JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Conformation-Dependent QSPR Models: $logP_{OW}$

Markus Muehlbacher,[†,‡] Ahmed El Kerdawy,[†] Christian Kramer,[§] Brian Hudson,[∥] and Timothy Clark*[,†,∥]

[†]Computer-Chemie-Centrum, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, 91052 Erlangen, Germany

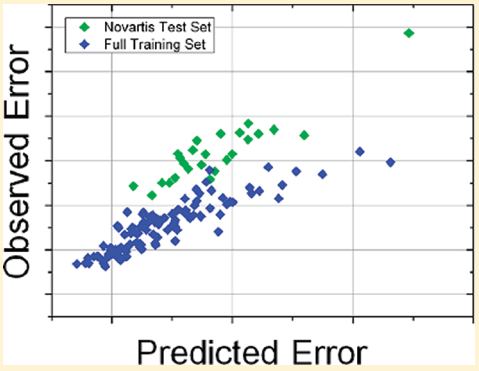
[‡]Psychiatrische und Psychotherapeutische Klinik, Friedrich-Alexander-Universität Erlangen-Nürnberg, Schwabachanlage 6, 91054 Erlangen, Germany

[§]Novartis Pharma AG, Forum 1, Novartis Campus, CH-4056 Basel, Switzerland

[∥]Centre for Molecular Design, University of Portsmouth, Mercantile House, Portsmouth PO1 2EG, United Kingdom

Supporting Information

ABSTRACT: Quantitative structure−property relationships for predicting the water−octanol partition coefficient, $logP_{OW}$, are reported. The models are based on local properties calculated at the standard isodensity surface using semiempirical molecular orbital theory and use descriptors obtained as the areas of the surface found in each bin in a predefined binning scheme. The effect of conformation is taken into account but was found to have little effect on the predictive power of the models. A detailed error analysis suggests that the accuracy of the models is limited by that of the experimental data and that the best possible performance is approximately ±0.5 log units. The models yield a local hydrophobicity function at the surface of the molecules.



INTRODUCTION

As part of our quest to develop more general quantitative structure−property relationships (QSPRs) and to understand the complex relationships between model performance, data, descriptors, and interpolation techniques, we recently described binned surface-integral models for the negative logarithm of the water−*n*-octanol partition coefficient, $logP_{OW}$.[1] These models separate local properties projected onto the molecular surface into bins and use the surface area assigned to the bins as descriptors. The models used isodensity[2] surfaces calculated at the standard ParaSurf[3] contour level of 0.0003 electrons $Å^{-3}$. The local properties include the molecular electrostatic potential (MEP),[4] the local ionization energy,[5] the local electron affinity (EA_L), polarizability (α_L), hardness (η_L), electronegativity (χ_L),[6,7] and the electrostatic field normal to the molecular surface. The models reported used isodensity surfaces and local properties calculated with Austin model 1 (AM1)[8,9] semiempirical molecular orbital (MO) theory.[10] In the following, we have extended the range of models to include the effect of different conformations and the modified neglect of differential overlap (MNDO),[11−13] MNDO/d,[14−18] parameterized model number 3 (PM3),[19−21] AM1*,[22−28] and PM6[29] semiempirical Hamiltonians. Calculations of the local electron affinity using Hamiltonians with polarization functions used the intensity-filtering technique.[30]

Surface-integral models for $logP_{OW}$ are intimately connected with the concept of local hydrophobicity, as discussed for MOLFESD surface-integral models by Brickmann and co-workers.[31,32] However, our original polynomial surface-integral models usually included a large constant term in the regression equation, which prevented their interpretation of the regression formula as a local hydrophobicity.[33] The binned surface-integral models[1] gave smaller constants and therefore became more interpretable in terms of hydrophobicity.

One aspect of the models reported in ref 1 was, however, particularly fascinating. Limiting the training set to rigid compounds (i.e., ones that can only have one possible conformation) removed the observed[1] tendency for the root-mean-square deviation (rmsd) between the model and the experiment to increase essentially linearly with increasing size of the molecules. This led us to conclude that the accuracy of the models is limited by (among other factors) the fact that a single conformation [derived from the CORINA[34] three-dimensional (3D) structure followed by optimization with AM1] was used for each compound. A very limited test in which a conformational search was conducted for three drug molecules and the predicted $logP_{OW}$ calculated by Boltzmann weighting the contributions for each conformation according to its AM1 energy calculated using a self-consistent reaction field (SCRF) correction for aqueous solvation[35] gave improved performance, but this test was by no means statistically significant.

$LogP_{OW}$ is one of the few physical properties for which sufficient data of high enough quality are available to be able to

Received: June 17, 2011
Published: August 15, 2011

Table 1. Definition of the Classification Scheme Used in the Quality Control of the $LogK_{OW}$ Data Set and the Numbers of Compounds Assigned to Each Class[a]

class	definition	compounds	recommended $logP_{OW}$ values[b]
A	compounds that passed all validation tests and for which there was a corresponding match between the CAS registry number and the open Babel[42] unique SMILES in at least one of the public databases	12 567	9043
B	compounds that passed all validation tests but for which there is no match in the public databases	8704	1826
C	compounds for which there is reason to suspect an error in the database SMILES through cross checking	1482	721
D	compounds that were rejected because of 'incorrect' SMILES	635	29
E	compounds that lack structural information (either a CAS registry number or a SMILES) and are thus impossible to cross check	308	4

[a] Of 23 696 compounds, 22 753 could be identified (classes A−C), resulting in 11 590 recommended $logP_{OW}$ values from a total of 11 623. Only the class A−C compounds were used for this work. [b] Recommended $logP_{OW}$ values are those judged to be reliable by the compilers of the database.[37]

Table 2. Numbers of Compounds in the Different Data Sets[a]

data set	single conformation	multi-conformation	conformations
0 rotatable bonds (rigid)	1356	1370	1.03 ± 0.34
0 + 1 rotatable bonds	2869	2888	1.55 ± 1.27
2 rotatable bonds	1407	1419	4.27 ± 4.56
3 rotatable bonds	1206	1199	8.91 ± 12.7
4 rotatable bonds	1018	1023	23.9 ± 35.4
0 + 1 + 2 + 3 + 4 rotatable bonds	6500	6529	6.99 ± 17.1
full = all compounds	11 056		

[a] Numbers of compounds vary slightly between the single- and multiconformation data sets because of a very small number of failures due different handling of the SMILES in the individual workflows (e.g. automatic desalt).

construct QSPR models that depend on the molecular conformation. Thus, most 3D-QSPR models reported in the literature use a single representative conformation, usually derived from a 2D−3D conversion program. However, it seems likely that the idea of deriving a local hydrophobicity function from a surface-integral model fitted by regression depends critically on the correct conformation or on correctly weighted descriptors derived from the conformational ensemble being used, as emphasized in the 4D-QSAR concept.[36] In order to test this assumption, we have taken the effect of the molecular conformation into account specifically when using and training 3D-QSPR models. Our aim is two-fold: to parametrize a function that represents a true local hydrophobicity within the model and to investigate the gain in accuracy likely to be obtained by considering the ensemble of molecular conformations in solution specifically, so that we can judge the possible benefits of compute-intensive workflows that involve conformational searching and weighting.

DATA SET AND DATA PREPARATION

All $logP_{OW}$ data were taken from the $logK_{OW}$ database (2010 version).[37] The identities (i.e., the structures) of the individual compounds were checked according to the following criteria:

- Whether the CAS Registry number is valid using the CAS checksum.
- Invalid SMILES.
- Unfeasibly large or small values of $logP_{OW}$.
- Duplicate CAS RNs with different SMILES.
- That the molecular weight entered in the database matches that from the unique SMILES.
- That the molecular formula in the database matches that from the unique SMILES.
- Whether the compounds could be processed successfully by CORINA,[34] VAMP,[38] and ParaSurf.[3] Many of the errors were simple typographical ones, which were corrected where possible. Where inconsistencies in the data were found, the compounds were assigned an error code.

In a second validation step, the compounds were cross checked against a number of public databases. These were the EPI Physprop,[39] NCI compound,[40] and eMolecules.com[41] databases. The compounds in the database were assigned into five classes, as defined in Table 1.

COMPOUND SELECTION FOR THE TRAINING SET

Permanently charged compounds and zwitterions were first removed from the 11 590 class A−C compounds with recommended $logP_{OW}$ values. Zwitterions were defined as compounds in which two opposite charges are separated by at least three bonds.

Additionally, only compounds that contained elements available for all the semiempirical Hamiltonians used (those containing only H, C, N, O, F, P, S, Cl, Br, or I) were retained in the final data set of 11 208 compounds. Our workflow failed to give 3D coordinates for 42 compounds, resulting in a final total of 11 056 compounds (denoted as "full" set below).

This data set was then split according to the number of rotatable bonds (as determined by MOE)[43] to give subsets of molecules containing 0−4 rotatable bonds, which made up the data sets whose sizes are given in Table 2. The different numbers of compounds for the single and multiconformation cases are again caused by some failures by different components of the workflow.

Table 3. Performance of the Different Models Using Different out-of-Bag Test Sets[a]

								largest errors	
model[b]	Hamiltonian	training set[c]	*b*	*N*	R^2	RMSE	MUE	+ve	−ve
single	AM1	0	1	1508	0.929	0.81	0.56	4.78	−9.87
			2	1405	0.815	0.95	0.69	2.40	−4.12
			3	1197	0.823	0.91	0.66	2.63	−4.29
			4	1012	0.826	0.92	0.66	2.24	−4.86
			1 + 2 + 3 + 4	**5122**	**0.874**	**0.89**	**0.66**	**4.78**	**−9.87**
multi	AM1	0	1	1518	0.882	1.06	0.62	4.74	−19.77
			2	1419	0.774	1.09	0.74	2.23	−5.69
			3	1198	0.790	1.02	0.71	2.21	−4.73
			4	1020	0.802	1.00	0.70	2.95	−5.23
			1 + 2 + 3 + 4	**5155**	**0.834**	**1.05**	**0.69**	**4.74**	**−19.77**
single	AM1	0 + 1	2	1405	0.906	0.62	0.49	2.31	−3.25
			3	1197	0.882	0.68	0.52	2.32	−4.23
			4	1012	0.872	0.73	0.55	2.44	−3.48
			2 + 3 + 4	**3614**	**0.890**	**0.67**	**0.51**	**2.41**	**−4.22**
multi	AM1	0 + 1	2	1419	0.911	0.61	0.48	2.12	−2.70
			3	1198	0.883	0.67	0.52	2.04	−4.11
			4	1021	0.872	0.74	0.55	2.46	−3.18
			2 + 3 + 4	**3638**	**0.892**	**0.67**	**0.51**	**2.46**	**−4.14**
multi	AM1*	0 + 1	2 + 3 + 4	3634	0.789	0.67	0.51	3.28	−3.18
	MNDO			3641	0.790	0.69	0.54	2.51	−3.96
	MNDO/d			3638	0.792	0.67	0.52	2.51	−4.34
	PM3			3640	0.725	0.76	0.59	4.91	−3.61
	PM6			3639	0.793	0.66	0.50	2.63	−3.04

[a] Full details are given for AM1 and only the results of validation with the entire data set ("all") for the other Hamiltonians. The complete data are given in the Supporting Information. [b] Single indicates models generated and applied using the single-conformation workflow defined above, and multi indicates that the multiconformation workflow was used to generate and apply the models. [c] Numbers indicate the number of rotatable bonds in the molecules of the data set, as defined above.

Two different workflows (defined in the Supporting Information) were then used to model these data sets and to assess the effect of an increasing number of rotatable bonds on the performance of the models:

Single-Conformation Workflow. The SMILES strings for each molecule were converted into a single set of 3D coordinates using CORINA.[34] The remainder of the workflow is that described for constructing binned surface-integral models in ref 1 (i.e., geometry optimization with VAMP[38] followed by surface and descriptor generation with ParaSurf 10[3] and model generation as described below).

Multiconformation Workflow. The second workflow included many conformations for each molecule in a Boltzmann-weighted fashion. The details of the individual calculations in the workflows are given in the Supporting Information, but the steps involved are as follows:

(1) Convert the SMILES string to a 3D structure with MOE.[43] The preprocessing of the molecules included a washing step, where we forced the neutralized protonation state. This was mainly done to obtain uncharged structures for the VAMP optimization in order to avoid complications caused by ionic compounds with different counterions. In addition, a molecular mechanics energy minimization by MOE (default settings) was performed to start the conformational search from a reasonable structure.
(2) Generate candidate conformations using the systematic search facility in MOE.[43]
(3) Optimize the geometries of the different conformations with VAMP.[38]
(4) Eliminate identical conformations using energetic and rmsd criteria.
(5) Calculate single-point energies for the remaining conformations using a polarized continuum model self-consistent reaction field (PCM-SCRF) treatment for aqueous solvent.[35]
(6) Construct Boltzmann-weighted descriptors for each compound using either the gas phase or SCRF calculated energies of the individual conformations.
(7) Generate the models as described below.

MODEL BUILDING

The models were constructed using bagged stepwise multiple linear regression with the descriptor pool size corrected F-value[44] as the stopping criterion. Each model consists of 100 independent multiple linear regression models that were built based on randomly chosen 75% fractions of the overall data set. The remaining 25% of the data set are used as a test set. On average every compound therefore occurs 25 times in the test set. The test set prediction is the average prediction for all cases in which the compound in question was part of the test set. Overfitting was avoided using two strategies: Descriptors are not included if they

Table 4. Model Similarities Defined As the Dot Products of the Vectors of Regression Coefficients for the Different Models[a]

single → / multi ↓	0	0 + 1	0 + 1 + 2	full
0	**1.0**	0.031	0.027	0.163
0 + 1	0.148	**0.590**	0.548	0.064
0 + 1 + 2	0.149	0.712	**0.664**	0.060

[a] Similarities in the lower triangle are those between multiconformation models and those in the upper triangle those between single-conformation models. The diagonal values represent similarities between corresponding single- and multiconformation models.

do not improve the predictions more than 95% of random descriptors would, and the 25% test set samples constitute an independent test set that has never been used to select or weight descriptors and can thus be used to assess the quality of the model independently.

This algorithm is independent of any additional parameter tuning and has been proven to deliver very robust models and avoid overfitting. Since it yields a linearly additive model, the final predictions can be mapped back to their origin. For surface-integral models, the contribution of each surface triangle to the final prediction can thus be identified exactly. This allows us to assess contributions of individual groups or moieties to $\log P_{OW}$ (within the context of a strictly local hydrophobicity model, see below). Because of the complex nature of the regression models, this is the most direct way of interpreting the physical meaning of the models.

MEASURES OF QUALITY

The root-mean-squared error (RMSE), the mean unsigned error (MUE), and the predictive R^2 were used to assess the quality of the relative predictions. RMSE and MUE measure the absolute accuracy of the prediction, i.e., how well the experimental value is reproduced by the model. The predictive R^2 measures the correlation of the predicted values with the experimental ones. In contrast to Pearson's R^2, it does not tolerate parallel shifts of the predictions (i.e., predictions that all have the same constant value added).

$$\mathrm{MUE} = \frac{1}{N}\sum_{i=1}^{N} |y_{i,\mathrm{predicted}} - y_{i,\mathrm{measured}}| \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_{i,\mathrm{predicted}} - y_{i,\mathrm{measured}})^2} \quad (2)$$

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{N} (y_{i,\mathrm{predicted}} - y_{i,\mathrm{measured}})^2}{\sum_{i=1}^{N} (y_{i,\mathrm{measured}} - \bar{y})} \right); \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_{i,\mathrm{measured}} \quad (3)$$

Here, N is the number of instances, $y_{i,\mathrm{predicted}}$ the predicted $\log P_{ow}$, and $y_{i,\mathrm{measured}}$ the measured $\log P_{ow}$. The average values are calculated from the complete subset, i.e., either training or validation set.

RESULTS

Model Performance. Models were constructed and applied using the single ("single") and multiconformation ("multi") workflows defined above and using the six semiempirical Hamiltonians (AM1, AM1*, MNDO, MNDO/d, PM3, and PM6). Table 3 summarizes the results obtained for the different types of model constructed for AM1. Results for all other Hamiltonians, performance data for the internal test sets, and coefficients of the averaged models are given in the Supporting Information. All results reported in Table 3 are for completely unseen ("out of bag") validation data sets.

Figure S1, Supporting Information, shows the performance of the models plotted against the number of rotatable bonds. The models trained with a combination of the compounds with zero and one rotatable bonds show state-of-the-art performance in predicting $\log P_{OW}$, especially considering that the most difficult compounds have been used for the validation data sets and that the ratio between the number of compounds in the validation and training/test data sets is unusually large (approximately 5:4). The models trained using the 0 data set perform significantly worse for the remainder of the data than those that used the 0 + 1 data set. This is most likely the result of better coverage of the chemical space of the validation set in the larger 0 + 1 training set, as indicated by the very large deviations for individual compounds using the 0 training set.

Models based on compounds with 0 + 1 rotatable bonds have two major benefits over those based on rigid compounds only: Even for large external test sets, the models perform well with a MUE of approximately 0.5 log units, and performance (estimation error) depends only weakly on the number of rotatable bonds.

All in all, the data shown in Table 3 suggest that the 0 + 1 models are robust and likely to be widely applicable. However, models trained using the compounds with more rotatable bonds may cover a wider applicability domain (see below). The 0 + 1 models are satisfactory, and the conservative model-building strategy used in their construction promises to give unusually robust models, but there may be some benefit from using a larger training set that covers more chemical space. We therefore also trained a single-conformation model using the full data set, as defined in Table 2. This model presumably has a similar coverage of the chemical space as most commercial models but does not allow comparison of single- and multiconformation models because we are not yet able to conduct conformational searches on the most flexible compounds. We have therefore investigated the similarities of our models.

Model similarities were calculated as the dot products of the vector coefficients. Table 4 shows the similarity of the models calculated as the dot product of the vectors of the regression coefficients (all descriptors were normalized and the constant term in the models was treated as a coefficient). The models (single vs multi) are identical by definition for the rigid compounds (0 rotatable bonds). The models using 0 + 1 and 0 + 1 + 2 rotatable bonds are apparently converging to a consensus model. The inclusion of the compounds with two rotatable bonds adds a further 1000 compounds and no additional chemical functionality. Gratifyingly, the multiconformation models for the 0 + 1 and 0 + 1 + 2 data sets are more similar (similarity = 0.71) than

Table 5. Performance (MUE) of the Binned Surface-Integral Models Compared to MOE's Topological Descriptor, ACD Labs (version 10.0), and CLogP[a]

		test set					
model	Hamiltonian	0 + 1	2	3	4	full	external
single 0 + 1	AM1	0.34[b]	0.49	0.52	0.55	0.71	1.05
	MNDO	0.35[b]	0.59	0.74	0.77	0.70	1.13
	PM3	0.41[b]	0.65	0.74	0.76	0.73	1.04
single full	AM1					**0.52**[b]	**0.82**
	MNDO					**0.54**[b]	**0.87**
	PM3					**0.56**[b]	**0.89**
MOE − SlogP[c]		0.47	0.49	0.55	0.55	0.55	0.90
ACD[d]		0.21	0.23	0.22	0.22	0.26	1.21
CLogP[e]		0.32	0.28	0.29	0.30	0.35	0.88

[a] The behavior observed previously[1] is seen once more. The ACD software performs extremely well for published data, most of which are presumably included in its training set. The external training set is, however, unpublished. In this case, the ACD model proves to be less robust than the others. [b] Training set. [c] MOE topological descriptor model, version 2009.10.[43] [d] ACD Laboratories, version 11.1.[49] [e] ClogP.[50]

the corresponding single-conformation equivalents (similarity = 0.55), suggesting that the conformational weighting is helping to converge to a common model. This is also true at a much lower level between the 0 and 0 + 1 models. The multiconformation models show a similarity of 0.15 compared with essentially 0 for their single-conformation equivalents.

Table 4 exhibits the behavior that we would expect if the multiconformation models were physically better founded than their single-conformation equivalents. The apparent convergence of the multiconformation models on extending the training set indicates a well-behaved model/training system. These suggest that the models should converge for training sets larger than 0. The very low similarities between the 0 and 0 + 1 models indicate inadequate coverage of the chemical space by the 0 training set and are consistent with the large maximum deviations shown for these models in Table 3. However, the low similarities between the single-conformation model for the full data set and those for 0, 0 + 1, and 0 + 1 + 2 rotatable bonds indicate that these models are not converging with increasing size of the training set. The limited evidence available in Table 4 suggests that a multiconformation model trained with the full data set may give evidence of model convergence, but constructing such a model is nor computationally feasible.

The constants in the regression equations (Table S1, Supporting Information) can also be taken as a measure of how well the model corresponds to a local hydrophobicity, for which the constant should be 0. Remarkably, the multiconformation models for all 6 Hamiltonians lie between −0.35 (AM1*) and −0.63 (PM6) with a mean value of −0.46 and a standard deviation for the 6 values of 0.10. However, the constants for the models built using the "full" data set are significantly smaller (between 0.05 for AM1 and −0.19 for PM3) with a mean value of −0.12 and a standard deviation of 0.08. Thus, the effect of increasing the applicability domain by extending the training set is to reduce the regression constant to a value close to 0, whereas the constants in the single- and multiconformation models obtained with the same training set and Hamiltonian are very similar. Thus, the hoped for effect that constructing multiconformation models should bring our models closer to a true local hydrophobicity cannot be observed. On the contrary, the effect of increasing the training set is much larger. Nevertheless, the absolute values of the regression constants are less than 0.5 logP units for all models except the PM6 and MNDO/d multiconformation ones, so that the regression expressions are a moderately good representation of a local hydrophobicity if judged by this criterion.

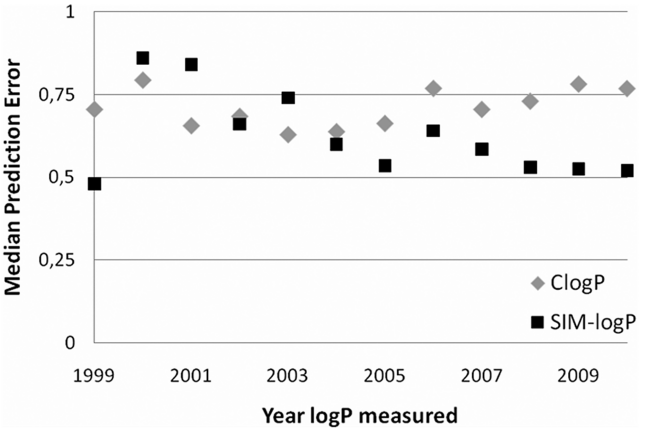


Figure 1. Median error of logP prediction versus year in which logP was measured for the in-house data set for the best commercial model (ClogP, gray) and the SIM logP model based on AM1 and the full training set (black). Figure S2, Supporting Information shows a more complete analysis of these data.

DISCUSSION

Disappointingly, the single models perform almost as well as their multi equivalents, so that we must conclude that the extra work involved in the multiconformation workflow is not justified for most compounds. This may have several causes. The most obvious is that the heats of formation used for Boltzmann weighting are not accurate enough to improve the prediction significantly. Semiempirical MO techniques without an orthogonalization correction generally do not give good results for relative conformational energies,[45] and PCM solvent models are not systematically improvable,[46] so that the Boltzmann weighting factors may be considerably in error.

However, it may also be that even the single-conformation model is performing at the limit of accuracy imposed by the experimental data. We have therefore analyzed the behavior of the observed errors in detail. Because many available methods are likely to have used the data contained in the $\log K_{OW}$ database (which is all available in the open literature) for training, we have included an in-house data set (denoted external in Table 5) from Novartis (N = 2217, measured by dual-phase potentiometric titration)[47,48] as a validation data set that is guaranteed not to be included in the training data. We believe[1] that such in-house data sets provide the most realistic and rigorous test for the real-life predictive performance of the models.

Error Estimation and Analysis. Table 5 shows the binned surface-integral model based on the full training set to be comparable to ClogP and MOE for the external training set and significantly better than ACD Laboratories.

All experimental data sets contain uncertainties due to external (compound purities, buffer purities, temperature, pressure, and

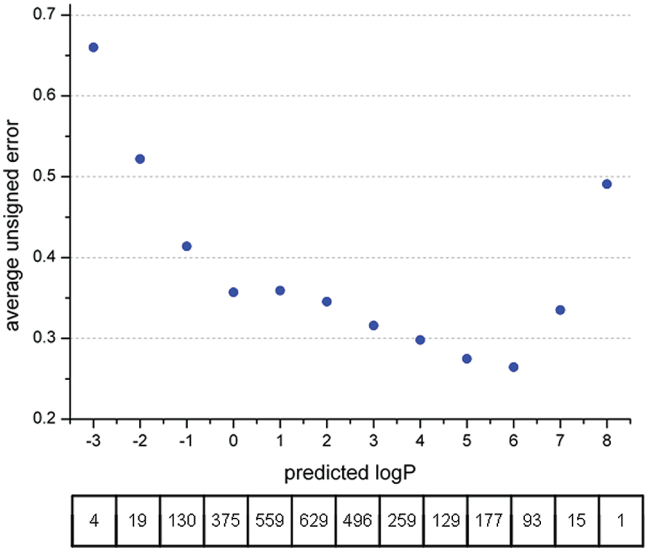


Figure 2. Scatter plot of the MUE of logP prediction with respect to the predicted $logP_{OW}$. The data points are calculated for bins with a range of one log unit in the predicted $logP_{OW}$ centered on integer values. The boxes below the horizontal axis give the number of compounds in each bin.

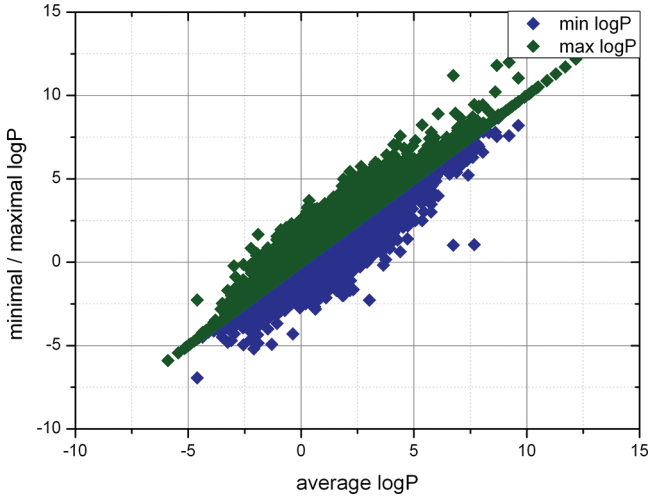


Figure 3. Scatter plot of the lowest and highest $logP_{OW}$ values given in the data set for all compounds with more than one reported value of $logP_{OW}$. The standard deviation of multiple $logP_{OW}$ measurements estimated from all compounds with more than one published value is 0.48.

different measuring methods) and internal factors (changes in experimental setup and differences in interpreting the experimental results). For an experiment run over a long time, the internal sources of error should become lower as specific knowledge of the experimental setup grows. In other words, experimental values from a well-known standardized experiment can often be more reliable than those from a newly introduced experimental setup. We therefore analyzed the predictive performance on the Novartis in-house logP data set of our best model (AM1, single conformation, full data set) and the best commercial model (ClogP) over time. Figure 1 shows the distribution of prediction errors in box-plots for every year (1999−2011) since the dual-phase potentiometric titration experiment[47,48] was implemented at Novartis.

Figure 1 shows that the prediction error of the SIM logP model on the in-house data set decreases over time to a median error of ~0.5 log units. The ClogP prediction error does not improve over the years. Nearly all the strongest outliers of the SIM logP prediction have been identified as measurement errors. Most come from the early years when the logP measurement method was introduced. Since the SIM logP predictive error has become stable in recent years, we anticipate that a median prediction error of 0.5 log units can be anticipated on a mixed drug-like data set.

The initial aim of this work was to investigate the effect of considering possible conformational errors in QSPR models for $logP_{OW}$. This effect turned out to be small, as outlined above. There are, however, other sources of error that contribute to model performance. These include:

Experimental Errors. These are likely to be smaller for $logP_{OW}$ than for other molecular properties because of the large number of data available and the relative reliability of the experimental techniques. The observed MUEs around 0.5 log units support this contention. Large errors for individual compounds can neither be ruled out nor detected easily as the source of disagreement between model and experiment unless more than one measurement exists for the compound in question. Figure 2 shows the dependence of the prediction errors on the absolute value of $logP_{OW}$. The steep increases at the extreme ends of the range indicate increasing uncertainty in the measured values for very hydrophobic or hydrophilic molecules. This behavior is inherent to any quantity that depends on measuring equilibrium concentrations, which become very large or very small and hence are prone to large errors at extreme values of the equilibrium constant. In this specific case, $logP_{OW}$ values higher than 7 or lower than −2 appear to be less reliable than those in the center of the $logP_{OW}$ range (0−6). Difficulties (and hence larger errors) in measuring $logP_{OW}$ values for compounds at the extreme ends of the polarity range are well-known,[51] so that the behavior documented in Figure 2 is to be expected.

Apart from this value-specific error distribution, the experimental measurements themselves have errors or inconsistencies that may affect their accuracy. Some inconsistencies (e.g., different pH values or buffers, different temperatures, etc.) could in principle be detected by examining the experimental protocols in detail. As this has not been done for the training data, we estimated the likely experimental/consistency error limits of the data from all multiple measurements. A plot of the highest and lowest reported $logP_{OW}$ values in the data set against each other is shown in Figure 3.

The scatter is large, and the likely experimental error calculated from these data (rmsd = 0.48 log units for all duplicate measurements, not just the highest and lowest) is close to the performance of most models for their training sets, which suggests that the binned surface-integral descriptors can reproduce the variance found for $logP_{OW}$ for this data set. We note with interest that the RMSEs for the ACD model are lower than our estimate of the likely experimental uncertainty.

Model Errors. The modeling techniques used are adequate for describing the data within the 0.5 log unit error range, so that large outliers can only result for compounds outside the applicability domain or those for which factors not considered in the model affect the measured $logP_{OW}$ strongly. Model errors can be judged from the deviation between the predictions of individual

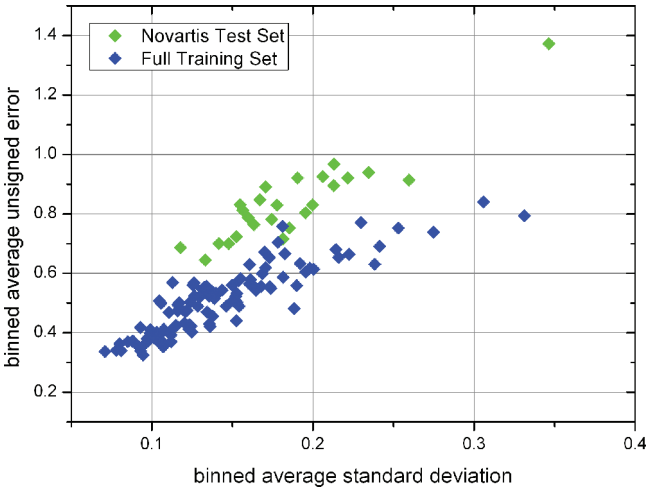


Figure 4. Correlation between average absolute prediction error and average prediction standard deviation for binned lots of 100 compounds for the $\log K_{OW}$ database test set and the Novartis in-house validation set.

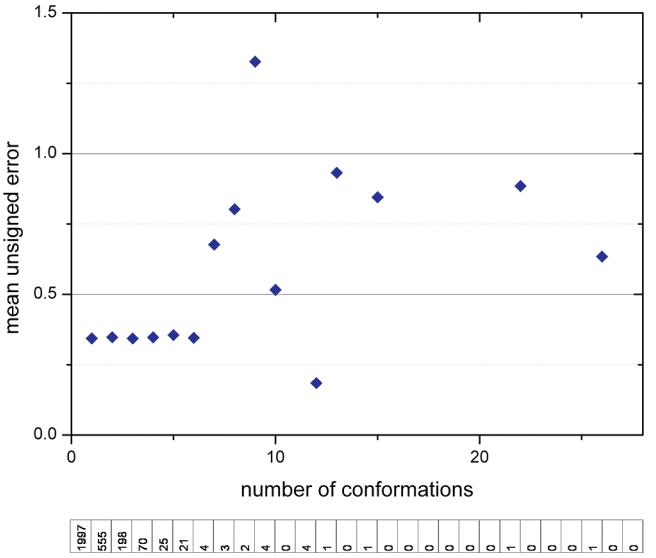


Figure 5. Plot of the average MUE with respect to the number of conformations. The histogram shows the number of compounds in each sample.

models in techniques that use an ensemble of models.[52,53] Because we have used 100 individual models to construct each complete model, we can estimate the overall error from the variance of the predictions from each individual model, as in the technique used earlier for the committee machine neural net models for boiling points.[52] The standard deviation of the predictions indicates the magnitude of the model error. We use a constant factor determined by the performance of the model for the training/validation set to estimate the contributions of the other error factors. This is clearly a significant approximation but does allow some estimate of the likely error of prediction in the absence of unusual error contributions. Figure 4 shows the relationship between the standard deviation of the predictions and the deviation between prediction and experiment for the test data set for the multiconformation model based on the 0 + 1 data set using the AM1 Hamiltonian. Corresponding plots for the other models are given in Figure S3, Supporting Information.

There is a clear and essentially linear relationship between the two quantities ($R^2 = 0.714$ for the $\log K_{OW}$ database, $R^2 = 0.813$ for the Novartis in-house validation set), although the one isolated point for each data set does suggest a better correlation than is actually present. This correlation allows us to estimate the error of prediction from the regression equation.[52] In this case, the estimated RMSE of prediction σ_{exp} is given by

$$\begin{aligned} \sigma_{exp} &= 0.136 + 1.364 \cdot \sigma_{models}(\log K_{OW}) \\ \sigma_{exp} &= 0.310 + 2.820 \cdot \sigma_{models}(\text{Novartis-in-house}) \end{aligned} \quad (4)$$

where σ_{models} is the standard deviation of the predictions of the 100 individual models.

This relationship allows an individual error estimation for each compound, so that the standard deviation of the predictions of the 100 individual models can be treated as a confidence level. It is tempting to interpret eq 4 to mean that the intercept (0.136 $\log P_{OW}$ units) represents the maximum achievable accuracy for the model type described here. The difference in the correlation equations for the training and the unseen data set corresponds to the behavior found previously with a factor of approximately two between training and test set correlations.[52,53]

Conceptual Errors in the Model. All models for distribution coefficients that are based on a single molecular structure must fail if the ensemble of conformations, the protonation state, or any other structural feature of the molecule changes significantly between the two phases. These errors represent an absolute limit of the applicability of such models but are detectable and could conceivably be considered in more complex models. The effect of differently protonated states might be indicated by the dependence of the prediction error on the reported pK_a values of the compounds. An analysis of the dependence of the prediction error on the predicted pK_a is shown in Figure S4, Supporting Information, but does not indicate any significant interdependence.

Generation of Conformational Ensemble. The spread of logP predictions for different conformations per compound is 0.39 log units on average. This explains why the multiconformational model does not generally perform much better than the single-conformation one. However, the spread is much larger for some conformationally flexible molecules, so that the selection of the right conformational ensemble and accurate values for the heat of formation are crucial for the multiconformation model. In this context, the solvent treatment is as important as the underlying Hamiltonian in estimating the conformation equilibrium in solution.

For the multiconformational approach, we were able to analyze $\log P_{OW}$ predictions for different conformations. Figure 5 shows the dependence of the prediction MUE on the number of conformations. The MUE remains approximately constant up to six conformations. Above this value, it rises dramatically, but the bins for more than six rotatable bonds contain very few compounds, as shown by the histogram in Figure 5.

Quantitative Error Estimation. The above analyses suggest that the distribution of errors given by the model depends on a number of factors, so that we can construct a further model to predict the likely prediction error in $\log P_{OW}$. Because the relationship between the descriptors (σ_{models}, predicted $\log P_{OW}$, number of conformations, total surface area, calculated pK_a, number of halogens, oxygens, nitrogens, carbons, see Figure

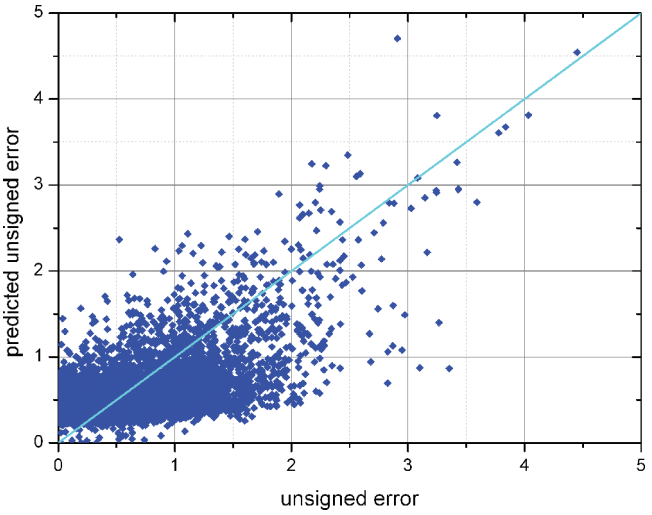


Figure 6. Scatter plot of the error for each individual compound predicted by the ANN against the observed signed error (predicted − experimental value) for the full data set. The line denotes the 1:1 correlation.

S5, Supporting Information) and the likely error is probably nonlinear, we used a simple feed-forward artificial neural net (ANN) trained to predict the model error for each individual compound in the entire validation data set. The results are shown in Figure 6.

Clearly the neural net can only predict systematic errors, so that the scatter in the correlation between predicted and observed $\log P_{OW}$ values is indicative of the random errors. The resolution of the predicted (ANN) error is poor for compounds with observed unsigned errors of two logP units or less, which are usually assigned predicted errors between zero and one. However, the neural net is very good at detecting large outliers that result from systematic weaknesses in the model.

CONCLUSIONS

Binned surface-integral models and the descriptors used to construct them are clearly well able to predict $\log P_{OW}$ as accurately as is allowed by the experimental data. Our estimate of the maximum possible accuracy obtainable is an RMSE of 0.48 $\log P_{OW}$ units. Interestingly, the data are sufficiently consistent that we can reach this accuracy using only 44% of the available data (the 0 + 1 data set) for training and 56% for validation. The resulting models are remarkably robust and generally perform best of the alternatives tested for completely unknown data.

The effect of Boltzmann-weighted conformational averaging is not significant for most compounds, although very flexible compounds may benefit from this effect. This is most likely because even the single conformation models attain the maximum possible accuracy, although errors in the calculated conformational energies in solution may also limit the multiconformation models. However, the available performance data suggest that the multiconformation models are more consistent and probably tend toward a converged consensus model as the number of training data is increased.

A detailed error analysis can relate the performance of the models to known types of error and provides the basis for error estimation for individual compounds.

ASSOCIATED CONTENT

Supporting Information. Two different workflows used to model these datasets, and details of the individual calculations in the workflows. Results for all other Hamiltonians, performance data for the internal test sets, and coefficients of the averaged models. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author
*E-mail: Tim.Clark@chemie.uni-erlangen.de.

ACKNOWLEDGMENT

We thank the Deutsche Akademische Austauschdienst (DAAD) for the award of a Fellowship to A.E.K. and Novartis Pharma AG for the award of an NIBR Presidential Postdoctoral Fellowship to C.K. and for providing unpublished data. We thank Lilya Sviridenko for calculating the ACD Labs logP values for the Novartis internal validation set. We are grateful for financial and software support from Cepos InSilico Ltd.

REFERENCES

(1) Kramer, C.; Beck, B.; Clark, T. A Surface-Integral Model for $\log P_{OW}$. *J. Chem. Inf. Model.* **2010**, *50*, 429–436.
(2) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*. Oxford University Press: Oxford, U.K., 1994.
(3) Clark, T.; Lin, J.; Horn, A. H. C. *ParaSurf 10*; CEPOS Insilico Ltd.: Kempston, Bedford, U.K., 2010.
(4) Politzer, P.; Truhlar, D. G. *Chemical Applications of Atomic and Molecular Electrostatic Potentials: Reactivity, Structure, Scattering, and Energetics of Organic, Inorganic, and Biological Systems*; Plenum Publishing Corporation: New York, 1981.
(5) Sjoberg, P.; Murray, J. S.; Brinck, T.; Politzer, P. A. Average local ionization energies on the molecular surfaces of aromatic systems as guides to chemical reactivity. *Can. J. Chem.* **1990**, *68*, 1440–1443.
(6) Ehresmann, B.; Martin, B.; Horn, A. H. C.; Clark, T. Local molecular properties and their use in predicting reactivity. *J. Mol. Model.* **2003**, *9*, 342–347.
(7) Clark, T.; Byler, K. G.; De Groot, M. J. Biological Communication via Molecular Surfaces. In *Molecular Interactions - Bringing Chemistry to Life*, Proceedings of the International Beilstein Workshop, Bozen, Italy, January 23−25, 2006; Logos Verlag: Berlin, Germany, 2008; pp 129−146.
(8) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
(9) Holder, A. J. AM1 In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons Ltd: Chichester, U.K., 1998, Vol. 1, pp 8−11.
(10) Clark, T.; Stewart, J. J. P. MNDO-like Semiempirical Molecular Orbital Theory and its Application to Large Systems. In *Computational Methods for Large systems: Electronic Structure Approaches for Biotechnology and Nanotechnology*; Reimers, J. R., Ed.; John Wiley & Sons Ltd: Chichester, U.K., 2010, Chapter 8.
(11) Dewar, M. J. S.; Thiel, W. Ground states of molecules. 38. The MNDO method. Approximations and parameters. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.
(12) Dewar, M. J. S.; Thiel, W. Ground states of molecules. 39. MNDO results for molecules containing hydrogen, carbon, nitrogen, and oxygen. *J. Am. Chem. Soc.* **1977**, *99*, 4907–4917.

(13) Thiel, W. MNDO In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons Ltd: Chichester, U.K., 1998, Vol. 3, pp 1599−1604.
(14) Thiel, W.; Voityuk, A. A. Extension of the MNDO formalism to d orbitals: integral approximations and preliminary numerical results. *Theor. Chim. Acta.* **1992**, *81*, 391–404.
(15) Thiel, W.; Voityuk, A. A. Extension of MNDO to d orbitals: parameters and results for the halogens. *Int. J. Quantum Chem.* **1992**, *44*, 807–829.
(16) Thiel, W.; Voityuk, A. A. Extension of MNDO to d orbitals: parameters and results for silicon. *J. Mol. Struct.* **1994**, *313*, 141–154.
(17) Thiel, W.; Voityuk, A. A. Extension of MNDO to d Orbitals: Parameters and Results for the Second-Row Elements and for the Zinc Group. *J. Phys. Chem.* **1996**, *100*, 616–626.
(18) Thiel, W. MNDO/d In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons Ltd: Chichester, U.K., 1998, Vol. 3, pp 1604−1605.
(19) Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.* **1989**, *10*, 209–220.
(20) Stewart, J. J. P. Optimization of parameters for semiempirical methods II. Applications. *J. Comput. Chem.* **1989**, *10*, 221–264.
(21) Stewart, J. J. P. PM3 In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons Ltd: Chichester, U.K., 1998, Vol. 3, pp 2080−2086.
(22) Winget, P.; Horn, A. H. C.; Martin, C. B.; Clark, T. AM1* Parameters for Phosphorous, Sulfur and Chlorine. *J. Mol. Model.* **2003**, *9*, 408–414.
(23) Winget, P.; Clark, T. AM1* Parameters for Aluminum, Silicon, Titanium and Zirconium. *J. Mol. Model.* **2005**, *11*, 439–456.
(24) Kayi, H.; Clark, T. AM1* Parameters for copper and zinc. *J. Mol. Model.* **2007**, *13*, 965–979.
(25) Kayi, H.; Clark, T. AM1* parameters for bromine and iodine. *J. Mol. Model.* **2009**, *15*, 295–308.
(26) Kayi, H.; Clark, T. AM1* parameters for vanadium and chromium. *J. Mol. Model.* **2009**, *15*, 1253–1269.
(27) Kayi, H.; Clark, T. AM1* parameters for cobalt and nickel. *J. Mol. Model.* **2010**, *16*, 29–47.
(28) Kayi, H.; Clark, T. AM1* parameters for manganese and iron. *J. Mol. Model.* **2010**, *16*, 1109–1126.
(29) 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.
(30) Clark, T. The Local Electron Affinity for Non-Minimal Basis Sets. *J. Mol. Model.* **2010**, *16*, 1231–1238.
(31) Jäger, R.; Schmidt, F.; Schilling, B.; Brickmann, J. Localization and quantification of hydrophobicity: the molecular free energy density (MolFESD) concept and its application to sweetness recognition. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 631–646.
(32) Jäger, R.; Kast, S. M.; Brickmann, J. Parameterization Strategy for the MolFESD Concept: Quantitative Surface Representation of Local Hydrophobicity. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 237–247.
(33) Ehresmann, B.; De Groot, M.; Clark, T. Surface-Integral QSPR Models: Local Energy Properties. *J. Chem. Inf. Model.* **2005**, *45*, 1053–1060.
(34) Sadowski, J.; Schwab, C.; Gasteiger, J.; *CORINA 3.4*, Molecular Networks GmbH: Erlangen, Germany, 2006.
(35) Rauhut, G.; Clark, T.; Steinke, T. A Numerical Self-Consistent Reaction Field (SCRF) Model for Ground and Excited States in NDDO-Based Methods. *J. Am. Chem. Soc.* **1993**, *115*, 9174–9181.
(36) Andrade, C. H.; Pasqualoto, K. F. M.; Ferreira, E. I.; Hopfinger, A. J. 4D-QSAR: Perspectives in Drug Design. *Molecules* **2010**, *15*, 3281–3294and references therein.
(37) Sangster, J. $LogK_{OW}$ - A databank of evaluated octanol-water partition coefficients (LogP). Sangster Research Laboratories: Montreal, Quebec, Canada, 2010; purchased April, 2010.
(38) Clark, T.; Alex, A.; Beck, B.; Burckhardt, F.; Chandrasekhar, J.; Gedeck, P.; Horn, A.; Hutter, M.; Martin, B.; Rauhut, G.; Sauer, W.; Schindler, T.; Steinke, T. *VAMP 10.0*; Accelrys Inc.: San Diego, CA, 2007.
(39) *US Environmental Protection Agency*, EPI Suite version 4.00, 2009; http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm (accessed November 23, 2010).
(40) *National Cancer Institute NCI Open Database*, release 3, 2003; http://cactus.nci.nih.gov/download/nci/ (accessed November 23, 2010).
(41) *eMolecules*; eMolecules, Inc.: Solana Beach, CA; http://www.emolecules.com (accessed November 23, 2010).
(42) *The Open Babel Package*, version 2.0.1; http://openbabel.sourceforge.net/ (accessed November 23, 2010); See also Rajarshi, G.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J. K.; Willighagen, E. The Blue Obelisk – Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991−998.
(43) *Molecular Operating Environment (MOE)*, version 2009.10; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2009.
(44) Kramer, C.; Tautermann, C. S.; Livingstone, D. J.; Salt, D. W.; Whitley, D. C.; Beck, B.; Clark, T. Sharpening the toolbox of computational chemistry: A new approximation of critical F-values for multiple linear regression. *J. Chem. Inf. Model.* **2009**, *49*, 28–34.
(45) Clark, T. Quo Vadis, Semiempirical MO-Theory? *J. Mol. Struct. (THEOCHEM)* **2000**, *530*, 1–10.
(46) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3094.
(47) Clarke, F. H.; Cahoon, N. M. Ionization constants by curve fitting: Determination of partition and distribution coefficients of acids and bases and their ions. *J. Pharm. Sci.* **1987**, *76*, 611–620.
(48) Avdeef, A. pH-Metric logP. Part 1. Difference Plots for Determining Ion-Pair Octanol-Water Partition Coefficients of Multiprotic Substances. *Quant. Struct.-Act. Relat.* **1992**, *11*, 510–517.
(49) *ACD/PhysChem Suite*, version 11.1; ACD Labs: Toronto, Canada, 2007.
(50) *ClogP 4.0*; Pomona College and BioByte, Inc.: Claremont, CA, 2008.
(51) Sangster, J. Octanol-water partition coefficients: fundamentals and physical chemistry In Wiley series in solution chemistry; John Wiley & Sons Ltd.: Chichester, England, 1997.
(52) Beck, B.; Breindl, A.; Clark, T. QM/NN QSPR Models with Error Estimation: Vapor Pressure and LogP. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046–1051.
(53) Chalk, A. J.; Beck, B.; Clark, T. A Quantum Mechanical/Neural Net Model for Boiling Points with Error Estimation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 457–462.