A Surface-Integral Model for Log P_{OW}

Christian Kramer,[†,‡] Bernd Beck,*[,‡] and Timothy Clark*[,†,§]

Computer-Chemie-Centrum and Interdisciplinary Center for Molecular Materials, Friedrich-Alexander Universität Erlangen-Nürnberg, Nägelsbachstrasse 52, 91052 Erlangen (Germany); Department of Lead Discovery, Boehringer−Ingelheim Pharma GmbH & Co. KG, 88397 Biberach (Germany); and Centre for Molecular Design, University of Portsmouth, Mercantile House, Hampshire Terrace, Portsmouth, PO1 2EG, United Kingdom

Received November 6, 2009

Log P_{OW}, the negative logarithm of the octanol−water partition coefficient, is omnipresent in computational drug design. Here, we present a surface-integral model for calculating log P_{OW}. The model is based on local properties calculated using AM1 semiempirical molecular orbital theory. These are the molecular electrostatic potential (MEP), local ionization energy (IE_L), local electron affinity (EA_L), local hardness (HARD), local polarizability (POL), and the local field normal to the surface (FN). We have developed a new scheme to calculate a local hydrophobicity based on binning the range of local surface properties instead of using polynomial expansions of the base terms. The model has been trained using ∼9500 compounds available from the literature. It was validated on ∼1350 compounds from the literature and an in-house validation set of 768 compounds from Boehringer−Ingelheim. The model performs similarly to or slightly better than the best commercially available models. We also introduce a model based purely on conformationally rigid compounds that performs well for flexible compounds if the Boltzmann weighted predictions for the different conformers are used. This is the first 3D QSPR model based on such a large databasis that is able to benefit from using conformational ensembles.

INTRODUCTION

Surface-integral models[1] are quantitative structure−property relationships (QSPRs) in which a local property is integrated over the entire surface of a molecule to obtain the target property:

$$P = \int_O f(l_1, l_2, ...l_n)dO \approx \sum_{i=1}^{ntri} f(l_1^i, l_2^i, ...l_n^i)A^i \quad (1)$$

where P is the target property, $f(l_1, l_2,..., ..., l_n)$is a function of the n local properties l_1, l_2, ..., ..., l_n and the integral runs over the molecular surface O. In practice, the integral is replaced by a numerical integration over a triangulated surface consisting of n_{tri} tesserae. The local properties at the center of triangle i of area A^i are given by l_1^i, l_2^i,..., ..., l_n^i. Surface-integral models offer a uniquely detailed view of the exact features that contribute to the physical property being modeled. Simple models based on atom or group additivity allow contributions to be assigned to individual atoms or groups, but Politzer et al.[2] have recently demonstrated that the surface properties of atoms can be very anisotropic, so that a surface-integral approach allows better resolution. Such a detailed view of the lipophilicity contributions of different areas of the surface is necessary to refine our knowledge of drug−receptor interactions, and eventually to design predictive scoring functions. Note that it has been demonstrated that surface integrals of the type used here can be equivalent to volume integrals in determining solvation energies.[3]

The first approach of mapping the lipophilicity to the surface was published by Audry, who introduced the Molecular Lipophilicity Potential (MLP) in 1986.[4] In this approach, the contribution of each Ghose−Crippen[5] type atom to the overall lipophilicity decreased linearly with the distance from the atom to the surface. In 1993 Heiden, Brickmann, and co-workers refined this approach by introducing a different functional form for the distance dependency, which they called MolFESD.[1,6] In the meantime Richards, Williams, and Tute[7] published a systematic foundation for using surface-integral models in modeling solution effects. In 2003, Brickmann and co-workers refined the MolFESD approach by using interaction potentials of an apolar probe on a grid around the molecule instead of Ghose−Crippen atom types as the basis descriptor.[8] In the approach presented in this work, we use local properties derived from semiempirical molecular orbital calculations.

In order to generate a surface-integral model for log P_{OW}, it is necessary to define a local hydrophobicity, whose integral is log P_{OW}. There are other possible definitions of local hydrophobicity, but that based on log P_{OW} has gained wide acceptance. In MLP and MolFESD, this was achieved by defining a surface density that depends on the distance from each atom to the surface point. We have previously used local properties[9] defined using semiempirical molecular

* To whom correspondence should be addressed. Phone: +49(0) 9131 85 22948 (T.C.). Fax: +49(0) 9131 85 26565 (T.C.). E-Mail: Bernd.beck@boehringer-ingelheim.com (B.B.), clark@chemie.uni-erlangen.de (T.C.).

† Computer-Chemie-Centrum and Interdisciplinary Center for Molecular Materials, Friedrich-Alexander Universität Erlangen-Nürnberg.

‡ Department of Lead Discovery, Boehringer−Ingelheim Pharma GmbH & Co. KG.

§ Centre for Molecular Design, University of Portsmouth, Mercantile House, Hampshire Terrace.

10.1021/ci900431f
Published on Web 01/28/2010

orbital (MO) theory and have used them to generate surface-integral models for solvation energies in a variety of solvents,[10] despite the fact that solvation energies cannot be considered local properties. This is made evident by the fact that the models each contained a significant constant, rather than being the simple integral of the local solvation function. This is usually the case for models derived by least-squares fitting as the training data set is usually quite local, so that the model actually fits relatively small deviations from a constant value. We now report a new approach, in which we have defined a true local hydrophobicity (i.e., one whose integral over the surface gives log P_{OW}) based on local properties derived from semiempirical MO calculations that therefore does not require a constant to be added to the surface integral.

Log P_{OW} is itself used as a model for the hydrophobicity of molecules and is used to predict the partitioning behavior of small molecules in living organisms.[11] It is the major descriptor in many QSAR/QSPR equations for predicting pharmacokinetic and pharmacodynamic parameters, toxicity, chemical fate, soil retention times, solubility, and many other properties. It is, for example, part of Lipinski's rule of five, a rule-of-thumb approach for estimating the drug-likeness of organic compounds.[11d] However, it also has a very fundamental importance in determining the binding constants of small molecules to protein receptors. It was the most significant descriptor in most early QSAR equations[12] and has more recently become prominent in the HYDE scoring function.[13] This is because the primary process involved in binding a drug-like (i.e., hydrophobic) molecule to a protein receptor is its transfer from aqueous solution to a less polar environment. Log P_{OW} is a better model for this process than, for instance the free energy of hydration because the ligand is transferred to an environment of approximately the same polarity as octanol, not completely desolvated. The need for reliable methods to predict log P_{OW} is therefore significant, a fact that has also been reflected in the secondary literature.[14]

Log P_{OW} is defined by the difference between the free energies of solvation of the compound in water and in water-saturated *n*-octanol.[15] Log P_{OW} naturally varies with the protonation state, but usually the Log P_{OW} for the neutral compound is used. In practice, this means that the log *P* is measured in a buffered solution in which the compound is mostly neutral (un-ionized or zwitterionic). If the distribution coefficient is given at a defined pH, then it is denoted log D_{pH}. Log $D_{7.4}$, which corresponds to the pH of the blood environment, is sometimes more useful than the log P_{OW} because it ensures that the compounds are being considered in the physiologically relevant protonation state.[16] However, there is no general agreement that either is superior and log P_{OW} is more common. In order to estimate log $D_{7.4}$, the pK_a values of all acid and base centers must be predicted, which introduces a further source of error.

It has also been suggested that a different model solvent than octanol should be used to represent the lipophilic environment.[17] For example, *n*-hexane has been proposed as an even more hydrophobic solvent.[18] *n*-Octanol can donate and accept hydrogen bonds, whereas *n*-hexane is completely nonpolar. However, *n*-octanol has remained the standard for lipophilicity models, so that very many more data for log P_{OW} are available than for partition coefficients between water and other organic phases.

LogP_{OW} can be measured by a variety of techniques[19] that are far less labor-intensive than the original shake-flask method. The most commonly used techniques are correlation with chromatography retention times[20] and automated titration with potentiometric measurements.[21] Results obtained with the different techniques agree well, so that it is justifiable to combine log P_{OW} data measured with different methods for model building and validation.

Many collections of log P_{OW} data exist, for instance in the PHYSPROP,[22] Beilstein,[23] and LOGKOW[24] databases. These data have been used to construct log P_{OW} models using, multiple linear regression,[25] partial least-squares,[26] support vector machines,[27] neural networks, and ensembles thereof[28] in conjunction with a variety of different descriptors. The best known log P_{OW} prediction approaches use topological indices,[28b,29] Ghose−Crippen atom types[5] or fragments.[25,30] Models based on descriptors calculated using ab initio MO-theory or density-functional theory (DFT) have also been published, for example by Haeberlein and Brinck[31] and Chuman et al.[32] However, these models require compute-intensive calculations and have thus only been applied to small to medium-sized data sets. A recent paper by Tetko et al. describes many different log P_{OW} modeling approaches.[14a]

We now describe a surface-integral model for logP that introduces a local hydrophobicity based solely on local properties derived from semiempirical (in this case AM1)[33] MO calculations.

MATERIALS AND METHODS

Data Set. The log P_{OW} training, test, and validation data were obtained from the LOGKOW database.[24] At the end of 2008, it consisted of 37 783 values for 23 479 compounds collected from the literature. The published values are checked and for each compound an average value of all those that pass a quality check is suggested if possible. This results in around 11 500 reliable data. The SMILES structures for those compounds were checked, compared to the structure stored in SciFinder or PubChem and corrected if necessary. Duplicate entries were removed. Only compounds containing H, C, N, O, F, P, S, Cl, Br, and I were kept. Compounds with permanent charges or unpaired electrons were removed. This resulted in 11 102 values. All zwitterionic compounds were removed later, so that 10 814 compounds remained for model generation. 12.5% of the compounds were chosen randomly and put aside as a validation set, which was not used for the model building process. 12.5% was considered adequate because an independent external validation set was also used.

The likely experimental accuracy of the measurements is of importance for assessing the performance of models because this is often limited by the data. Lombardo et al.[34] give data that allow us to determine the errors between two different but accurate experimental methods (shake flask and HPLC) for determining log P_{OW}. Their data gives a calculated mean unsigned error (MUE) of 0.20 and a root-mean-square error (RMSE) of 0.24 log P_{OW} units. It is therefore probably realistic to assume that the experimental error is of the order of ±0.3−0.4 log P_{OW} units. This is likely to be the limit for the attainable accuracy in computational models.

A set of 768 log P_{OW} measurements carried out in-house at Boehringer−Ingelheim was used as a second validation

set that does not overlap with the training/test/validation sets. This set was measured with the GPLK method.[21]

Calculations. First for each compound an initial 3D structure was obtained from the SMILES string using CORINA.[35] All molecules were then subjected to semiempirical geometry optimization using the AM1 Hamiltonian in VAMP.[36] The molecular surfaces and descriptors were calculated with ParaSurf'09[37] based on the VAMP output. In the first approach, the standard ParaSurf descriptors as described in the ParaSurf manual were used. For the surface integral model, the SIM descriptors as described in reference[10] were calculated. These are listed in the Supporting Information.

ParaSurf can calculate several different types of molecular surface. The models described below used the marching-cube surface[6] based on the default isodensity value.[38] The local properties calculated are the molecular electrostatic potential (MEP), local ionization energy (IE_L), local electron affinity (EA_L), local hardness (HARD), local polarizability (POL), and the local field normal to the surface (FN).[39]

The conformational ensemble and the number of rotatable bonds was calculated with the Molecular Operating Environment (MOE).[40]

Fitting Algorithm. All models were generated with a bagging version of stepwise multiple linear regression.[41] Bagging alleviates the effects of noise and outliers in the training set on model generation. It also allows reasonable test set estimates to be obtained for a test set as large as the training set.[42] The critical 95% F-value as calculated by the formula described in ref 43 was used as stopping criterion. In this way, we can avoid overtraining and ensure that only significant variables are included in the models. As the results show, this strategy leads to robust models. Forward and backward stepping were enabled. For each model, 50 independent bagging samples were generated. Each bagging sample was based on a random selection of 75% of the compounds from the total training set. Thus, each compound was part of the test and the training set many times. The results in terms of R^2 and RMSE are given for the test- and training sets. The final formula was created by averaging the coefficients from the 50 single formulae.

RESULTS

In order to test the level of performance that we can expect for this data set using ParaSurf descriptors or properties, we first constructed two models using techniques that we have used before. The first was a descriptor-based model that used the standard ParaSurf descriptor set described in ref 44. The correlation coefficient obtained was R^2(test) = 0.74, the root mean squared error (RMSE) is 0.89 log P units and the mean unsigned error for the test set (MUE(test)) is 0.68 log P units. A total of 34 of the 84 ParaSurf descriptors were used in more than 25 of the 50 bagging equations. The average constant used is 28.52.

The second model was a surface-integral model (SIM) constructed using the techniques described previously.[10] This model gave a squared correlation coefficient for the test set (R^2(test)) = 0.74, a root-mean-square error for the test set (RMSE(test)) = 0.90 and MUE(test) = 0.69. Fifteen of the 126 SIM descriptors are used in more than 25 of the 50 bagging equations, with the most important descriptors being

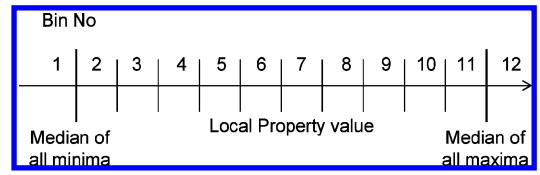

Figure 1. Binning approach.

powers of MEP × EA_L, EA_L, and MEP. The average constant used is 0.58.

In the approach used above to create the SIM model,[10] a polynomial expansion of the local properties (including cross terms) was integrated and the integrals fit to the target. However, this procedure often leads to a fitted equation with a significant constant and a local function that only describes relatively small deviations from this constant value, although the constant in the model described above is small. If the constant is significant, then the function obtained cannot be considered a local hydrophobicity. We have now used a different and more flexible approach; the local properties and their cross-products were binned. For each local property or product, the median values of the maxima and minima for all training compounds were used as the outer binning thresholds. The intervening range was divided into ten bins of equal width. The binning scheme is shown in Figure 1.

This results in twelve bins and eleven thresholds. This binning scheme is similar to the surface histogram bin approach used by Breneman and co-workers for their PEST descriptors,[45] which build upon earlier work on the transferable atom equivalent (TAE) approach by the Breneman group.[46] The six local properties and fifteen cross-products result in 252 new surface-bin descriptors for the twelve bins. The value of each bin descriptor is given by

$$D^N = \sum_{i=1}^{Ntri} \delta_i^N A_i \qquad (2)$$

where D^N is the descriptor associated with bin N of the local property or the cross-product P, whose value for triangle i is P_i, and δ_i has the value one if P_i lies within bin N and zero otherwise. Thus, the final models are built using a pool of 252 descriptors, of which 145 and 154 are retained in the two models reported below.

This set of descriptors was used in a stepwise multiple linear regression to fit experimental log P values. The model has a performance of R^2(test) = 0.84, RMSE(test) = 0.70 and MUE(test) = 0.52. A total of 44 of the 252 descriptors are used in at least 25 of the 50 bagging equations. Among these, MEP × EA_L, local hardness and IE_L bins occur the most often. The most important descriptors according to the sum of the absolute values of the coefficients are (in decreasing order) IE_L × local hardness, IE_L × EA_L, local hardness, IE_L and MEP × EA_L. The average constant used is 0.23. The overall equation is given in the Supporting Information.

Upon analysis of the outliers of this model it turned out that the log P_{OW} of zwitterions (which were calculated in the nonionized form) is predicted on average to be 1.26 log units too high. More general models can be made by weighting the zwitterions higher in the fitting procedure, but this leads to a slight degradation of the performance for the other compounds. Thus, all zwitterions that were identified with the ACDlabs pK_a prediction tool[47] were removed from

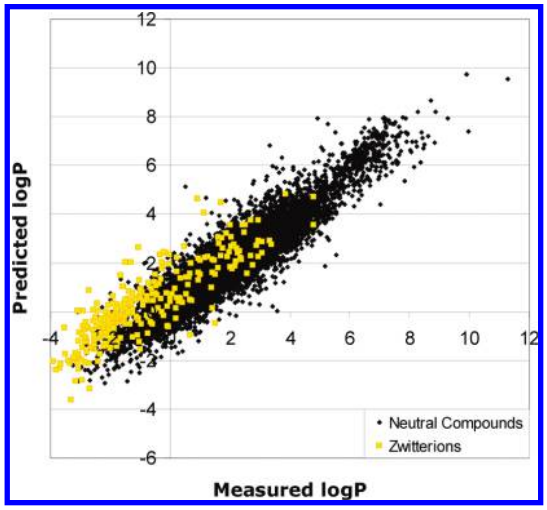

Figure 2. Measured versus predicted log P_{OW} values for the test set.

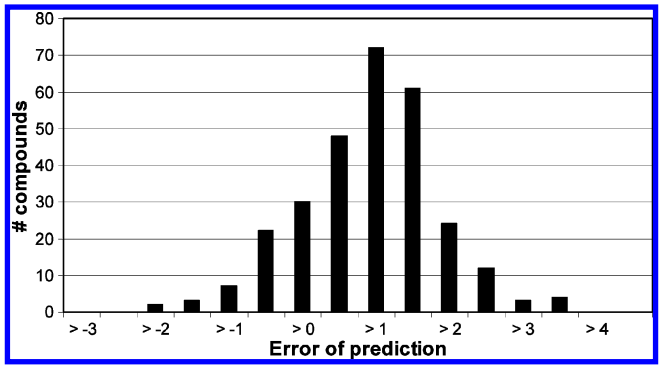

Figure 3. Error of prediction for zwitterions in the training set. Importance of the different basic descriptors as sum the absolute values of the corresponding bin coefficients.

the data set. The remaining 10 814 compounds were used to generate a new stepwise multiple linear regression model. This model has a performance of R^2(test) = 0.86, RMSE(test) = 0.64 and MUE(test) = 0.48. Here, 46 of the 252 descriptors are used in at least 25 of the 50 bagging equations. Among these, local hardness, MEP × EA_L, IE_L × EA_L and IE_L bins occur the most often. The most important descriptors according to the sum of the absolute values of the coefficients are (in decreasing order) IE_L × EA_L, local hardness, MEP × EA_L, IE_L and IE_L × local hardness. The average constant used is 0.12. The overall equation is given in the Supporting Information. A plot of predicted versus measured values for the out-of-bag test set predictions is shown in Figure 2.

The log P_{OW} of zwitterions is on average predicted to be 1.26 log units too high with a normal distribution. Subtracting 1.26 from each zwitterion prediction gives a specific zwitterion model with R^2 = 0.74, RMSE = 0.93 and MUE = 0.66. The original errors of prediction of zwitterions are shown in Figure 3.

Figure 3 shows that the error of $logP_{OW}$ prediction for zwitterions is approximately normally distributed, as would be expected for an otherwise functional model with a systematic error for a given compound type.

The performance for all four models generated for training and out-of-bag test sets are shown in Table 1. As the models do not allow estimates of likely errors for individual compounds, a standard model error equal to the root-mean-square error (RMSE) must be assumed.

The bin thresholds and coefficients of the final formula are given in the Supporting Information. The constant term is very close to zero, which means that if a molecule is reduced to having no surface the likeliness for distributing into water and octanol becomes equal. Thus, the model represents a true local hydrophobicity.

The model generated above was validated with the validation set split off before model generation. The validation set is also predicted with MUE = 0.48, RMSE = 0.64, R^2 = 0.86. The model was further validated with a second validation set consisting of compounds measured in-house at Boehringer−Ingelheim. This validation set was predicted with MUE = 0.85, RMSE = 1.10, R^2 = 0.53.

Comparison with Publicly Available log *P* Models. The two validation sets were predicted with other publicly available log *P* prediction tools. These were ACDlabs log *P* prediction,[30d] *C*log *P* from Biobyte,[48] log $P_{o/w}$ and *S*log *P* available in MOE[40] and AlogP available within TSAR.[49] We used all predictions for comparison irrespective of the predicted accuracy. The performance is summarized in Table 2.

As might be expected from solubility studies,[50] there is a striking difference in the performance of the models for the public and on the in-house data sets. These data sets are very different in nature; the public data set consists mostly of compounds with a low molecular weight. The in-house data set consists of drug-like molecules from project work which have a smaller variance (1.89 vs 2.89) and a higher average value (3.69 vs 2.05) than the public validation set. It is also not clear whether some of the public models were trained with data from the public validation set, so that we can only be certain that the in-house data set is a real validation set for these models.

Variable Importance. The variable importance can be obtained as the sum of the absolute values of coefficients of the descriptor derived from a given property or a cross-product of two properties. These importance measures are shown in Figure 4.

The molecular electrostatic potential, MEP, local ionization energy, and electron affinity, IE_L, EA_L, respectively, and the local hardness, HARD and cross-products between them are the most important descriptors. No single local property dominates, but the local polarizability, POL, and the field normal to the surface, FN, do not play a significant role.

Conformational Dependence. A recurring question about QSPR-models of any sort based on 3D-molecular structures is that of the conformational dependence of the results. Mobley et al.[51] for instance have recently shown that in molecular dynamics simulations conformational entropy changes at room temperature for relatively small molecules can be as high as 2.3 kcal mol^{-1} and that there is no correlation with the number of rotatable bonds. The situation must be similar of log *P*. The usual procedure in constructing such models is to start with 2D structures (usually SMILES strings) and to use one of several commercial 2D → 3D conversion programs to produce a single conformation for each compound. This is usually an unavoidable source of error (because the real conformations of the training compounds are not known), whose effect is demonstrated in Figure 5, which shows the MUE of subsets of the training

Table 1. Measures of Performance of All Models Generated (MUE = Mean Unsigned Error, RMSE = Root Mean Square Error, R^2 = the Square of the Correlation Coefficient)

	training set			test set		
model	MUE	RMSE	R^2	MUE	RMSE	R^2
ParaSurf descriptors	0.67	0.88	0.75	0.68	0.89	0.74
polynomial SIM descriptors	0.68	0.89	0.75	0.69	0.90	0.74
binned SIM descriptors	0.51	0.68	0.85	0.52	0.70	0.84
binned SIM descriptors, no zwitterions	0.48	0.62	0.86	0.48	0.64	0.86

Table 2. Performance of Other Publicly Available log $P_{o/w}$ Models for the Public and in-House Validation Sets (MUE = Mean Unsigned Error, RMSE = Root Mean Square Error, R^2 = the Square of the Correlation Coefficient)

	public validation set			in-house validation set		
model	MUE	RMSE	R^2	MUE	RMSE	R^2
SIM-log *P*	0.48	0.64	0.86	0.85	1.10	0.53
ACDlabs log *P*	0.26	0.45	0.94	1.03	1.36	0.47
*C*log *P*	0.31	0.52	0.92	0.86	1.14	0.59
*S*log *P*	0.53	0.68	0.85	0.92	1.19	0.51
log $P_{o/w}$	0.53	0.77	0.82	1.00	1.28	0.49
*A*log *P*	0.62	0.86	0.79	0.97	1.24	0.50

set sorted by surface area and divided into bins of 200 compounds each.

For the complete data set (gray squares), the RMSE rises essentially linearly with increasing surface area. This behavior could result from increasing uncertainty about the experimental conformation(s) with increasing compound size, but could also be a simple result of the uncertainty in the fitted regression coefficients, which can also give larger errors as the individual descriptors (binned areas) become larger.

The log P_{OW} data set used here allows us to resolve this question. Of the total data set, 1581 compounds contain no rotatable bonds. When potentially flexible macrocycles and compounds with extremely high log P_{OW} values, which are hard to measure, are also removed, a data set of 1563 compounds without significant conformational flexibility remains. The black points in Figure 5 show the MUE for these compounds binned as for the complete data set. The MUE for these compounds (0.35 − 0.40 log *P* units) is lower than for the complete data set but, more significantly, it does not increase with increasing surface area. It also corresponds to our estimate of the maximum possible model performance given above. Thus, the increasing error level shown for the complete data set is most likely due to increasing conformational uncertainty. Table 3 shows the performance of the "single conformation" model for its own out-of-bag test set, the full training set, the public validation set and the in-house validation set.

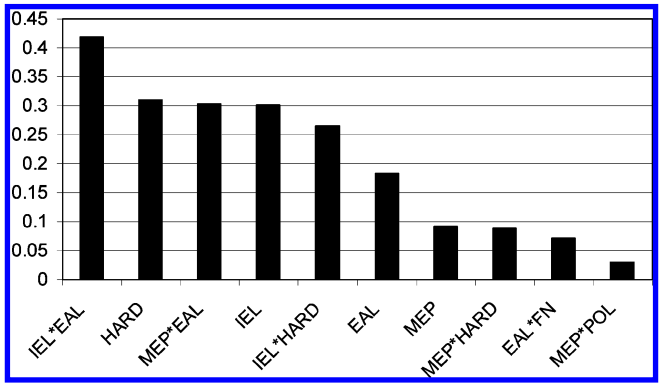

Figure 4. Importance of the different basic descriptors as a sum of the absolute values of the corresponding bin coefficients.

DISCUSSION

The log P_{OW} model generated is a 3D QSPR model and as such conformation-dependent. However, in contrast to most QSPR target properties, enough log P_{OW} data exist for conformationally rigid molecules that we can construct a model (denoted "single conformation" above) that contains specific conformationally dependent information. We expect this model to perform well for compounds whose conformation is known, but less so for flexible molecules, for which such a model should be used with Boltzmann-weighted conformational ensembles. We have examined the prediction performance of the two SIM-models outlined above for Omeprazole, Risperidone, and Haloperidol, for which reliable log P_{OW} values have been reported and which all include central rotatable bonds that make large conformational changes possible. They are shown in Figure 6.

All possible conformations were generated systematically with MOE by using all combinations of 120° rotations around the single bonds that do not produce steric clashes. All conformations were optimized semiempirically with AM1 in vacuo. Two different energies, the heats of formation in vacuo and those obtained using a self-consistent reaction field (SCRF) correction for solvation in water,[52] were used in two alternative Boltzmann-weighting schemes. The SCRF cal-

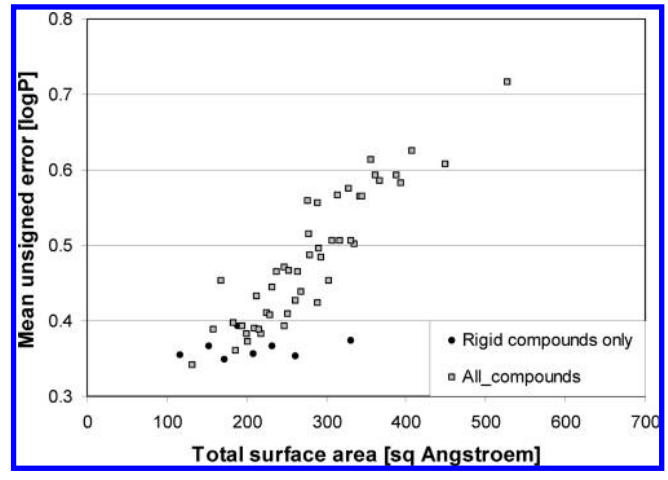

Figure 5. Error of prediction versus total surface area.

Table 3. Performance of the "Single Conformation" Model for All Data(Sub)Sets Examined[a]

	single-conformation training set	full training set	public validation set	in-house validation set
no. of compounds, N	1563	9459	1355	768
MUE	0.37	0.66 (0.48)	0.66 (0.48)	0.93 (0.85)
RMSE	0.51	0.91 (0.62)	0.90 (0.64)	1.20 (1.10)
R^2	0.93	0.76 (0.86)	0.78 (0.68)	0.52 (0.53)

[a] Values in parentheses are those given by the conformationally averaged model for comparison (MUE = mean unsigned error, RMSE = root mean square error, R^2 = the square of the correlation coefficient).

Table 4. Comparison of Log P_{OW} Values Calculated with the Alternative SIM Models Using a Single Conformation Obtained from CORINA and Conformational Ensembles

				boltzmann weighted	
compound	experiment	model	CORINA conformation	gas-phase	SCRF
omeprazole	2.23[53]	total data set	2.94	2.97	2.65
	2.38[54]	single conf.	2.92	2.68	2.51
risperidone	3.04[55]	total data set	2.48	3.20	3.14
		single conf.	2.78	3.17	3.03
haloperidol	4.30[55]	total data set	4.52	4.23	4.19
		single conf.	4.17	4.20	4.18

culations used the geometries optimized in vacuo and did not include the dispersion correction. Log P_{OW} predictions were calculated for each microspecies and the overall log P_{OW} was obtained as the Boltzmann-weighted sum of these values. The results are summarized in Table 4.

The performance of the models confirms our expectations. The prediction of the single-conformation model improves with Boltzmann weighting and is usually best using the SCRF-corrected energies for the weighting. There is usually a rather large difference between the CORINA conformation and the conformation with the lowest energy; the CORINA conformation is always extended, while the best conformation according to SCRF energy is U-shaped or folded.

The performance of the model obtained from the total data set also improves with Boltzmann weighting. Serendipitously, the change in prediction on Boltzmann weighting is negative for omeprazole, positive for risperidone and essentially zero for haloperidol using the single-conformation model. However, the calculated log P_{OW} for haloperidol using the conformationally averaged (total data set) model does decrease significantly using Boltzmann weighting. These three examples suggest, perhaps surprisingly, that Boltzmann averaging can also improve the predictions of conformationally averaged models, which clearly encode some conformational information simply because the rigid compounds

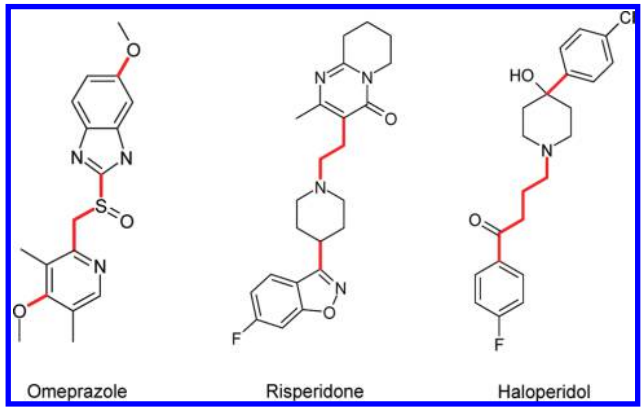

Figure 6. Drugs for which a conformational ensemble log P evaluation was carried out. Rotated bonds are shown in red.

are also part of their training sets. The single-conformation model performs marginally better than the conformationally averaged one for these three compounds, but the difference is not large and the sample far too small to draw any conclusions. However, the single-conformation model has a firmer physical basis and is also statistically very conservative. It was only trained on ~1500 compounds, rather than ~9500 for the conformationally averaged model and still performs acceptably for the remaining 8000 compounds.

Clearly, a systematic conformational search followed by Boltzmann weighting in aqueous solution represents far more computational work than would normally be invested in a log P_{OW} prediction. Nonetheless, the question of the conformational dependence of properties predicted by QSPR models has often been posed and seldom answered. We[56] reported a systematic study of the effect of conformation on the predicted boiling point of a single compound using a conformationally averaged model, but the fluctuations were well within the error limits of the predictions. In the present case, we have the luxury of adequate data for conformationally rigid compounds, so that more concrete conclusions can be drawn. For the three drugs used for the test, the single-conformation model with AM1-SCRF Boltzmann weighting gives a mean unsigned error of zero and a standard error of 0.2 log units. The standard error of the experimental data is likely to be of the same order of magnitude.

Thus, we estimate that the single-conformation model is probably as accurate as currently possible with the exception of compounds, such as those with nitro groups, which are not included in the rigid compound training set and occupy a specific region in the descriptor space. Limitations are likely to arise from the approximation that molecules exist in the same conformation and protonation state in water and in wet *n*-octanol. This assumption will necessarily limit all in silico models that do not treat the two phases separately.

We have emphasized that the small constant term allows us to consider the function to be integrated in a surface-integral model as a true local hydrophobicity. However, it is derived from a regression scheme and may therefore not

represent physical reality. We will analyze this aspect of the model in depth in a later manuscript.

CONCLUSIONS

We have generated a surface integral model for log *P* based on quantum-mechanical surface properties. The performance of the model with its RMSE(test) = 0.64 and MUE = 0.48 is as good as the best conservative models published based on other modeling strategies. The model is based on only six different basic descriptors, which can be computed for each compound. This makes it very robust. The robustness can be seen from the very similar performances for test- and validation data sets.

Zwitterions and compounds containing permanent charges cannot be predicted, but a reasonable model for zwitterions involves subtracting a constant 1.26 from the calculated log P_{OW} for the nonionized protonation state.

At first sight, the overall error of prediction seems to be related to the total surface area of the compound. However, we also generated a single-conformation model based on rigid compounds. In this case, the error of prediction remains constant with increasing total surface area.

The single-conformation SIM does not perform much worse than the full model and appears to be very accurate if complete conformational searches and Boltzmann weighting are used. However, this conclusion is based on test calculations for only three compounds. A fully Boltzmann-weighted model would require considerable computer resources.

The approach demonstrated here is also feasible for most of the other QSPR properties and we will explore more applications in future. It may also be possible to improve the models systematically by using more modern semiempirical Hamiltonians such as AM1*[57] or PM6.[58] We also plan to investigate the benefits which can be obtained from them.

ACKNOWLEDGMENT

This work was supported by Boehringer−Ingelheim Pharma & Co. KG.

Supporting Information Available: Parameters and thresholds for the different SIM models are included in the SI. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Pixner, P.; Heiden, W.; Merx, H.; Moeckel, G.; Moeller, A.; Brickmann, J. Empirical Method for the Quantification and Localization of Molecular Hydrophobicity. *J. Chem Inf Comp Sci* **1994**, *34*, 1309–1319.

(2) Politzer, P.; Murray, J. S.; Concha, M. C. σ-Hole bonding between like atoms; a fallacy of atomic charges. *J. Mol. Model.* **2008**, *14*, 659–665.

(3) Abraham, R. J.; Hudson, B. D.; Kermode, M. W.; Nines, J. R. A General Calculation of Molecular Solvation Energies. *J. Chem. Soc. Faraday Trans.* **1988**, *84*, 1911–1917.

(4) Audry, E.; Dubost, J. P.; Colleter, J. C.; Dallet, P. A new approach of structure activity relationships: The "potential of molecular lipophilicity". *Eur. J. Med. Chem.* **1986**, *21* (1), 71–72.

(5) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. I. Partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–677.

(6) Heiden, W.; Goetze, T.; Brickmann, J. Fast generation of molecular surfaces from 3D data fields with an enhanced "marching cube" algorithm. *J. Comput. Chem.* **1993**, *14*, 246–250.

(7) (a) Richards, N. G. J.; Williams, P. B.; Tute, M. In *Empirical Methods for Computing Molecular Partition Coefficients. I. Upon the Need to Model the Specific Hydration of Polar Groups in Fragment-Based Approaches*; Löwdin, P., Ed.; International Journal of Quantum Chemistry: Quantum Biology Symposium, 1991; pp 299−316; (b) Richards, N. G. J.; Williams, P. B.; Tute, M. Empirical methods for computing molecular partition coefficients: II. Inclusion of conformational flexibility within fragment-based approaches. *Int. J. Quantum Chem.* **1992**, *44*, 219–233.

(8) 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* (1), 237–247.

(9) Ehresmann, B.; de Groot, M. J.; Alex, A.; Clark, T. New Molecular Descriptors Based on Local Properties at the Molecular Surface and a Boiling-Point Model Derived from Them. *J. Chem. Inf. Comp. Sci.* **2004**, *43*, 658–668.

(10) Ehresmann, B.; de Groot, M. J.; Clark, T. Surface-Integral QSPR Models: Local Energy Properties. *J. Chem. Inf. Model* **2005**, *45*, 1053–1060.

(11) (a) Dearden, J. C. Partitioning and lipophilicity in quantitative structure-activity relationships. *Environ. Health Perspect.* **1985**, *61*, 203–228. (b) Hansch, C.; Leo, A.; Hoekman, D., *Hydrophobic, Electronic, and Steric Constants*; American Chemical Society: Washington DC, 1995; (c) Kubinyi, H. Lipophilicity and drug activity. *Prog. Drug Res.* **1979**, *23*, 97–198. (d) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Res.* **1997**, *23*, 3–25. (e) Pliska, V.; Testa, B.; van de Waterbeemd, H., *Lipophilicity in Drug Action and Toxicology*; Wiley-VCH: Weinheim, 1997; (f) van de Waterbeemd, H.; Smith, D. A.; Beaumont, K.; Walker, D. K. Property-based design: optimization of drug absorption and pharmacokinetics. *J. Med. Chem.* **2001**, *44* (9), 1313–1333. (g) Tetko, I. V., Prediction of physicochemical properties. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; John Wiley & Sons, Inc: New Jersey, 2007; pp 241−275; (h) Dearden, J. C. In silico prediction of ADMET properties: how far have we come. *Expert Opin. Drug Metab. Toxicol.* **2007**, *3* (5), 635–639. (i) Gleeson, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008**, *51* (4), 817–34. (j) Sangster, J. Octanol−water partition coefficients of simple organic compounds. *J. Phys. Chem. Ref. Data* **1989**, *18* (3), 1111–1229. (k) Hansch, C.; Bjorkroth, J. P.; Leo, A. Hydrophobicity and central nervous system agents: on the principle of minimal hydrophobicity in drug design. *J. Pharm. Sci.* **1987**, *76*, 663–687.

(12) Hansch, C.; Leo, A., *Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(13) Reulecke, I.; Lange, G.; Albrecht, J.; Klein, R.; Rarey, M. Towards an Integrated Description of Hydrogen Bonding and Dehydration: Decreasing False Positives in Virtual Screening with the HYDE Scoring Function. *ChemMedChem* **2008**, *3*, 885–897.

(14) (a) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity; State-of-the-art and comparison of log P methods on more than 96 000 compounds. *J. Pharm. Sci.* **2008**, *98*, 861–893. (b) Klopman, G.; Zhu, H. Recent methodologies for the estimation of n-octanol/water partition coefficients and their use in prediction of membrane transport properties of drugs. *Mini Rev. Med. Chem.* **2005**, *5* (2), 127–133.

(15) Sangster, J., *Octanol-Water Partition Coefficients: Fundamentals and Physical Chemistry*; John Wiley & Sons Ltd.: Chichester, 1997; Vol. 2.

(16) Bhal, S. K.; Kassam, K.; Peirson, I. G.; Pearl, G. M. The Rule of Five revisited: applying log *D* in place of log *P* in drug-likeness filters. *Mol. Pharm.* **2007**, *4* (4), 556–560.

(17) (a) Leo, A.; Hansch, C.; Elkins, D. Partition coefficients and their uses. *Chem. Rev.* **1971**, *71* (6), 525–616. (b) Leahy, D. E.; Taylor, P. J.; Wait, A. R.; Model Solvent Systems for QSAR Part, I. Propylene Glycol Dipelargonate (PGDP). A new Standard Solvent for use in Partition Coefficient Determination. *Quant. Struct.-Act. Relat.* **1989**, *8* (1), 17–31.

(18) Schulte, J.; Dürr, J.; Ritter, S.; Hauthal, W. H.; Quitzsch, K.; Maurer, G. Partition Coefficients for Environmentally Important, Multifunctional Organic Compounds in Hexane + Water. *J. Chem. Eng. Data* **1998**, *43* (1), 69–73.

(19) Dearden, J. C.; Bresnen, G. M. The Measurement of Partition Coefficients. *QSAR Comb. Sci* **2006**, *7* (3), 133–144.

(20) Valko, K. Application of high-performance liquid chromatography based measurements of lipophilicity to model biological distribution. *J. Chromatogr. A* **2004**, *1037* (1−2), 299–310.

(21) Takacs-Novak, K.; Avdeef, A. Interlaboratory study of log *P* determination by shake-flask and potentiometric methods. *J. Pharm. Biomed. Anal.* **1996**, *14* (11), 1405–1413.

(22) The physical properties database (PHYSPROP). Syracuse research corporation.
(23) *CrossFire Beilstein*; Elsevier: Frankfurt, 2009; Vol. 7.1.
(24) Sangster, J., *LOGKOW -A databank of evaluated octanol-water partition coefficients (Log P)*; Sangster Research Laboratories: Montreal, Quebec, accessed 11/23/2008.
(25) Nys, G. G.; Rekker, R. F. The concept of hydrophobic fragmental constants (f-values). II. Extension of its applicability to the calculation of lipophilicities of aromatic and hetero-aromatic structures. *Chim. Ther.* **1973**, *9*, 361–374.
(26) Liu, R.; Zhou, D. Using Molecular Fingerprint as Descriptors in the QSPR study of Lipophilicity. *J. Chem. Inf. Model.* **2008**, *48* (3), 542–549.
(27) Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log *P*. *J. Chem. Inf. Model.* **2008**, *48* (1), 220–232.
(28) (a) Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. Prediction of the *n*-octanol/Water Partition Coefficient, logP, Using a Combination of Semiempirical MO-Calculations and a Neural Network. *J. Mol. Model.* **1997**, *3*, 142–155. (b) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. Prediction of *n*-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J. Chem. Inf. Comp. Sci.* **2001**, *41*, 1407–1421.
(29) (a) Huuskonen, J. J.; Villa, A. E.; Tetko, I. V. Prediction of partition coefficient based on atom-type electrotopological state indices. *J. Pharm. Sci.* **1999**, *88* (2), 229–233. (b) Livingstone, D. J.; Ford, M. G.; Huuskonen, J. J.; Salt, D. W. Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. *J. Comput. Aided Mol. Des.* **2001**, *15* (8), 741–752.
(30) (a) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated log *P* Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comp. Sci.* **1994**, *34* (4), 752–781. (b) Meylan, W. M.; Howard, P. H. Estimating log *P* with atom/fragments and water solubility with log *P*. *Perspect Drug Discov. Des.* **2000**, *19* (1), 67–84. (c) Leo, A.; Jow, P. Y.; Silipo, C.; Hansch, C. Calculation of hydrophobic constant (log *P*) from π and f constants. *J. Med. Chem.* **1975**, *18* (9), 865–868. (d) Petrauskas, A. A.; Kolovanov, E. A. ACD/Log *P* method description. *Perspect Drug Discov. Des.* **2000**, *19* (1), 99–116.
(31) Haeberlein, M.; Brinck, T. Prediction of water-octanol partition coefficients using theoretical descriptors derived from the molecular surface area and the electrostatic potential. *J. Chem. Soc., Perkin Trans.* **1997**, *2*, 289–294.
(32) Chuman, H.; Mori, A.; Tanaka, H.; Yamagami, C.; Fujita, T. Analyses of the partition coefficient, log *P*, using ab initio MO parameter and accessible surface area of solute molecules. *J. Pharm. Sci.* **2004**, *93* (11), 2681–2697.
(33) (a) 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. (b) 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.; Wiley: Chichester, 1998; pp 8−11.
(34) Lombardo, F.; Shalaeva, M. Y.; Tupper, K. A.; Gao, F.; Abraham, M. H. ElogPoct: A Tool for Lipophilicity Determination in Drug Discovery. *J. Med. Chem.* **2000**, *43*, 2922–2928.
(35) *CORINA 3.4*; Molecular Networks Inc.: Erlangen, Germany, 2006.
(36) Clark, T.; Alex, A.; Beck, A.; Burkhardt, F.; Chandrasekhar, J.; Gedeck, P.; Horn, A. H. C.; Hutter, M.; Martin, B.; Rauhut, G.; Sauer, W.; Schindler, T.; Steinke, T. *VAMP 8.2, available from Accelrys Inc.*; Erlangen: San Diego, USA, 2002.
(37) *ParaSurf09*, CEPOS Insilico Ltd.: Erlangen, Germany, 2009.
(38) Meyer, A. Y. The size of molecules. *Chem. Soc. Rev.* **1985**, *15*, 449–475.
(39) 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, May 15−19, 2006* (Logos Verlag:), Berlin, 2008; pp 129−146.
(40) Labute, P. *Molecular Operating Environment*, 2008.10; Chemical Computing Group: Montreal, Quebec, Canada, 2008.
(41) Efroymson, M. A., Multiple regression analysis. In *Mathematical Methods for Digital Computers*; Ralston, A., Milf, H. A., Eds. Wiley: New York, 1960; Vol. 1, pp 191−203.
(42) Polikar, R. Ensemble based systems in decision making. *IEEE Circ. Sys. Mag.* **2006**, *03/06*, 21–45.
(43) 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* (1), 28–34.
(44) Kramer, C.; Beck, B.; Kriegl, J. M.; Clark, T. A composite model for hERG blockade. *ChemMedChem* **2008**, *3* (2), 254–265.
(45) Breneman, C. M.; Sundling, C. M.; Sukumar, N.; Shen, L.; Katt, W. P.; Embrechts, M. J. New Developments in PEST Shape/Property Hybrid Descriptors. *J. Comp. Aided Mol. Des.* **2003**, *17*, 231–240.
(46) Breneman, C.; Rhem, M. QSPR analysis of HPLC column capacoty factors for a set of high-energy materials using electronic van der Waals surface property descriptors computed by the transferable atom equivalent method. *J. Comput. Chem.* **1997**, *18*, 182–197.
(47) *ACD/PhysChem Suite 11.0*, ACD/Labs: Toronto, Canada.
(48) *BioByte Inc. ClogP, 4.0*; Daylight Chemical Information Systems Inc.: Aliso Viejo, CA, 2008.
(49) *TSAR, 3.3*; Oxford Molecular Ltd., Oxford, UK; now Accelrys Inc., San Diego, CA: 2000.
(50) Schroeter, T. S.; Schwaighofer, A.; Mika, S.; Ter Laak, A.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Muller, K. R. Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput. Aided Mol. Des.* **2007**, *21* (12), 651–64.
(51) Mobley, D. L.; Dill, K. A.; Chodera, J. D. Treating Entropy and Conformational Changes in Implicit Solvent Simulations of Small Molecules. *J. Phys. Chem. B* **2008**, *111*, 938–946.
(52) 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.
(53) Craig, P. N., *Drug Compendium*; Pergamon Press: New York, 1990; Vol. 6.
(54) Ungell, A.-L.; Nylander, S.; Bergstrand, S.; Sjöberg, A.; Lennernäs, H. Personal Communication, cited in: Membrane Transport of Drugs in Different Regions of the Intestinal Tract of the Rat. *J. Pharm. Sci.* **1998**, *87* (3), 360–366.
(55) Leysen, J. E.; Janssen, P. M. F.; Gommeren, W.; Wynants, J.; Pauwels, P. J.; Janssen, P. A. J. Vitro and In Vivo Receptor Binding and Effects on Monoamine Turnover in Rat Brain Regions of the Novel Antipsychotics Risperidone and Ocaperidone. *Mol. Pharmacol.* **1992**, *41* (3), 494–508.
(56) Chalk, A. J.; Beck, B.; Clark, T. A Quantum Mechanical/Neural Net Model for Boiling Points with Error Estimation. *J. Chem. Inf. Comp. Sci.* **2001**, *41*, 457–462.
(57) Winget, P.; Horn, A. H. C.; Selcuki, C.; Martin, B.; Clark, T. AM1* Parameters for Phosphorous, Sulfur and Chlorine. *J. Mol. Model.* **2003**, *9*, 408–414.
(58) 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.

CI900431F