Application of ALOGPS 2.1 to Predict log *D* Distribution Coefficient for Pfizer Proprietary Compounds

Igor V. Tetko*,† and Gennadiy I. Poda‡

Biomedical Department, Institute of Bioorganic and Petroleum Chemistry, Ukrainian Academy of Sciences, Murmanskaya 1, Kyiv, 02094, Ukraine, and Structural and Computational Chemistry, Pfizer Global Research and Development, 700 Chesterfield Parkway West, Chesterfield, Missouri 63017

Received June 22, 2004

Abstract: Evaluation of the ALOGPS, ACD Labs LogD, and PALLAS PrologD suites to calculate the log *D* distribution coefficient resulted in high root-mean-squared error (RMSE) of 1.0−1.5 log for two *in-house* Pfizer's log *D* data sets of 17 861 and 640 compounds. Inaccuracy in log *P* prediction was the limiting factor for the overall log *D* estimation by these algorithms. The self-learning feature of the ALOGPS (LIBRARY mode) remarkably improved the accuracy in log *D* prediction, and an rmse of 0.64−0.65 was calculated for both data sets.

Oral bioavailability of chemicals is a very important pharmacokinetic parameter in drug development. To reach the target enzyme in the human body, drugs have to cross barriers by passive diffusion or carrier-mediated uptake. The 1-octanol−water partition coefficient, log *P*, is well-known as one of the principal parameters to estimate lipophilicity (or solubility in lipids) of chemical compounds and, to a large degree, determines their pharmacokinetic properties. The log *P* is also used as one of the standard properties identified by Lipinski in the "rule of 5" for druglike molecules.[1] By definition log *P* refers to neutral molecules. If a molecule contains basic or acidic groups, it becomes ionized and its distribution in octanol−water becomes pH-dependent. The pH-dependent distribution coefficient, log *D*, was shown to correlate with a number of biological parameters, such as the effective permeability in human jejunum,[2] blood−brain barrier (BBB) permeability,[3] plasma protein binding,[4] CYP 450 oxidation,[5] and volume of distribution (V_D).[6,7] Oral drugs, to be able to be absorbed by passive diffusion through the gut wall, should have their lipophilicity within a given range (usually between 1 and 4 on the log *D* scale).

Both coefficients log *P* and log *D* are very important parameters in drug development,[8] and thus, there is a need to develop new methods to accurately calculate them from chemical structures. Currently, the amount of publicly available experimental log *P* data comprises tens of thousands of compounds.[9] These resources stimulated development of a number of programs to calculate it.[10−15] The problem of predicting log *D* is more complicated. As a rule, it is computed from log *P* and pK_a assuming that only the neutral form partitions into the organic phase as[12,16]

$$\log D(\mathrm{pH}) = \log P - \log(1 + 10^{(\mathrm{pH}-\mathrm{pK_a})\Delta_i}) \qquad (1)$$

where $\Delta_i = \{1, -1\}$ for acids and bases, respectively.

If several groups can be ionized, the equation is modified accordingly to incorporate correction terms for all of them. Thus, the log *D* prediction potentially accumulates errors due to the log *P* and pK_a predictions. Development of computational approaches is further complicated because of the absence of publicly available large data sets with experimental log *D* values. As a result, only a few programs are available to estimate the log *D*.[12] A recent evaluation of two commercial programs calculated a root-mean-squared error (rmse) of 1.4−1.9 log units for a data set of about 20 000 compounds[17] that is not accurate for practical usage. Therefore, large pharmaceutical companies such as Pfizer and AstraZeneca have established their own techniques to experimentally determine log *D* for their proprietary compounds.

The ALOGPS program[18−20] (http://www.vcclab.org) was developed using the associative neural network (ASNN) method.[21,22] The ASNN provides a possibility to include new data into the memory of neural nets without retraining the neural networks themselves in the so-called LIBRARY mode (further LIBRARY).[19] The LIBRARY dramatically improved prediction of the ALOGPS program for the log *P* prediction using in-house data sets from BASF,[21] Pfizer,[23] and AstraZeneca.[24,25] The current study demonstrates that the ALOGPS is also able to reliably predict the pH-dependent distribution coefficient, log *D*.

The octanol−water partition data used in this study was collected at two Pfizer sites and contributed to two data sets. The first data set included 669 legacy Pharmacia compounds with log *D* values measured by a medium-throughput method using a nitrogen detector (called the NlogD set). A typical experimental error in log *D* measurements is about 0.3−0.5 log units. The second data set (ElogD set) included 18 889 compounds measured using the ElogD method.[26,27] An inspection of compounds indicated that both sets were not overlapping. For compounds that had multiple measurements average values were used. Also, because the ALOGPS method does not take into account stereoselectivity, average values were used for stereoisomers. After removal of structural duplicates and stereoisomers, the numbers of compounds decreased to 640 and 17 861 for NlogD and ElogD data sets, respectively.

For comparison, ACD Labs LogD v.7.19[28] and PALLAS PrologD software[29] was used to calculate log *D* values at pH 7.4 for ElogD and NlogD data sets. The stand-alone graphical-based interface versions of ALOGPS and ASNN were used to perform analysis of compounds using three protocols.

In the first protocol, the ALOGPS program was used "as is" to calculate a blind prediction of molecules from each data set.

In the second protocol, the self-learning feature implemented as a "LIBRARY" mode of ALOGPS 2.1 was

* To whom correspondence should be addressed. Address: Institute for Bioinformatics GSF, Forschungszentrum für Umwelt und Gesundheit, GmbH, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany. Phone: +49-89-3187-3575. Fax: +49-89-3187-3585. E-mail: itetko@vcclab.org.

† Institute of Bioorganic and Petroleum Chemistry.

‡ Pfizer Global Research and Development.

10.1021/jm049509l CCC: $27.50
Published on Web 10/05/2004

Table 1. Prediction Performance of Programs for ElogD Data Set

methods description	n^a	rmse	MAE	% compds within rmse range 0−0.3	0−0.5	0−1.0	0−2.0
ACD Labs LogD, pH 7.4	17 341	1.32	0.97	21	35	63	89
ACD Labs logP	17 848	1.38	1.08	19	30	55	85
Pallas PrologD, pH 7.4	17 800	1.41	1.06	19	31	58	87
Pallas PrologP	17 860	1.52	1.21	15	25	50	80
ALOGPS "as is" blind prediction	17 861	1.17	0.92	21	35	62	91
ALOGPS LOO for all compds used in the LIBRARY	17 861	0.64	0.43	50	70	91	98
ALOGPS LOO for 50% compds used in the LIBRARY	8 931	0.69	0.48	45	65	88	98
ALOGPS prediction of 50% remaining compds	8 930	0.69	0.48	46	66	88	98

[a] Different numbers of compounds in this column are due to failure in processing some chemical structures by ACD Labs LogD or Pallas PrologD suites.

Table 2. Prediction Performance of Programs for NlogD Data Set

methods description	n	rmse	MAE	% compds within rmse range 0−0.3	0−0.5	0−1.0	0−2.0
ACD Labs LogD, pH 7.4	576	0.99	0.69	27	48	79	95
ACD Labs logP	639	1.14	0.80	27	45	74	92
PALLAS PrologD, pH 7.4	640	1.52	1.29	8	15	41	84
PALLAS PrologP	640	1.46	1.20	10	19	46	86
ALOGPS "as is" blind prediction	640	1.33	1.09	15	22	50	89
ALOGPS LOO for all compds used in the LIBRARY	640	0.65	0.42	54	70	90	98
ALOGPS LOO for 50% compds used in the "random" LIBRARY	320	0.66	0.44	52	68	88	98
ALOGPS prediction of 50% remaining compds	320	0.68	0.45	52	73	89	98
ALOGPS blind prediction using ElogD set as the LIBRARY	640	1.58	1.29	14	23	43	77
ASNN LOO for 50% compds used to retrain neural networks	320	0.49	0.37	52	75	95	100
ASNN prediction of 50% test set compds	320	0.57	0.42	49	73	94	99

used to train the program using 50% randomly chosen compounds from each data set. The other 50% were used for prediction. The input data set for the LIBRARY was a flat file with a SMILES string per molecule per line. Each SMILES was followed by the log *D* experimental value. The program selected smoothing parameters for the nearest neighbors from the LIBRARY and then predicted log *D* for new compounds. It also estimated statistical parameters (in the "leave-one-out", or LOO, mode) for the compounds used in the LIBRARY mode. The calculation of the LIBRARY for the ElogD set and prediction of all molecules required about 10 min on a Pentium M 1.5 GHz IBM ThinkPad T40 computer.

In the third protocol, a file with 75 E-state indices[30,31] used to train the ALOGPS program was generated for the NlogD set. These indices were applied to train the stand-alone version of ASNN and, thus, create new models to predict log *D*. For this study exactly the same ASNN parameters were utilized that were used to develop the ALOGPS program: 5 hidden neurons, sigmoid activation function, 64 neural networks per ensemble, Levenberg−Marguardt training algorithm.[19] We used 50% of the randomly chosen molecule to develop the model. At the end of training the ASNN performance was evaluated on 50% of the remaining molecules.

In the "as is" mode the ALOGPS program showed rmse of 1.17 and 1.33 for the ElogD and NlogD data sets, respectively (Tables 1 and 2). Both sets were difficult to predict using the PALLAS PrologD and ACD Labs LogD programs, and they showed similar or even lower performance. The poor performance of these methods indicates difficulties arising from the prediction of the *in-house* collection of compounds that contain structural features not covered by molecules in the training set. It is worth noting that the ALOGPS was developed to predict the 1-octanol−water partition coefficient, log *P*, for neutral compounds but still demonstrated similar or superior performance even in the blind prediction mode.

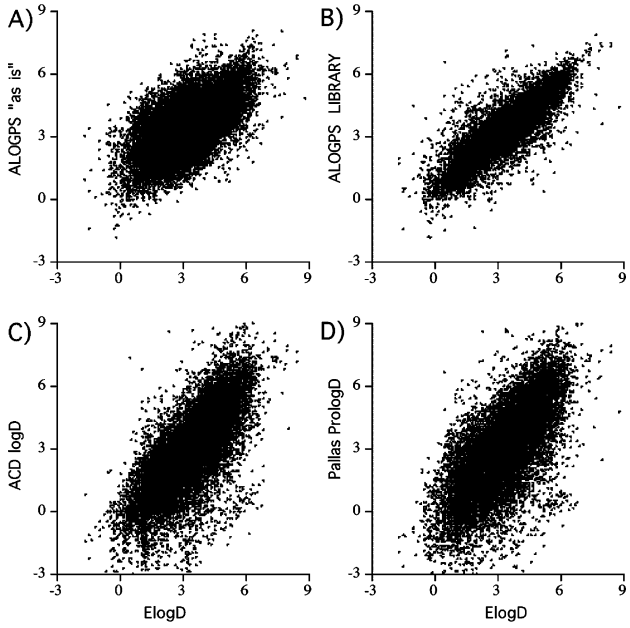

Figure 1. Calculated versus experimental ElogD values for ALOGPS blind prediction (A), ALOGPS LIBRARY mode (B), ACD Labs LogD (C) and Pallas PrologD (D) suites.

The use of LIBRARY mode dramatically improved prediction ability of the ALOGPS and gave rmse of 0.65 and 0.64 for both sets according to the LOO cross-validation test, respectively (Tables 1 and 2). Figure 1 indicates that the distribution of errors for the LIBRARY mode does not contain any bias for particular log *D* values over a range of 9 log units.

When only 50% of compounds were used the LIBRARY calculated a similar performance in the LOO mode and for the blind prediction of the remaining 50% of compounds that were not used in the LIBRARY.

Thus, the LOO results provide an unbiased estimation of the program performance. The use of ElogD data set as the LIBRARY decreased the performance of the program for the NlogD set to rmse of 1.58. This result can be explained by different chemical space covered by compounds from NlogD and ElogD data sets.

The third analysis was done to determine if the performance of the ALOGPS could be further improved by developing a new model for the investigated set of compounds. After the training, the neural network showed better results for the 320 molecules in the test set (Table 2). The rmse and mean average error (MAE) decreased by about 15%. This change of the model performance is significant but not dramatic for most applications. Thus, retraining the ASNN using the same set of descriptors following the original procedure did not radically improve the prediction ability of the model.

On the basis of the performance of ALOGPS, PALLAS PrologD, and ACD Labs LogD suites to blindly predict analyzed data sets, the main factor determining the program performances could be attributed to the poor accuracy of log *P* prediction modules. Indeed, the accuracy of log *D* modules of the same programs was only about 10% better compared to their log *P* predictions for the same compounds (Tables 1 and 2). These results comprised neutral and ionizable compounds. Performance of all three programs for the ElogD subset of compounds that do not have ionizable groups (1920 compounds) gave rmse of 1.19, 1.10, and 0.98 for PALLAS PrologD, ACD Labs LogD, and ALOGPS, respectively. The performance of these programs for the remaining set of compounds produced rmse of 1.43, 1.30, and 1.19, respectively. Both subsets had approximately the same number of heavy atoms, i.e., 29.8 for neutral and 30.6 for ionizable compounds. Thus, there was no bias in the size of molecules. If we assume that errors in the prediction of log *P* and $\Delta(\log P - \log D)$ are independent and equal in magnitude, we would expect an increase of about 40% of rmse for the subset of ionizable compounds compared to the subset of neutral ones. The observed increase in errors of all three models is less than 20%, and thus, the errors in prediction of log *P* values dominate in the model performance.

This observation explains the high prediction ability of the ALOGPS in the LIBRARY mode. In this mode the ASNN identifies nearest neighbors in the chemical space as shown in our previous publications.[19,21,22] The errors for identified nearest neighbors are used to calculate correction term for target compounds, and thus, this improves prediction power of the method. Since log *P* but not log *D* data were used to develop the ALOGPS method, the contribution of calculated indices based on ionizable groups is probably underrepresented and characteristic nearest neighbors are not always found for ionizable compounds. This may decrease the performance of the method. However, since accuracy of log *P* models dominates in the total log *D* error, this factor is not so important for the final accuracy of the model.

The LIBRARY mode provides only local correction. Thus, if the target compound does not have any similar compounds in the LIBRARY, the accuracy of the prediction may not be improved. This was the case for prediction of the NlogD set using ElogD as LIBRARY. Since both NlogD and ElogD sets contained different compounds, there were no improvements in the prediction of the NlogD set using the ElogD as LIBRARY.

The current need of the pharmaceutical industry is not limited to log *P* or log *D* prediction. There is a great need to accurately predict other physical properties such as aqueous solubility and other ADME properties. Known QSAR approaches have limitations to the prediction of these properties for only chemical classes of compounds used in the training set. These sets are usually represented by publicly available databases or published data sets and do not cover a wide range of chemical space. Likewise, in this case, the PHYSPROP database[32] used to train the ALOGPS program is quite different from the ElogD and NlogD data sets. Indeed, the median molecular weight (MW) was 231, 424, and 493 Da for PHYSPROP, ElogD, and NlogD data sets, respectively. The median values of the sum of H-bond donors/acceptors were 4, 5, and 9 for the same three sets. Thus, the molecules in the two test data sets were quite different from those of the PHYSPROP set with respect to these rule of 5 parameters.[1] This difference explains a general trend in medicinal chemistry to synthesize more complex and larger molecules to achieve higher affinity for lead compounds.[33,34] The absence of druglike molecules in the training set can bias developed tools toward simple structures as it was argued elsewhere.[35] However, even the use of a well-balanced set could hardly change this tendency. Indeed, the linear increase of error as a function of the number of nonhydrogen atoms was observed for log *P* [18] and aqueous solubility[36] programs. This tendency appeared to be independent from the nature of the used algorithms (linear or nonlinear) or types of calculated descriptors (atom- or fragment-based or E-state indices). Thus, it is practically impossible to develop the "magic bullet", i.e., a global model that would work for all imaginable chemical classes. As a result, many local models are being developed to predict these properties for proprietary series. The ALOGPS develops such models in a completely automatic fashion. Usually, just a few experimental values are required to generate reliable prediction for the whole series of compounds.[19,25]

Neural network based ALOGPS with its self-learning feature of local correction combines the best properties of both global and local models and in most cases significantly improves the accuracy of prediction for *in-house* compounds.[19,21,22,24] Within this approach a statistical ensemble of neural networks is trained to correlate input parameters with the target property. The ASNN procedure globally maps input parameters (calculated descriptors) to the target property, i.e., does global land-shafting of the property space. The final tuning of the performance of ASNN for compounds is done in the LIBRARY mode. In this step the program uses nearest-neighbors technique to determine local corrections according to the specific features of the analyzed chemical series or/and the property. Since only one or two parameters are used in this step, the ASNN does not overfit even small data sets. Thus, as it was shown, a combination of global and local land-shafting in the LIBRARY mode makes it possible to predict the pH-dependent distribution coefficient, log *D*, although the program was initially developed to predict the

partition coefficient, log *P*, for only neutral compounds. The final correction performs only local tuning of the model and will not generalize to chemical series outside the LIBRARY, as was demonstrated above by the poor prediction of the NlogD data set using ElogD as LIBRARY.

To summarize, we tested several algorithms for prediction of the log *D* distribution coefficient and demonstrated that the neural network based ALOGPS program gives similar or superior results compared to the well-known PALLAS PrologD software and ACD Labs LogD suite in the "as is" mode on two Pfizer proprietary data sets. Performance of ALOGPS in the LIBRARY mode significantly reduced rmse for log *D* prediction to 0.64 and 0.65 (compared to 1.17 and 1.33 in the "as is" mode) for data sets of 17 341 and 640 compounds, respectively. It was shown that the inaccuracy of log *P* predictions was the limiting factor in the performance for all three algorithms. ALOGPS is very fast, works in a completely automated fashion, and does not require any user intervention or extended knowledge in computational chemistry. It can be used to create local models with high prediction abilities for the *in-house* data sets. QSAR approaches such as ALOGPS that can improve prediction ability by self-learning on the user-specific data may find significant applications in the pharmaceutical industry in the near future.

Acknowledgment. This study was partially supported by "Virtual Computational Chemistry Laboratory" INTAS Grant 00-0363. This study would not be possible without the invaluable work of Franco Lombardo and Marina Shalaeva (Pfizer Global Research and Development, Groton Laboratories, CT) who collected experimental data for the ElogD set.

References

(1) 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 Delivery Rev.* **2001**, *46*, 3−26.
(2) Winiwarter, S.; Bonham, N. M.; Ax, F.; Hallberg, A.; Lennernas, H.; Karlen, A. Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach. *J. Med. Chem.* **1998**, *41*, 4939−4949.
(3) Liu, X.; Tu, M.; Kelly, R. S.; Chen, C.; Smith, B. J. Development of a computational approach to predict blood−brain barrier permeability. *Drug Metab. Dispos.* **2004**, *32*, 132−139.
(4) Colmenarejo, G. In silico prediction of drug-binding strengths to human serum albumin. *Med. Res. Rev.* **2003**, *23*, 275−301.
(5) Ishigami, M.; Honda, T.; Takasaki, W.; Ikeda, T.; Komai, T.; Ito, K.; Sugiyama, Y. A comparison of the effects of 3-hydroxy-3-methylglutaryl-coenzyme a (HMG-CoA) reductase inhibitors on the CYP3A4-dependent oxidation of mexazolam in vitro. *Drug Metab. Dispos.* **2001**, *29*, 282−288.
(6) Poulin, P.; Theil, F. P. Prediction of pharmacokinetics prior to in vivo studies. 1. Mechanism-based prediction of volume of distribution. *J. Pharm. Sci.* **2002**, *91*, 129−156.
(7) Lombardo, F.; Obach, R. S.; Shalaeva, M. Y.; Gao, F. Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data set and leave-class-out statistics. *J. Med. Chem.* **2004**, *47*, 1242−1250.
(8) Kerns, E. H. High throughput physicochemical profiling for drug discovery. *J. Pharm. Sci.* **2001**, *90*, 1838−1858.
(9) Tetko, I. V. The WWW as a tool to obtain molecular parameters. *Mini-Rev. Med. Chem.* **2003**, *3*, 809−820.
(10) Mannhold, R.; van de Waterbeemd, H. Substructure and whole molecule approaches for calculating log P. *J. Comput.-Aided. Mol. Des.* **2001**, *15*, 337−354.
(11) Japertas, P.; Didziapetris, R.; Petrauskas, A. Fragmental methods in the analysis of biological activities of diverse compound sets. *Mini-Rev. Med. Chem.* **2003**, *3*, 797−808.
(12) Livingstone, D. J. Theoretical property predictions. *Curr. Top. Med. Chem.* **2003**, *3*, 1171−1192.
(13) 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*, 741−752.
(14) Huuskonen, J. J.; Livingstone, D. J.; Tetko, I. V. Neural network modeling for estimation of partition coefficient based on atom-type electrotopological state indices. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 947−955.
(15) Huuskonen, J. J.; Villa, A. E. P.; Tetko, I. V. Prediction of partition coefficient based on atom-type electrotopological state indices. *J. Pharm. Sci.* **1999**, *88*, 229−233.
(16) Hansch, C.; Leo, A. J. *Subsistent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
(17) Lombardo, F.; Shalaeva, M. Y.; Bissett, B. D.; Chistokhodova, N. ElogD(7.4) 20,000 Compounds Later: Refinements, Observations and Applications. In *LogP2004, The 3rd Lipophilicity Symposium*; ETH: Zurich, Switzerland, 2004; p L-22.
(18) 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. Comput. Sci.* **2001**, *41*, 1407−1421.
(19) Tetko, I. V.; Tanchuk, V. Y. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136−1145.
(20) ALOGPS is available free on-line at http://www.vcclab.org.
(21) Tetko, I. V. Neural network studies. 4. Introduction to associative neural networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717−728.
(22) Tetko, I. V. Associative neural network. *Neural Proc. Lett.* **2002**, *16*, 187−199.
(23) Tetko, I. V.; Tanchuk, V. Y.; Poda, G. I. Evaluation of ALOGPS 2.1 on Pfizer Datasets. In *LogP2004, The 3rd Lipophilicity Symposium*; ETH: Zurich, Switzerland, 2004; p C-30.
(24) Tetko, I. V.; Bruneau, P. Evaluation of ALOGPS 2.1 in AstraZeneca. In *LogP2004, The 3rd Lipophilicity Symposium*; ETH: Zurich, Switzerland, 2004; p C-17.
(25) Tetko, I. V.; Bruneau, P. Application of ALOGPS to predict 1-octanol/water distribution coefficients, log *P* and log *D*, of AstraZeneca in-house database. *J. Pharm. Sci.*, in press.
(26) 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.
(27) Lombardo, F.; Shalaeva, M. Y.; Tupper, K. A.; Gao, F. ElogD(oct): a tool for lipophilicity determination in drug discovery. 2. Basic and neutral compounds. *J. Med. Chem.* **2001**, *44*, 2490−2497.
(28) ACD Labs suite is available at http://www.acdlabs.com.
(29) Tsantili-Kakoulidou, A.; Panderi, I.; Csizmadia, F.; Darvas, F. Prediction of distribution coefficient from structure. 2. Validation of Prolog D, an expert system. *J. Pharm. Sci.* **1997**, *86*, 1173−1179.
(30) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: London, 1999; p 245.
(31) Hall, L. H.; Kier, L. B. Electrotopological state indices for atom types—a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
(32) The Physical Properties Database (PHYSPROP) is a trademark of Syracuse Research Corporation, www.syrres.com.
(33) Oprea, T. I. Current trends in lead discovery: are we looking for the appropriate properties? *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 325−334.
(34) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided. Mol. Des.* **2000**, *14*, 251−264.
(35) Goodnow, R. A.; Gillespie, P.; Bleicher, K. Chemoinformatic tools for library design and the hit-to-lead process: A user's perspective. *Cheminformatics in Drug Discovery*; Wiley-VCH: Weinheim, Germany, 2005.
(36) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.

JM049509L