and equal importance to the model. The local semiempirical parameters, such as partial charges and electrostatic potentials, were shown to be less influencing and by omitting these the statistical significance of the model could be improved.

To improve the performance of model III variables were deleted based on their VIP values. The new models were calculated including descriptors with VIP values above certain levels. As seen in Table 3 the R^2Y remains stable as variables are deleted or even decreases until only variables with VIPs above 1.1 are left in the model. In contrast to the R^2Y, the Q^2 value increases steadily as descriptors are excluded and a total increase in Q^2 of 0.2 was reached from the model with all descriptors to one including only descriptors with VIP values above 1.4. At the VIP value of 1.1, a cutoff seems to have been reached considering R^2Y, RMSEE, and RMSEP. The performance of the model including only variables with VIP values above 1.1 surpass the less refined models in RMSEP with about 0.1 log units. In this model only 16 of the original 62 descriptors were kept. An explanation to the found cutoff value of 1.1 can be seen in Figure 4 as the VIP values of the descriptors increase noticeably at this level.

Model IV. If 3D-structures are available, applying the DRAGON program package can easily extend the number of theoretical molecular descriptors. The program calculates based on input structures in total 853 diverse descriptors, such as constitutional, topological, WHIM, GETAWAY, and Randic molecular profile descriptors (Table 1).[10,23,24] Concerning interpretation of descriptors of importance, the situation is obviously even worse here than for model III. Hence, a variable selection procedure is needed to increase the interpretability of the model but also due to correlation among the descriptors. Nevertheless, a PLS model based on

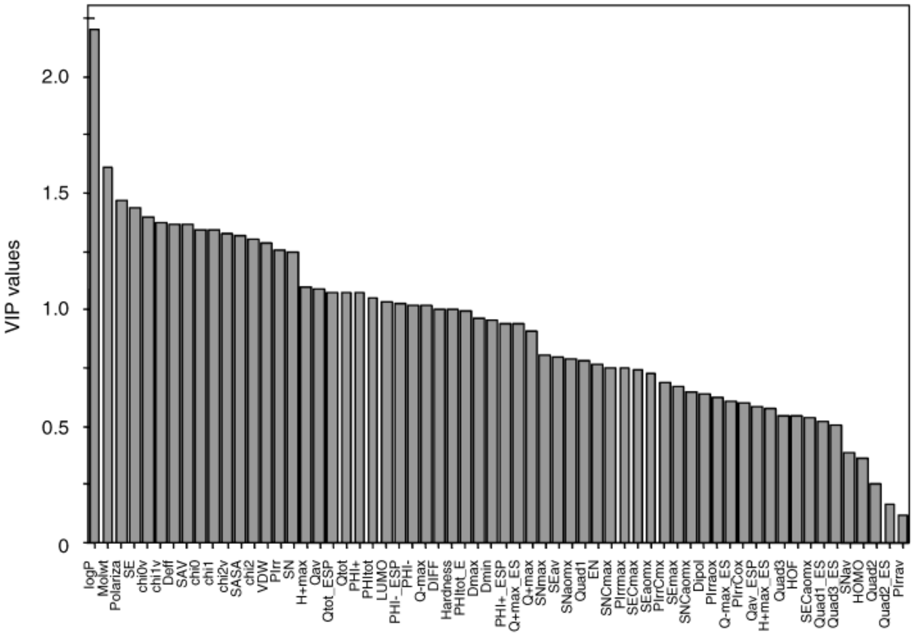


Figure 4. Variable influence on projection (VIP) values from model III showing the significance of the 62 physicochemical descriptors as summarized over 2 components.

Table 3. Statistical Data from the Variable Selection Procedure of Model III[a]

model	R^2X	R^2Y	Q^2	RMSEE	RMSEP
all	0.49	0.69	0.49	0.49	0.56
>0.6	0.59	0.68	0.54	0.50	0.56
>0.7	0.64	0.67	0.55	0.50	0.56
>0.8	0.74	0.65	0.56	0.52	0.58
>1	0.77	0.66	0.58	0.51	0.57
>1.1	0.90	0.73	0.62	0.46	0.48
>1.3	0.91	0.72	0.63	0.46	0.48
>1.4	0.96	0.72	0.68	0.47	0.50

[a] The models include two significant components and descriptors with VIP values above certain cutoffs. Following measures are reported; R^2X the explained variation in the descriptor set, R^2Y the explained variation of K_{oc}, Q^2 the cross-validated explained variation, RMSEE the root-mean-square error of estimate, and RMSEP the root-mean-square error for the predictions.

all variables was calculated, and it performs similar to model III in regards to most statistics (Table 2). The most significant parameters in this model were logP, molecular weight, and certain descriptors from following classes, GETAWAY, 3D-Morse, 2D-Autocorrelation, Geometrical, and BCUT. In total 311 parameters showed a VIP value between 1.7 and 1.1, which indicates that more refined variable selection criteria as well as methods are necessary but were not explored in this study.

Model V. In total 726 descriptors were generated in CODESSA of which Simca only kept 390 after an automatic variable removal. Descriptors were excluded due to more than 50% missing values or if they showed a variance close to zero. As seen in Table 2, the model displayed the highest R^2Y but a relatively low Q^2 and high RMSEP. The 10 descriptors of highest significance according to their VIP values were logP, molecular weight, WNSA-1 Weighted PNSA (PNSA1*TMSA/1000) (Zefirov's PC), gravitation index (all pairs), Kier&Hall index (order 0), ALFA polarizability (DIP), Randic index (order 3), molecular surface area, WNSA-2 Weighted PNSA (PNSA2*TMSA/1000) (Zefirov's PC), and Randic index (order 1).[18] Not surprisingly these are all closely related to the size of the molecules. More interesting is perhaps descriptors of significance in the second component, i.e., those that presumably contributed to explain polar characteristics of the compounds related to soil sorption. To acquire a general score of their importance, the VIP values of the second component were calculated as the difference between the cumulative VIP values and those of the first component. A new sorption model was then calculated by using the 10 most significant descriptors in the first component complemented by the 10 with highest VIP values in the second component. This model with 2 significant components showed improvements in all statistical values except R^2Y (R^2X 0.73, R^2Y 0.73, Q^2 0.68, RMSEE 0.47, RMSEP 0.55). The most significant descriptors in the second component reflect specific characteristics of the compounds, e.g. max e−e repulsion for a C−H bond, max e−e repulsion for an O atom, min atomic orbital electronic population, and min net atomic charge for an O atom.[18] In summary, CODESSA clearly provides descriptors of significance for soil sorption models and might be even more useful for class specific models but need a careful variable selection procedure and interpretation. This issue will, however, be dealt with in a separate publication.

Class-Specific Models. Separate models of the 14 chemical classes predefined by Sabljic et al.[7] were calculated using descriptors from models I to III. The derived models were validated by their R^2Y and Q^2 values, as shown in Table 4. Models II and III, i.e. those based on the descriptor sets of the general models II and III, were compared independent of the number of significant components. The average R^2Y values indicate that model III showed highest internal correlation followed by models I and II. This also applies to number of classes where model III resulted in highest R^2Y. However, concerning predictive capacity the logP models scored the highest average Q^2 (0.61) and were most significant for 9 of the 14 classes of chemicals. An alternative model II was calculated, which included logP from ClogP instead of logP from HyperChem. These models with average

Table 4. Statistical Data and Number of Significant Components from the Class Specific Models Based on Descriptors from Models I, II, and III

class	I	R2Y	Q2	II	R2Y	Q2	III	R2Y	Q2	III*[a]	R2Y	Q2
acetanilides	1	0.53	0.51	1	0.43	0.36	2	0.89	0.67	2	0.89	0.72
alcohols	1	0.82	0.78	1	0.90	0.87	4	0.97	0.89	1	0.89	0.84
amides	1	0.27	0.26	1	0.43	0.40	1	0.54	0.42	1	0.51	0.47
anilines	1	0.78	0.72	2	0.75	0.58	2	0.90	0.77	2	0.91	0.85
carbamates	1	0.56	0.54	1	0.38	0.26	1	0.54	0.33	2	0.64	0.43
dinitroanilines	1	0.81	0.79	2	0.80	0.75	1	0.67	0.42	2	0.80	0.70
esters	1	0.70	0.66	2	0.69	0.58	1	0.52	0.39	3	0.93	0.63
nitrobenzenes	1	0.78	0.77	2	0.84	0.63	1	0.91	0.77	1	0.93	0.90
organic acids	1	0.72	0.69	2	0.73	0.66	2	0.80	0.61	2	0.79	0.70
phenols + benzonitriles	1	0.75	0.73	2	0.74	0.68	1	0.52	0.41	1	0.51	0.43
phenylureas	1	0.50	0.46	1	0.47	0.43	1	0.57	0.44	1	0.55	0.46
phosphates	1	0.61	0.59	2	0.45	0.41	1	0.43	0.33	2	0.67	0.50
triazines	1	0.38	0.32	1	0.39	0.29	1	0.38	0.18	1	0.38	0.31
triazoles	1	0.74	0.70	1	0.60	0.57	2	0.92	0.67	1	0.84	0.79
average		0.64	0.61		0.61	0.53		0.68	0.52		0.73	0.62

[a] III* indicates models using descriptors with VIP values above 1.1 from model III.

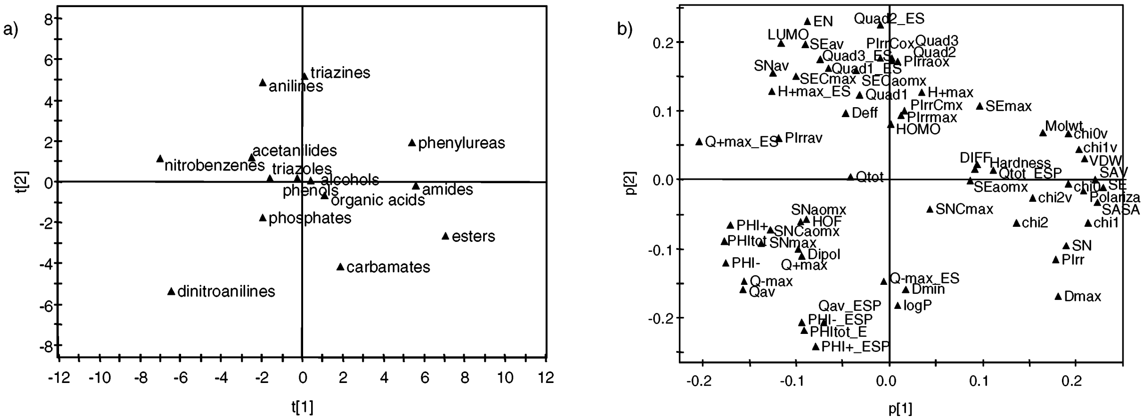


Figure 5. PCA of the 14 chemical classes described by normalized VIP values for the 62 physicochemical descriptors. (a) Score plot of principal component 1 versus 2 and (b) corresponding loading plot.

R^2Y and Q^2 of 0.67 and 0.61, respectively, perform similar to the logP models. This indicates that the logP values, calculated using ClogP are more accurate and useful than those from HyperChem. Addition of the constitutional descriptors to model II, as tested for the general counterpart, did improve neither the internal correlation nor the predictive capacity.

The finding that the logP models showed higher Q^2 values than the more sophisticated models was unexpected. It was hypothesized that class-specific models need descriptors reflecting specific polar characteristics of the compounds. To refine and study influencing descriptors from model III, new models were calculated including only descriptors with VIP values above 1.1. As seen in Table 4, these showed improvements concerning Q^2 for 13 of the 14 studied chemical classes. The Q^2 value of the models varied between 0.31 for triazines to 0.90 for nitrobenzenes with an average value of 0.62. By this variable selection procedure, models were calculated that go beyond the log P models in terms of internal correlation but still not in terms of Q^2. To advance also this measure of the models, a more demanding variable selection strategy was required. If only the 5 most significant descriptors per compound class, according to their VIP values, were kept in the models the Q^2 values improved. The average Q^2 value of models III increases by this procedure from 0.52 to 0.67 and compared to the logP models only esters, phenols and benzonitriles, and phosphates still show lower Q^2. For these classes of chemicals the difference in Q^2 was less than 0.05, thus in summary the multivariate models surpass the univariate but only after a tough variable selection. This procedure can, however, be risky as the models become less stable and more dependent on the quality of the remaining descriptors. The new model may show a better internal correlation, but external compounds may be less well described by the selected descriptors. Thus, it is recommended to use an external set of compounds in the variable selection procedure.[36]

The most significant parameter in all 14 models, except for the acetanilides, alcohols, amides, nitrobenzenes, phenylureas, and triazoles was logP. The result stresses that logP generally is capable of describing the soil sorption phenomena, but that for certain types of chemicals other descriptors are essential. This was also indicated by a large variation in VIP values between the various chemical classes. To study these variations in detail, a PCA was calculated including the VIP values of the descriptors as variables and the chemical classes as objects. Before the calculations the VIP values were normalized to the sum of 100. The first and second PCA components explained 28% and 14% of the variation in VIP values, respectively. The score plot as shown in Figure 5 indicates that a few groups of compounds show distinctive VIP patterns, such as the amides, anilines, carbamates, dinitroanilines, esters, nitrobenzenes, phenylureas, and triazines. An interpretation of the most dominant pattern in the PCA shows that descriptors of specific significance for dinitroanilines and nitrobenzenes were PHI, Q+, and Q-av, i.e., electrostatic potentials and partial atomic charges. These two classes of compounds have in common the functional group $-NO_2$, which explains the significance of such descriptors. In the models for anilines and triazines, descriptors such as the electronegativity (EN), the energy of the lowest unoccupied molecular orbital (LUMO), the average electrophilic superdelocalizability (SEav), and the quadropole moments (Quad) 2 and 3 show high VIP values. The energy of LUMO and EN, which is formed through the outer orbital energies as well as SEav, are indicators of reactivity. The importance of these descriptors could be due to the amino groups of the anilines and triazines, which may differ in susceptible to donor/acceptor interactions depending on e.g. position and neighboring groups. The amides, esters, and phenylureas display high significance for size related descriptors, such as molecular weight (Mw), van der Waals volume (VDW), and the valence corrected connectivity indices chi0v and chi1v. This indicates that the differences in soil sorption potential within these groups of compounds are predominantly dependent on bulk properties. Models that showed largest improvements in the variable selection were in general those with the most specific VIP pattern. The descriptors related to the specific chemical features of each class will in such models be given a higher influence.

In particular five chemical classes resulted in poor models even after the variable selection procedure, viz. amides, carbamates, phenols and benzonitriles, phenylureas, and triazines. Outlying behavior both in K_{oc} and chemical characteristics were considered in detail in these models using descriptors from model III. The normalized distance to the model in both physicochemical properties (DModX) and soil sorption (DModY) were used to identify extreme compounds. After identification of special compounds, these were deleted and models recalculated. The new models were calculated either without compounds of high DmodX or high DmodY. In general, the new models improved most if compounds with high DModY were excluded. This indicates inaccuracy in the K_{oc} values and that differences in chemical properties were less influencing. In total, the new models showed after deleting the extreme compounds an increase in Q^2 from 0.42 to 0.64, 0.33 to 0.66, 0.41 to 0.57, 0.44 to 0.65, and 0.18 to 0.41 for amides, carbamates, phenols and benzonitriles, phenylureas, and triazines, respectively. The molecular structures of the extreme compounds were examined, but no special functional groups or atoms could be isolated as the cause of the high DmodX or DmodY. From the considered chemical classes following seven compounds with high DmodY were deleted, viz. 2,6-dichlorobenzamide (2008−58−4), EPTC (759−94−4), asulam (3337−71−1), 2,3,5-trimethylphenol (697−82−5), 3-hydroxyphenol (108−46−3), diflubenzuron (35367−38−5), and chlorsulfuron (64902−72−3).

CONCLUDING REMARKS

QSAR models to predict the potential of chemicals for sorption to soil are important for assessing their fate and risk in the environment. Models provide a means to fill the huge gaps of data in the flood of new chemicals entering the market and possibly the environment. Besides, models are capable to generate comparable and continuous scales of the soil sorption coefficients. In this study we have explored five different sets of chemical descriptors for use in QSAR models aimed to predict soil sorption of a range of organic compounds. These include various types of descriptors, such as constitutional, geometrical, connectivity, and quantum chemical indices. The multivariate sets of descriptors were in general equally useful for predicting K_{oc}, which makes sense since they all include measures reflecting hydrophobicity, bulk, and polar characteristics. To identify differences between the descriptor sets applied in the present study, more specific endpoints than partition coefficients should be studied. In fact, diverse chemical descriptors have previously been shown to contain similar information,[37] although they may differ on a local level.[38] However, in a recent comparison of descriptors for use in lead search in drug discovery, both logP and connectivity indices were ranked low, and much more specific parameters, such as comparative molecular field analysis (CoMFA) parameters, were considered as the most useful.[39] There are clear advantages of using calculated molecular descriptors as compared with experimental properties given that they are relatively simple and rapid to achieve. They also provide information before the actual compound is synthesized or available. Some of the calculated descriptors are though not easy to interpret and link to an actual physicochemical characteristics or functional group of the compound. In addition, the amount of calculated molecular descriptors, possibly and easily generated may hide the significant information in the model and thus decrease its interpretability. This stresses the importance of multivariate approaches and sensitive variable selection procedures in the development of QSAR models.

Since organic carbon (OC) normalized sorption coefficients were modeled in this study, logP was as expected the most dominant chemical descriptor. This normalization procedure reduces largely effects from other constituents in the soil, such as clay content and metal oxides. Although it is accepted that partition in soils with an OC level above 0.1% is dominated by the content of soil organic matter,[2,40,41] other factors may be of importance in particular for relatively polar compounds.[42,43] Molecular interactions between soil and solute are thus still a factor of concern including π bonding, hydrogen bonding, ligand exchange, and covalent bond formation through reaction with the functional groups of the compounds.[2,40,41] In a study with atrazine it was reported that the soil sorption potential of the compound in various soils was strongly correlated with the OC content; however, the study also indicated the significance of the composition of the organic phase.[42] Most of the variation in the soil sorption of atrazine was explained by the fraction of humic acids. This study also showed that oxides such as aluminum and manganese influence only to a minor extent the partition coefficients. Recently interactions to clay and organic matter were studied for seven pesticides with varying physicochemical properties.[43] Three of the compounds, viz. 4,6-dinitro-o-creosol, dichlobenil, and carbaryl showed stronger sorption to clay than to organic matter. Examples of chemical features of the compounds, thought to correlate with interactions to the clay, were planar aromatic structures and electron withdrawing substituents, such as $-CN$ or $-NO_2$.[43] Another factor that may contribute to variations in K_{oc} is the humidity in the soil.[44] A strategy to avoid the discussed variations in K_{oc} is to use common testing protocols and reference soils, such as the OECD soil[45] and EUROSOILS.[46] A HPLC screening method to measure soil sorption of organic chemicals was recently tested using the second generation of the EUROSOILS.[47]

The performance of QSAR models has to be evaluated in relation to the expected experimental error and as discussed above the various factors of uncertainty. An indication of the size of this error has been given for atrazine, which was tested in 109 soils resulting in a K_{oc} that varied with a factor of 4.[42] This uncertainty agrees well with the RMSEP values reported in the present study for the general models, viz. 0.5 to 0.6 log units. Furthermore, a critical step in constructing the model and to recognize before applying the model in risk assessment procedures is the chemical domain covered. This is basically determined in the physicochemical variation and representativity of the compounds in the training set. In the present study, the applied training and validation sets were selected in a thorough selection procedure to achieve representative compounds from a structurally diverse group of compounds.[16] A certain number of compounds, depending on the complexity of its chemical domain, were selected from each of the 14 chemical classes. By this procedure the chemical domain of the 342 diverse chemicals was covered. Untested compounds to be predicted should possess similar chemical properties as these, which can be examined in score plots or DmodX values. The applied practice yields information on the calculated model to set its limits for use as predictive tool in risk assessment procedures. Further, the use of an external validation set is crucial to assess the predictive capacity of the model. This set of compounds should represent the chemical domain equally well as the training set. In the present study, 274 compounds were used to validate the models established on only 68 compounds. The presented RMSEP values are thus based on a robust and conservative validation, and these could certainly be improved if using a larger set of compounds to train the model. The relatively small training set could be argued not to capture the chemical variation, thus models were also calculated where the training and validation sets were altered. In summary, these models showed improved external predictivity and decreased internal correlation performance. To note was also that the statistics of the univariate model was no longer better but similar to the multivariate models. In general, however, this alternative analysis of the data did not affect our main conclusions.

The five sets of descriptors explored in the present study included various categories and numbers of structural and physicochemical descriptors. Uni- and multivariate models were developed to validate the performance of the descriptors and in order to reach a QSAR model for screening of the compounds soil sorption potential. Generally the more sophisticated and complex models, including a multitude of descriptors, did not show superior performance compared to the simple logP model. This indicates that time-consuming modeling, such as geometrical optimization procedures to achieve quantum chemical measures, can be avoided. On the other hand, the transparent and univariate model heavily depends on the quality of the logP estimate, which directly determine the outcome. In addition, logP has shown in recent studies not to be enough to explain the chemical features related to soil sorption of the compounds.[9] It was reported that connectivity indices quantifying size and shape have dominant effect in models of soil sorption for various persistent organic pollutants. Furthermore, certain chemical properties of the compounds have been shown to affect the sorption,[43] which may not be well described by logP. Examples of chemical classes studied here with functional groups, which need additional descriptors for characterization are e.g. acetanilides, amides, carbamates, phenylureas, and triazines. It should be emphasized that in this study no particular descriptor was found more useful to describe polar characteristics. It is more likely that a selection of descriptors compose the inherent capacity to reflect the crucial properties. The descriptors tested in the present study have various advantages and provide crucial information in constructing QSAR models to predict the K_{oc} coefficients of specific classes of organic substances. However, the results of the present study indicate that a basic univariate logP model is sufficient as a first screen of the compounds soil sorption potential.

ACKNOWLEDGMENT

Financial support is greatly acknowledged from the research training network entitled “Intelligent Modeling Algorithms for the General Evaluations of TOXicities” (IMAGETOX) (HPRN-CT-1999-00015). We also thank Aynur Aptula and Gerrit Schüürmann, UFZ Centre for Environmental Research, Leipzig, Germany for assistance in checking CAS numbers and SMILES of the compounds and advice and Martin Müller, Fraunhofer Institute for Environmental Chemistry and Ecotoxicology, Schmallenberg, Germany and Aleksandar Sabljic, Institute Rudjer Boskovic, Zagreb, Croatia for providing the K_{oc} database and physicochemical descriptors and comments on the manuscript.

REFERENCES AND NOTES

(1) Gawlik, B. M.; Sotiriou, N.; Feicht, E. A.; Schulte-Hostede, S.; Kettrup, A. Alternatives for the determination of the soil adsorption coefficient, K_{oc}, of nonionic organic compounds − A review. *Chemosphere* **1997**, *34*, 2525−2551.

(2) Northcott, G. L.; Jones, K. C. Experimental approaches and analytical techniques for determining organic compound bound residues in soil and sediment. *Environ. Pollut.* **2000**, *108*, 19−43.

(3) Van den Berg, M.; Van de Meent, D.; Peijnenburg, W. J. G. M.; Sijm, D. T. H. M.; Struijs, J.; Tas, J. W. Transport, accumulation and transformation processes. In *Risk Assessment of Chemicals: An Introduction*; van Leeuwen, C. J., Hermens, J. L. M., Eds.; Kluwer Academic Press: Dordrecht, The Netherlands, 1995; Chapter 3, pp 37−102.

(4) Gerstl, Z. Estimation of organic chemical sorption by soils. *J. Contam. Hydrology.* **1990**, *6*, 357−375.

(5) Müller, M. Quantum chemical modelling of soil sorption coefficients: Multiple linear regression models. *Chemosphere* **1997**, *35*, 365−377.

(6) Sabljic, A. On the prediction of soil sorption coefficients of organic pollutants from molecular structure: Application of molecular topology model. *Environ. Sci. Technol.* **1987**, *21*, 358−366.

(7) Sabljic, A.; Gusten, H.; Verhaar, H.; Hermens, J. QSAR modelling of soil sorption. Improvements and systematics of logK_{oc} vs logK_{ow} correlations. *Chemosphere* **1995**, *31*, 4489−4514.

(8) Baker, J. R.; Mihelcic, J. R.; Shea, E. Estimating K_{oc} for persistent organic pollutants: limitations of correlations with K_{ow}. *Chemosphere* **2000**, *41*, 813−817.

(9) Baker, J. R.; Mihelcic, J. R.; Sabljic, A. Reliable QSAR for estimating K_{oc} for persistent organic pollutants: correlation with molecular connectivity indices. *Chemosphere* **2001**, *45*, 213−221.

(10) Gramatica, P.; Corradi, M.; Consonni, V. Modelling and prediction of soil sorption coefficients of nonionic organic pesticides by molecular descriptors. *Chemosphere* **2000**, *41*, 763−777.

(11) Meylan, W.; Howard, P. H.; Boethling, R. S. Molecular topology/fragment contribution method for predicting soil sorption coefficients. *Environ. Sci. Technol.* **1992**, *26*, 1560−1567.

(12) Tao, S.; Lu, X.; Cao, J.; Dawson R. A comparison of the fragment constant and molecular connectivity indices models for normalized sorption coefficient estimation. *Water Environ. Res.* **2001**, *73*, 307−313.

(13) Sjöström, M.; Eriksson, L. Applications of Statistical Experimental Design and PLS Modeling in QSAR. In *Chemometric methods in molecular design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Switzerland, 1995; Vol. 2, Chapter 3.2, pp 63−90.

(14) Goss, K.-U.; Schwarzenbach, R. P. Linear free energy relationships used to evaluate equilibrium partitioning of organic compounds. *Environ. Sci. Technol.* **2001**, *35*, 1−9.

(15) *HyperChem Pro 6.02;* Hybercube, Inc.: Gainesville, FL, U.S.A., 2000.

(16) Eriksson, L.; Johansson, E.; Müller, M.; Wold, S. On the selection of the training set in environmental QSAR analysis when compounds are clustered. *J. Chemometrics* **2000**, *14*, 599−616.

(17) *DRAGON 1.11.* http://www.disat.unimib.it/chm/Dragon.htm, Todeschini, R.; Consonni, V. Milano Chemometrics and QSAR Research Group: Milan, Italy, 2001.

(18) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Reference manual version 2.0*; Gainesville, FL, U.S.A., 1994.

(19) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24*, 279−287.

(20) Dunn, W. J., III.; Wold, S.; Edlund, U.; Hellberg, S.; Gasteiger, J. Multivariate structure−activity relationships between data from a battery of biological tests and an ensamble of structure descriptors: The PLS method. *Quant. Struct. −Act. Relat.* **1984**, *3*, 131−137.

(21) Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(22) *ClogP for Windows*; Version 1.0, Biobyte Corp, Claremont, U.S.A., 1995.

(23) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids. *J. Comput. Aided Mol. Des*. **1997**, *11*, 79−92.

(24) Di Marzio, W.; Galassi, S.; Todeschini, R.; Consolaro, F. Traditional versus WHIM molecular descriptors in QSAR approaches applied to fish toxicity studies. *Chemosphere* **2001**, *44*, 401−406.

(25) Stewart, J. J. P. *MOPAC 6.0 Program Package*; QCPE, No 455, 1989.

(26) 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**, *107*, 3902−3909.

(27) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027−1043.

(28) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, U.S.A., 2000.

(29) Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. QSPR and QSAR models derived using large molecular descriptor spaces. A review of CODESSA applications. *Collect. Czech. Chem. Commun.* **1999**, *1551*, 1−1571.

(30) Jackson J. E. *A users guide to principal components*; John Wiley & Sons: New York, U.S.A., 1991.

(31) Wold, S. Cross-validatory estimation of the number of components in factor and principal component analysis. *Technometrics* **1978**, *20*, 397−405.

(32) *SIMCA-P 8.0;* Umetrics AB, Umeå, Sweden, 2000.

(33) Wold, S. PLS for Multivariate Linear Modeling. In *Chemometric methods in molecular design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Switzerland, 1995; Vol. 2, Chapter 4.4, pp 195−218.

(34) *ChemIDplus*; http://chem.sis.nlm.nih.gov/chemidplus/, Specialized Information Services, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, U.S.A.

(35) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three-dimensional structure directed quantitative structure−activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(36) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis: Principles and Applications*; Umetrics, Inc: Umeå, Sweden, 1999.

(37) Andersson, P. M.; Sjöström, M.; Wold, S.; Lundstedt, T. Comparison between physicochemical and calculated molecular descriptors. *J. Chemometrics* **2000**, *14*, 629−642.

(38) Benigni, R.; Gallo, G.; Giorgi, F.; Giuliani, A. On the equivalence between different descriptors of molecules: Value for computational approaches. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 575−578.

(39) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behaviour: A useful concept for validation of “molecular diversity” descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(40) Karickhoff, S. W. Semiempirical estimation of sorption of hydrophobic pollutants on natural sediments and soil. *Chemosphere* **1981**, *10*, 833−846.

(41) Means, J. C.; Wood, S. G. Hasset, J. J.; Banwart, W. L. Sorption of polynuclear aromatic hydrocarbons by sediment and soils. *Environ. Sci. Technol.* **1980**, *14*, 1524−1528.

(42) Payá-Pérez A. B.; Cortés A.; Sala M. N.; Larsen B. Organic matter fractions controlling the sorption of atrazine in sandy soils. *Chemosphere* **1992**, *25*, 887−898.

(43) Sheng, G.; Johnston, C. T.; Teppen, B. J.; Boyd, S. A. Potential contributions of smectite clays and organic matter to pesticide retention in soils. *J. Agric. Food Chem.* **2001**, *49*, 2899−2907.

(44) Canan Cabar, H. Effects of humidity and soil organic matter on the sorption of chlorinated methanes in synthetic humic-clay complexes. *J. Hazard. Mater.* **1999**, *68*, 6217−226.

(45) OECD guidelines for testing of chemicals: Earthworm, Acute toxicity test, Organisation for Economic Cooperation and Development, OECD guideline 207, 15, 1984.

(46) Gawlik, B. M.; Bo, F.; Kettrup, A.; Muntau, H. Characterisation of a second generation of European reference soils for sorption studies in the framework of chemical testing − Part I: chemical composition and pedological properties. *Sci. Tot. Environ.* **1999**, *229*, 99−107.

(47) Gawlik, B. M.; Kettrup, A.; Muntau, H. Estimation of soil adsorption coefficients of organic compounds by HPLC screening using the second generation of the European reference soil set. *Chemosphere* **2000**, *41*, 1337−1347.

CI025540P