A Universal Molecular Descriptor System for Prediction of LogP, LogS, LogBB, and Absorption

Hongmao Sun*

Discovery Chemistry, Hoffmann-La Roche Inc., 340 Kingsland Street, Nutley, New Jersey 07110

Received May 16, 2003

Predictive models for octanol/water partition coefficient (logP), aqueous solubility (logS), blood-brain barrier (logBB), and human intestinal absorption (HIA) were built from a universal, generic molecular descriptor system, designed on the basis of atom type classification. The atom type classification tree was trained to optimize the logP predictions. With nine components, the final partial least-squares (PLS) model predicted logP of 10850 compounds in Starlist with a regression coefficient (r^2) of 0.912, cross-validated r^2 (q^2) of 0.892, and root-mean-square error of estimation (RMSEE) of 0.50 log units. The PLS models for solubility (logS), blood-brain barrier (logBB), and a PLS-DA (discrimination analysis) model for HIA were established from the same atom type descriptors. The seven-component PLS model derived from a diverse set of 1478 organic compounds predicted a 21-compound test set designed by Yalkowsky with $r^2 = 0.88$ and RMSEP (RMS error of prediction) = 0.64. A predictive $r^2 = 0.90$ and RMSEE = 0.26 were achieved for logBB of a 57-compound "Abraham data set" with a three-component model. The first three components of a five-component PLS-DA model were sufficient to clearly separate the 169 drug molecules, collected by Abraham, into three classes, according to their percentage human intestinal absorption.

INTRODUCTION

The fact that more compounds fail in clinical trials due to unfavorable ADME properties, than for any other single reason, has attracted increased attention to the field of ADME prediction.[1] Such predictive approaches are needed, because experimental ADME studies are generally low-throughput and cannot match the pace at which compounds are produced in a modern drug discovery effort enhanced by high-throughput synthetic technologies. A reliable ADME predictive model can be useful for identifying and removing compounds with poor ADME properties in the early stages of the process, thus improving the overall quality of the compounds that proceed along the discovery pathway. Numerous procedures for ADME prediction have been reported recently.[2−7] Each procedure addresses one, or a limited set, of the molecular properties that are important for a potential drug candidate. A more generic approach would be beneficial, since it would allow a broad collection of properties to be assessed via the simple implementation of a single tool. ADME-related properties can be grouped into two categories: (1) properties determined by the inherent physicochemical attributes of the compound, such as logP, solubility, and passive diffusion-controlled permeability, and (2) properties related to ligand−receptor interactions, such as plasma protein binding, CYP450 inhibition, and multidrug resistance. It would be difficult to envision a single approach which covers both categories. For the properties in the first category, molecular descriptor based methods have been widely used to derive the predictive models,[8] while for the second category, pharmacophore models or expert systems are needed.[9,10] Here I describe a method that allows prediction of multiple properties from the first category, which was made possible by the development of a simple, yet versatile, molecular descriptor system.

The term "molecular descriptor" refers to any property, measured or calculated, of a molecule, such as molecular weight, melting point, polar surface area (PSA), and so forth. To derive a predictive model for a target property, the typical procedure is to select a collection of molecular descriptors that are believed to be related to that target property and build the model through regression. As expected, for different target properties, the molecular descriptors that are employed are different. In fact, since the developers of the predictive models are likely to bring unique insights to a problem, there are numerous cases where alternate models have been proposed to predict the same property by employing different molecular descriptors.[11] It would be highly beneficial if a generic molecular descriptor system could be found which would allow prediction of most, and possibly all, of the relevant properties of a molecule. This should be possible, since any property of a particular molecule is encoded in its structure. However, it has proven difficult, in practice, to effectively decode the information embedded in the structure. For example, comparing three-dimensional structures, even neglecting the flexibility of molecules, is an extremely challenging task. Reducing the 3D structure to a 2D Markush structure does not help much, and completely simplifying the structure into a 1D formula, while suitable for efficient computation, results in a great loss of structural information. Molecular fingerprinting[12] is a potentially useful approach for a pairwise comparison of molecules and for structural clustering, yet it has not found much use in molecular property predictions. So far, atom type classification has emerged as the most effective way to retain structural information, while reducing the problem to a simple frame-

* Corresponding author phone: (973)562-3870; e-mail: hongmao.sun@roche.com.

10.1021/ci030304f CCC: $27.50
Published on Web 01/23/2004

work. An atom type classification system was presented by Wildman and Crippen for use in predicting logP and molar refractivity (MR) from atomic contributions.[13] In the current study, I address the possibility of using atom types as general molecular descriptors to build predictive models for a wide variety of important properties—namely: logP, logS, logBB, and HIA. For each of these properties, previous attempts have been made to develop prediction tools, and examining these accomplishments is useful for pointing to areas where improvements may be made.

LogP, the logarithm of the partitioning coefficient between *n*-octanol and water, has been widely accepted as a descriptor of the lipophilicity of a molecule. Lipophilicity is one of the most important physicochemical parameters determining both the pharmacokinetic and pharmacodynamic behaviors of a molecule. Many different calculation methods have been published to estimate the logP values from chemical structures.[14−17] Mannhold and Dross[18] divided the methods into three categories: fragmental, atom-based, and conformation-dependent. According to their analysis, fragmental methods produced the best predictive accuracy in a test using 48 drugs with known experimental logP values. Atom-based methods are similar to fragmental methods in the sense of breaking the molecule into smaller groups and summing up the contributions to yield the molecular logP, sometimes including the use of correction factors. The major difference between the two methods is that fragmental contributions are mostly derived from experimental logP values of the selected small molecules, while the atomic contributions are derived from regressions. In the former approach, lack of data can result in missing fragments, as seen in the early versions of ClogP. Another distinguishing factor is that the choice of correction factors makes the most significant difference among the fragment-based methods, while the atom type definition plays the most critical role in differentiating the various atom-based methods. Compared with fragmental methods, one of the limitations of atom-based methods is that the atom types are more or less intercorrelated. To overcome this limitation and to improve the robustness and predictivity of the model, PLS[19] has been applied in the present study as a regression method.

LogS, a measure of aqueous solubility, is another crucial physical property closely related to the ADME profile of a compound. Compounds with poor solubility tend to have poor absorption, low stability, and fast clearance. Prediction methods which included an experimental melting point as a molecular descriptor generally gave more accurate estimations of solubility,[20] partly resulting from the similarity of the processes of dissolving and melting a compound. Both processes involve breaking the intermolecular interactions in solid phase, either through increasing the entropy of the system and Brownian motion of the solvent molecules or by increasing the enthalpy of the system. The obvious limitation of these methods is the dependence on experimentally determined properties, which may be difficult to obtain or not available because compounds have not yet been synthesized. Several methods have been published recently to predict the aqueous solubility without using a melting point. Klopman and co-workers[21] used a group distribution approach, and their best model estimated a test set of 21 commonly used drugs and pesticides designed by Yalkowsky[22] with RMSEP = 1.25 log units. Huuskonen[23] developed a 30−12−1 artificial neural network (ANN) model for estimating aqueous solubilities of a diverse set of 1297 organic compounds, which predicted the same 21-compound test set with $r^2 = 0.91$ and RMSEP = 0.63. Tetko et al.[24] obtained a similar result with fewer variables, a 33−4−1 ANN with 33 electrotopological state (E-state) indices at the input level. Using seven easy-to-compute 1D and 2D descriptors as the input, Liu and So[25] have developed an ANN model to predict the 21-compound test set with $r^2 = 0.79$ and RMSEP = 0.91. In the current study, I further extend this trend of simplification while retaining reasonable predictive success.

The blood-brain barrier (BBB) is a key physiological filter between the systemic circulation and the brain. Its role is to maintain the homeostasis of the central nervous system (CNS). Effective BBB penetration is desirable for drugs targeting CNS diseases, while non-CNS drugs must have a limited BBB penetration to avoid adverse side effects.[11] Due to the experimental difficulties in measuring BBB penetration, the available data sets for logBB are relatively small. Abraham and co-workers[26] combined 30 compounds of Young's histamine H_2 receptor antagonists[27] and their own data set to obtain a 65-compound "Abraham data set". Abraham's model predicted this 65-compound data set with $r^2 = 0.75$ and RMSEE = 0.397, and removal of eight strong outliers yielded a better model with $r^2 = 0.91$ and RMSEE = 0.197. A reasonable model was derived to predict logBB values of 45 drugs (Kelder data set) by using dynamic polar surface area (PSA) as the only descriptor,[28] yet the drawback of this model was its inability to distinguish the difference of hydrocarbon compounds. Clark[29] developed a method for rapid calculation of PSA from SMILES and evaluated the 55 out of 57 compounds in the "Abaraham data set" with $r^2 = 0.79$ and RMSEE = 0.354, using PSA and ClogP as the only descriptors. Both Abraham and Clark's models predicted a 13-compound test set with high accuracy.[29] Interestingly, by adding the number of oxygen and nitrogen atoms in the molecules, a simple model proposed by Norinder and Haeberlein[30] outperformed the sophisticated VolSurf model on the combined Clark and Kelder data sets. The following two rules of thumb given by Norinder and Haeberlein[30] served as a useful guideline to quickly estimate the brain uptake of a compound:

Rule 1: if the sum of N and O atoms (N + O) ≤ 5, then the molecule has a high chance of entering the brain;

Rule 2: if logP > (N + O), then logBB is positive.

In the present work, I seek a more quantitative approach, while retaining a simplified rule system.

Human intestinal absorption (HIA) is a critical parameter for developing orally available drugs. Since most drugs are absorbed via passive diffusion through the cell membrane, it would seem feasible to predict the intestinal absorption from the molecular structure. Lipinski's "Rule of Five" is the most widely applied qualitative model for a quick estimation of the absorption of a compound.[7] Quantitatively, Palm et al. observed an excellent sigmoidal relationship between the absorbed fraction and dynamic polar surface area (PSA) for 20 structurally diverse drugs.[31] Kelder and co-workers found that most of the orally administrated drugs had a PSA of less than 150 Å.[28] A larger data set was collected by Abraham and co-workers with the human intestinal absorption of 241 drugs, among which 169 drugs were considered as having reliable data.[32] The predictive

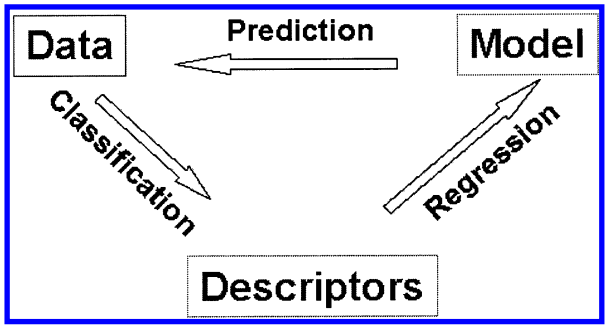

Figure 1. The cycle of model optimization.

model, based on five Abraham descriptors, estimated the absorption of 169 drugs with a r^2 of 0.74. In the current study, I seek an approach of comparable simplicity that yields a better prediction.

METHODS

Atom Type Classification. An atom type is related to its own chemical properties and its environment. A particular atom type is primarily determined by the following factors: its element, its aromaticity, its neighboring atoms, and whether it is in a ring. In the case of a conjugated system, the property of an atom can be affected by another atom or functional group several bonds away. On the basis of the experience and chemical intuition, the primary classification tree was constructed. The classification tree was implemented by using Daylight Toolkits.[33] To determine where to split and where to stop splitting the tree more accurately, logP was employed to train the tree. The reasons of choosing logP lie in the following considerations: first, logP can be measured accurately with an experimental error of less than 0.2 log units in most cases; second, partitioning coefficients span over 16 orders of magnitude, and logP values range from −5 to over 11; third, a high quality data set, Starlist, with 10 974 compounds is available through Daylight; fourth, the compounds in the data set are structurally diverse, with chemical entities ranging from simple methane to complex cyclic peptides. Compounds containing elements other than H, C, N, O, P, S, and halides were not included in the final data set of 10 850 compounds. Figure 1 illustrates the cycle for the training of the classification tree. The atom types identified from the original classification tree served as molecular descriptors to generate the predictive model for logP. Analysis of the errors in the prediction yielded clues for further modification of the classification tree. By running several cycles of atom type classification, regression, and prediction, a final model was achieved.

Analysis of the structures of the outliers provided information for refinement of the classification tree. A basic assumption guiding this study is that outliers are largely caused by misclassification of atom types, instead of experimental errors. Therefore, no data point was removed from the original data set throughout this study. In the initial model, it was found that compounds containing a positively charged aromatic nitrogen atom were all strong outliers. Separation of this atom type improved the model substantially. Analysis of the value of the "variable importance in projection" (VIP),[34] which estimates the influence of each atom type on the model, was another source for improving the classification. A low VIP value could be due to two possibilities: the particular atom type itself has little affect on the target property, or the atom type is poorly defined. A poorly defined atom type might correlate with the target property positively in some compounds and negatively in other compounds. In such a case, further splitting the atom type might lead to a better model. The poorly defined atom types could also be conformed by analyzing the standard errors of the coefficients with cross-validation, where a large standard error indicates that the same atom type contributes differently in different cross-validation models.

As mentioned in a previous section, the disadvantage of using atom types as molecular descriptors is that the variables are often intercorrelated, making it inappropriate to apply multiple linear regression. For example, the carbon atom and the nitrogen atom in a nitrile group (−C≡N) are highly correlated, since the atom type of a carbon triply bonded to a nitrogen tends to occur simultaneously with a nitrogen triply bonded to a carbon in the same molecule. To overcome this limitation, partial least squares (PLS) was chosen as the regression method for this study.

Partial Least Squares. PLS, also known as "projections to latent structures", is a powerful multivariate analytical tool.[34] PLS is a regression extension of principal component analysis (PCA), combining the strengths of PCA and of ordinary least squares regression. Projection is the procedure to extract the related information and reduce the dimensionality. With multivariate techniques, the original variables are summarized by calculation of new variables, called latent variables. The latent variables, which are linear combinations of the original variables, are orthogonal to each other. Furthermore, the latent variables are interpretable and rich in practical information. The residuals-based diagnostics and VIP analysis supply useful information to instruct the atom type classification. PLS analysis was performed by using SIMCA-P10.[35]

Both X and Y variables were mean-centered and scaled to unit variance (UV) before the PLS analysis. Preprocessing the variables with UV-scaling is the most objective way to scale the data, giving all variables an equal opportunity to influence the data analysis.[34] The scaling weight employed in this study was $1/S_k$, where S_k represented the standard deviation of variable k.

The goodness of fit of a PLS model was assessed by a regression coefficient r^2. The goodness of prediction was evaluated by a cross-validated r^2, designated as q^2. Cross-validation was carried out by keeping part of the data out of model development, then having it predicted by the model and compared with the actual values. Every data point was left out once and only once. The q^2 value is the main criterion for assessing the quality of a model. In general, a model with a q^2 of 0.3 or higher is statistically meaningful, while a q^2 greater than 0.5 is regarded as a good model and 0.9 or above is excellent.[34] VIP estimates the influence of every original variable on the matrix Y. Variables with larger VIPs are the most relevant for explaining Y, and those with VIPs less than 0.5 are of the least importance.

PLS-discriminant analysis (PLS-DA) was the extension of the PLS method. The objective of PLS-DA is to develop a model with the capability of separating "tight" classes of observations on the basis of their X variables. The Y matrix encodes class membership by a set of "dummy" variables,

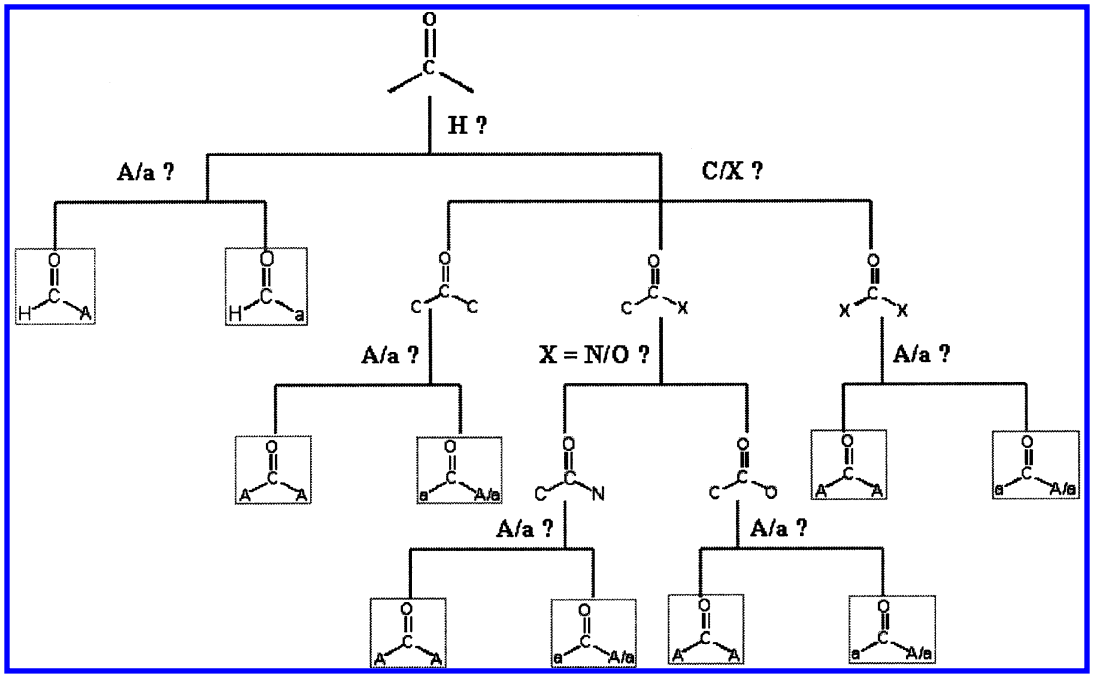

Figure 2. A subtree of an aliphatic carbonyl carbon. **H**: hydrogen, **A**: aliphatic, **a**: aromatic, **C**: carbon, and **X**: a non-carbon atom.

and a PLS model is derived by finding the discriminant plane to effectively separate data into different classes.

Prediction of logS, logBB, and Absorption. The models to predict aqueous solubility, blood-brain barrier, and HIA were built from the same atom types and correction factors identified from the classification tree. PLS regression was utilized to derive the models for logS and logBB with 7-fold cross-validations. PLS-DA was applied to establish a qualitative model for HIA.

For logS, the experimental aqueous solubility values of 1487 organic compounds were obtained by combining 1297 compounds from the AQUASOL database of the University of Arizona and SCR's PHYSPROP Database,[23] with 211 drug molecules extracted by Huuskonen et al.,[36] and then removing any duplicated compounds. The solubility values were expressed as logarithms of molar solubility (mol/L) and varied from 1.58 to −11.62 with a mean value of −2.76. The test set was 21 drug molecules and environmentally interesting compounds collected by Yalkowsky.[22]

For logBB, the "Abraham data set" of 57 compounds was selected as the training set. The test set contained the 13 compounds used by Clark[29] and Liu et al.[11]

For HIA, the 169 drug molecules collected by Abraham and co-workers[32] were used and classified as "good", "medium", and "poor", according to the percentage intestinal absorption. HIA was determined from three methods: bioavailability (for the cases of highly bioavailable drugs), percentage of urinary excretion of drug-related material following oral administration, and the ratio of cumulative urinary excretion of drug-related material following oral and intravenous administration.[32]

RESULTS AND DISCUSSIONS

Atom Type Classification. The advantage of using the atom types as generic molecular descriptors is to avoid preselecting the properties presumably related to the target property, thus reducing the opportunity of introducing bias at the beginning of model establishment. The emphasis of atom type classification is placed on maximizing the difference among the molecules by using as few atom types as possible. The contributions of different atom types to different target properties are automatically computed by PLS analysis.

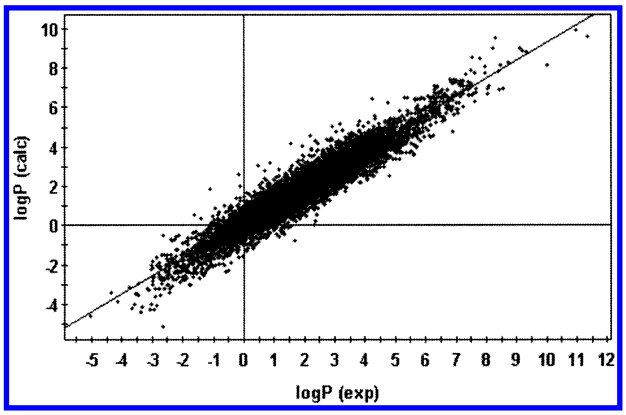

Figure 3. Relationship between experimental and calculated logP for 10 850 compounds in Starlist (9-component model, $r^2 = 0.912$, $q^2 = 0.894$, RMSEE = 0.509).

In total, 218 atom types were identified, including 88 types of C, 7 types of H, 55 types of N, 31 types of O, 8 types of halides, 23 types of S, and 6 types of P, together with 26 correction factors, as listed in Table 1. All of the correction factors are easy to calculate. Figure 2 illustrates a subtree of an aliphatic carbonyl carbon not in a ring, where the carbonyl carbon was further divided into 10 atom types, aliphatic and aromatic aldehyde, ketone, amide, ester, and urea or carbamate.

LogP Prediction. The final model used nine components and was able to predict logP of 10 850 compounds in the Starlist with a r^2 of 0.912, q^2 of 0.894, and RMSEE of 0.509 (Figure 3). Cross-validation was carried out by a leave-group-out procedure, and the data set was randomly split into seven even groups. The model is better than both the AlogP program as implemented in *Cerius2*[37] which predicted the

Table 1. Atom Types and Correction Factors[a]

var ID	atom type and correction factor	contribution for logP	logS	logBB	HIA	var ID	atom type and correction factor	contribution for logP	logS	logBB	HIA
C1	xc(H)x	−0.3235	−0.3136	−0.1185	−0.0060	C76	XC[X,x]=C	0.2927	−0.3307	−0.0886	0.1258
C2	xc(H)c	−0.3113	0.4226	−0.0816	0.0138	C77	CC(X)=C	0.2768	−0.2771		−0.0276
C3	cc(H)c	0.0864	−0.0591	0.0051	0.0055	C78	CC(X)=C	0.5261	−0.5020		0.0274
C4	ac(−a)a	0.1968	−0.5872	−0.0673	0.0402	C79	C=S	−0.2069	−0.0283		0.0550
C5	xc(=O)x	0.0249	−0.2328		−0.0067	C80	CC(C)=N	0.1276	−0.1912		0.1214
C6	xc(=O)c	−0.5577	0.5733	−0.2441	−0.0192	C81	XC(A)=N	−0.3921	−0.2717	−0.0510	−0.0333
C7	cc(=O)c	−0.7760	−0.0245			C82	A=C=A	1.2096	−0.8383		
C8	xc(X)x	0.0817	−0.5190	−0.0820	−0.0128	C83	cC#N	−0.1378	0.0417		
C9	xc(X)c	−0.0740	−0.2411	0.1030	0.0055	C84	A=C−C#N	0.1632	0.0675		
C10	cc(X)c	0.1767	−0.1353	−0.0137	−0.0022	C85	C−C#N	−0.6519	0.3039		−0.0326
C11	cc(C)x	−0.0529	0.6032	−0.0913	−0.0067	C86	X−C#N	−0.2619	0.2856	−0.1753	−0.1108
C12	cc(C)c	0.1588	−0.2343	0.0225	−0.0020	C87	HC#C	−0.2678	0.0199		0.0550
C13	cc(c)c	0.2735	−0.4369	−0.0598	0.0010	C88	AC#C	0.1210	−0.2182		−0.0023
C14	ac(a)a	0.1298	0.0058	−0.0071	−0.0132	H1	HC, C with all single bonds	0.0305	−0.0168	0.0070	0.0002
C15	CCH2C;R	0.2014	−0.0811	0.0471	0.0030						
C16	XCH2X;R	−0.2494	0.1523		−0.1092	H2	Hc	0.0681	−0.0463	−0.0001	0.0059
C17	CCH2X;R	−0.1387	0.2960	0.0086	−0.0022	H3	HC, C with nonsingle bonds	0.0233	−0.0508	−0.0260	0.0168
C18	C(C)CHC;R	0.1623	−0.0882	0.3464	0.0077						
C19	C(c)CHC;R	0.3642	−0.5496		0.0233	H4	HO[C,S,P](=O)	−0.2059	0.2223		−0.0175
C20	C(C)CHX;R	0.0405	−0.0502	−0.0064	−0.0036	H5	HO[A,a]	−0.3084	0.2686	−0.1350	−0.0132
C21	C(c)CHX;R	0.0233	−0.0290	−0.0604	0.0245	H6	Hn	−0.0805	−0.5368	−0.0959	0.0033
C22	CC(C)(C)C;R	−0.0165	−0.0600		−0.0037	H7	HN	−0.2155	0.1053	−0.0352	−0.0079
C23	CC(c)(C)C;R	−0.3415	0.1558		0.0447	N1	[n+]	−3.5460	1.2303		
C24	CC(A)(A)A;R	0.1185	−0.2644		−0.0028	N2	nC(=O)	−0.2324	0.0794	−0.2441	−0.0004
C25	CC(a)(A)A;R	0.0158	−0.2758		0.0415	N3	nn	−0.2584	0.5507	−0.1820	0.0306
C26	CC(=O)C;R	−0.2861	0.0954		0.0460	N4	an(A)a	−0.1034	0.1003	−0.3398	−0.0248
C27	XC(=O)X;R	0.6306	−0.2421			N5	cnc	−0.1309	0.0493	−0.0724	−0.0050
C28	CC(=O)X;R	−0.5903	0.0726	0.1092	−0.0773	N6	anc	−0.0588	−0.1970	−0.0737	0.0152
C29	AC(H)=C;R	0.1316	−0.0796		0.0260	N7	HNC(=O);R	−0.1949	0.0691		0.0421
C30	CC(C)=C;R	0.1207	−0.0650		0.0063	N8	ANC(=O);R	0.2091	−0.3011		
C31	AC(X)=C;R	0.0816	−0.4375		−0.2069	N9	aNC(=O);R	−0.7147			0.0194
C32	AC(=C)A;R	−0.0173	−0.4211		0.0267	N10	AN(A)C(=O);R	−0.0530	−0.1069	−0.0295	−0.0564
C33	AC(H)=N;R	−1.7565	−0.0848			N11	NS(=O);R	−0.3903	−0.0554		−0.1092
C34	AC(A)=N;R	−0.0789	0.3864			N12	NN;R	−0.6385	−0.5738		
C35	AC(A)=N;R	−0.3294	0.8159			N13	AN(=O)A;R	−1.0553			
C36	AC(=N)A;R	−0.3342	0.0403	−0.0064	0.0271	N14	C=AN(A)A;R	0.1657	−0.2052	0.4432	−0.0052
C37	AC=[P,S]A;R	0.0220	−0.5404			N15	X=AN(A)A;R	0.0592	−0.9367		−0.0265
C38	CH4,CCH3	0.2250	−0.0962	0.0430	0.0087	N16	AN(A)A;R	−0.0607	0.3813	0.0270	−0.0425
C39	cCH3	0.2104	−0.1066	−0.0124	0.0081	N17	ANHA=A;R	−0.2409	0.3930		−0.0275
C40	OCH3	−0.1796	0.1696	−0.1105	0.0087	N18	CNHC;R	−0.4892	0.5721	−0.0064	−0.0405
C41	NCH3	−0.1990	0.3827	0.0164	−0.0045	N20	AN=A;R	−0.2981	0.5889	−0.0064	0.0077
C42	[S,P]CH3	−0.1535	0.2041		0.0022	N21	H2NC(=O)	−0.4098	0.2408	−0.1209	−0.0057
C43	CCH2C	0.3072	−0.1130	0.0794	0.0090	N22	CNHC(=O)	−0.2710	0.0725	−0.3023	−0.0191
C44	[C,c]CH2c	0.2213	−0.1721	−0.1810	−0.0253	N23	cNHC(=O)	−0.2272	0.0135	−0.4290	0.0103
C45	XCH2x	−0.0054	0.2255		−0.0969	N24	[X,x]NHC(=O)	−0.5266	0.5008		0.0974
C46	CCH2X	−0.1654	0.1684	−0.0448	−0.0077	N25	CN(C)C(=O)	−0.3126	−0.0789		0.0508
C47	cCH2X	−0.1833	0.3481	−0.0134	−0.0381	N26	CN(c)C(=O)	−0.5649	0.1091		−0.0075
C48	XCH(X)X,XCH(C)X	−0.0040	0.1123	0.0172	−0.0161	N27	XN([A,a]C(=O)	−0.2560	0.0815		−0.0787
C49	XCH(C)C	−0.0010	0.1281	−0.1811	−0.0055	N28	H2NS(=O)	−0.5149	0.0966		−0.0826
C50	XCH(c)C	−0.0070	0.5477		−0.0717	N29	NHS(=O)	−0.3634	−0.0507		−0.0541
C51	CCH(C)C	0.2907	−0.2428	0.2090	0.0114	N30	NS(=O)	−0.0850	−0.8520		0.0236
C52	CCH(C)C	0.2283	−0.3993		0.0476	N31	N=N=N	−0.3757	−0.3093		
C53	XC(X)(X)X	0.6763	0.1983			N32	N=N	0.4346	−0.4598		0.1214
C54	CC(X)(X)X	0.1027	−0.3992	0.0076	0.0299	N33	NN	−0.1726	0.4420	−0.0443	0.0938
C55	CC(x)(X)X	0.1098				N34	aN(=O)=O	−0.1299	−0.1199	0.0663	0.0312
C56	CC(X)(X)C	−0.0634	−0.2870		0.2066	N35	ON(=O)=O	0.4780	0.2308		
C57	CC(C)(X)C	−0.1881	0.1947	−0.0737	−0.0325	N36	AN(=O)=O	−0.9062	−0.1542	−0.3334	0.1258
C58	CC(C)(C)C	0.3715	−0.4685	0.3286	0.0284	N37	D4N	−3.7649	−0.6684		
C59	aC(=O)H	−0.6683	0.8312			N38	A[=,#]CN(C)C	−0.0846	−0.4538		0.0125
C60	AC(=O)H	−0.5941	0.7069		−0.0807	N39	CN(C)C	0.0907	0.5230	0.1178	0.0135
C61	CC(=O)C	−0.4311		−0.2336	0.0411	N40	[C,c]N(c)[C,c]	0.4215	−0.0169	0.0627	0.0819
C62	cC(=O)[C,c]	−0.4898	0.3291		0.0566	N41	XN(A)A	0.2170			
C63	XC(=O)X	0.1157	0.5884	−0.1209	0.0033	N42	A[=,#]CNHC	−0.0193	0.1572	−0.1160	−0.0085
C64	xC(=O)(X,x)	0.1538	−0.0780			N44	CNHC	−0.4459	1.0798		−0.0081
C65	AC(=O)N	−0.3779	0.2720	−0.4283	−0.0089	N45	cNHC	0.2782	0.1456	0.0199	−0.0140
C66	aC(=O)N	−0.3057	0.0539	−0.1299	0.0260	N46	C=NC	−0.1852	0.7697		
C67	AC(=O)O	−0.1659	0.2528		0.0063	N47	C=Nc	−0.0583	−0.6181	−0.1753	0.0225
C68	aC(=O)O	0.0503	0.0222		0.0382	N48	A=NX	−0.0988	0.5175	0.0136	−0.0066
C69	CH2=C	0.1246	−0.0832	−0.0426	0.0620	N49	aC#N	−0.1754	0.1571		
C70	cCH=C	0.3276	−0.5656		0.0325	N50	AC#N	−0.1883	0.0417		
C71	CCH=C	0.2356	−0.1765		0.0302	N51	HN=	−0.0980	0.2481	−0.1753	−0.0593
C72	[X,x]CH=C	0.1191	0.0116	−0.0749	0.1140	N52	cNH2	−0.3016	−0.9371		−0.0363
C73	CC(C)=C	0.3303	−0.3064		−0.0016	N53	A=CNH2	−0.1608	0.3656	−0.3016	−0.0015
C74	cC(c)=C	−0.0848	−0.0057		−0.0415	N54	aA=CNH2	−0.3429		0.0068	−0.0153
C75	CC(c)=C	0.1464	−0.5095		−0.0404	N55	CNH2	−0.6088	0.0187		

Table 1. (Continued)

var ID	atom type and correction factor	contribution for logP	logS	logBB	HIA	var ID	atom type and correction factor	contribution for logP	logS	logBB	HIA
N56	XNH2	−0.3237		−0.5938	−0.0128	S2	csc	0.6981	−0.7253	0.0324	0.0363
N58	XNHA	−0.0906	0.1762			S3	xsa	0.4905	−0.5788		−0.0919
O1	AOC(=O)−;R	−0.7670	0.6197		−0.0795	S4	AS(=O)A;R	−0.4043	0.1289		0.0176
O2	AO[S,P](=O)−;R	−0.3197				S5	aSa;R	0.3486	−0.6230		
O3	COC;R	−0.2019	0.2583	−0.1209	0.0024	S6	ASA;R	0.4344	−0.3005		0.0788
O4	cOc;R	0.1181	−0.0140	−0.0231	−0.0670	S7	ASa;R	0.5195			
O5	XOA;R	0.0936	−0.1223	−0.0737	−0.0748	S8	ASH	0.1796	−0.4235		−0.0372
O6	AO[C,S,P](=O);R	0.0282	0.0365		0.0298	S9	aSH	−0.1924	−1.2367		−0.2359
O7	AO[C,S,P](=O);R	−0.2884	0.2756		−0.0222	S10	CSC	0.4474	−0.1936	−0.0922	−0.0637
O8	cOH	−0.1699	0.5052		−0.0877	S11	XSX	0.2723			
O9	COH	−0.3466	0.2012	−0.1350	−0.0097	S12	CSX	0.3753	−0.7733		0.0550
O10	[X,x]OH	−0.1736	0.1470		−0.0787	S13	AS(=O)A	−1.1744	1.3012		
O11	AO[C,S,P](=O)	−0.0377	0.2446		0.0076	S14	aS(=O)A	−1.0116	1.6380		0.0392
O12	aO[C,S,P](=O)	−0.3907	0.2164		0.0522	S15	cS(=O)(=O)c	−0.5584	0.4288		
O13	COC	−0.1292	0.3758	−0.0383	−0.0123	S16	CS(=O)(=O)C	−0.9799	0.3754		
O14	XOX	−0.1171				S17	cS(=O)(=O)C	−0.7697			0.1107
O15	COX	−0.0397	−0.0353			S18	X=S	0.5656	0.1516		−0.0297
O16	aOa	−0.1177	0.2868		0.1146	S19	NS(=O)(=O)	−0.2205	−0.6110		−0.2032
O17	COc	0.0129	0.2990	−0.0147	0.0191	S20	XS(=O)(=O)A	−0.3783	0.4244		
O18	AOa	−0.4180	−0.2398			S21	c=S	−0.3409	−0.2974		0.0550
O19	c(=O)	−0.3010	0.1208	−0.2441	−0.0084	S23	CC(=S)X	0.1915	−0.2162		
O20	x(=O)	−0.9118	0.1980		0.1356	S24	XC(=S)X	−0.0094	−0.1853		
O21	CC(=O)C	−0.4445	0.2951	−0.2336	0.0141	M1	(int) (molecular wt/100)	0.1604	−0.3433	−0.0301	−0.0051
O22	cC(=O)c	0.1566	−0.1726		0.0319	M2	intramolecular HB	0.1543	−0.0643		−0.0241
O23	CC(=O)C	−0.1841	−0.0746		0.0682	M3	adjacent halides	0.0732	−0.0631	0.0057	0.0121
O24	XC(=O)X	0.0517	−0.0970	−0.1209	−0.0033	M4	(C15 > 4) ? (C15−4)/2:0	−0.0128	−0.2240		−0.0035
O25	[X,x]C(=O)x	0.1538				M5	(C43 > 5) ? (C43−5)/2:0	−0.0787	−0.0223		
O26	CC(=O)X	−0.2871	0.1937	−0.2531	−0.0032	M6	frac of rotatable bonds	0.3333	−0.3645	0.1951	0.0183
O27	[X,x]C(=O)c	−0.1429	0.0543	−0.1299	0.0105	M7	alpha amino acid	−0.5496	−0.5471		
O28	[S,P]=O	−0.2040	0.0895		−0.0096	M8	amide	0.0123	−0.1398		−0.0592
O29	a−N(=O)=O	−0.0580	−0.0600	0.0331	0.0156	M9	salic acid	0.1171	−0.1318	−0.0200	0.0034
O30	A−N(=O)=O	−0.0963	0.0036	−0.1667	0.0629	M10	1,4-dioxane	0.9914	−0.5657		0.1179
O31	A(A)N(=O)A	−0.5977	0.2112			M11	acetyl urea	−0.1695	0.0710		0.0044
F1	aF	0.0752	0.0468		0.0392	M12	no. of aromatic rings	0.6343	−0.2623	0.0859	
F2	AF	0.1207	0.0179	−0.0004	0.0108	M13	no. of aliphatic rings	0.2158	−0.2808	−0.0242	0.0087
F3	aCl	0.5651	−0.3674	−0.0176	0.0086	M14	mult oxygen atoms (>8)	−0.0210	−0.0428	0.0601	−0.0003
F4	ACl	0.3547	−0.3303	0.0384	0.0273	M15	zwitterions	−0.2573	−0.9014		−0.0145
F5	aBr	0.5903	−0.6772	−0.4242	0.0296	M16	mult hydroxyl groups	−1.6210	−0.1986		
F6	ABr	0.4909	−0.5048	0.0139	0.1009	M17	mult acids	−0.1405	−0.0372	−0.0737	−0.0601
F7	aI	0.5618	−0.0360			M18	linear zwitterion	0.0946	−0.6628		−0.1236
F8	AI	0.9442	−1.1070			M19	ring no. = 0	−2.1109	1.5172		
P1	P;R	−0.6467				M20	no. of fused bonds (fb)	0.1603	−0.1967	−0.0514	0.0209
P2	D3P	0.1324				M21	no. of fused atoms (fa)	−0.0404	−0.1131	0.0019	0.0054
P3	AAP(=O)A	−0.7751	1.3597		−0.0909	M22	2 * fb − fa	−0.0158	0.1312		0.0076
P4	aAP(=O)A	−0.2931	−2.6575		0.0738	M23	pyrazine	0.0869	−0.0705	−0.0369	−0.0523
P5	aP(=O)	−0.8205				M24	ortho functional groups	−0.0882	−0.1187	0.0387	0.0128
P6	P=[S,N,C]	0.5041	−0.1853			M25	meta functional groups	0.0755	−0.0677	−0.0064	−0.0090
S1	D3s	0.0026				M26	para functional groups	0.0399	−0.1256		-0.0016

[a] Annotation: **A**: aliphatic, **a**: aromatic, **C**: aliphatic carbon, **c**: aromatic carbon, **X**: non-carbon aliphatic, **x**: non-carbon aromatic. −: single bond, =: double bond, #: triple bond.

same data set with a r^2 of 0.672, and the AlogP program in MOE,[38] which gave a r^2 of 0.302. The latest version of ClogP[14] produced the best prediction of the same data set, with a r^2 of 0.949 and RMSEE of 0.404. The values of r^2 and RMSEE, evaluating the fitting quality of a model, are not the only parameters to determine the quality of a model—q^2 and the number of variables (NOV) are equally important in assessing the robustness and predictivity of the model. ClogP used more than 150 variables,[16] compared to nine independent variables (components) used in this study. More components, in general, will improve the fitting quality of a model, by increasing the complexity of the model, but this decreases the predictive quality of the model. At a certain point, q^2 will start to drop as the complexity of the model increases, indicating that the model begins to simulate the noise. Since no measured property is free of experimental errors, overfitting the noise will lower the predictive power of the model.

LogP is the partition coefficient of octanol and water, where logP of 10 means the fraction of the compound in aqueous phase is 10 billion fold less than that in the organic phase. It is extremely difficult to accurately measure this trace amount of a compound in water, and as a result the higher logP values tend to be more noisy. On the other hand, the further a logP value is away from the mean value, the more influence the data point has on the r^2, according to the definition of r^2. Consequently, there exists an upper limit for r^2, depending on how many high logP data points were included in developing the model, since experimental noise is not predictable. Another source of experimental error is the low solubility of the compounds with high logP values. Besides, no matter how complex a model is, it cannot cover

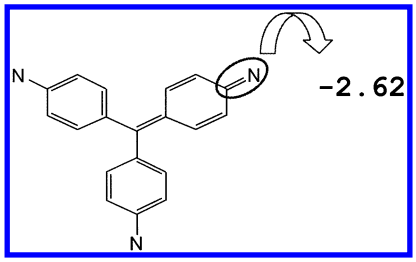

Figure 4. The outlier structure of the logP prediction. The double-bonded nitrogen has an arbitrary fragmental logP of −2.62 in the ClogP method.

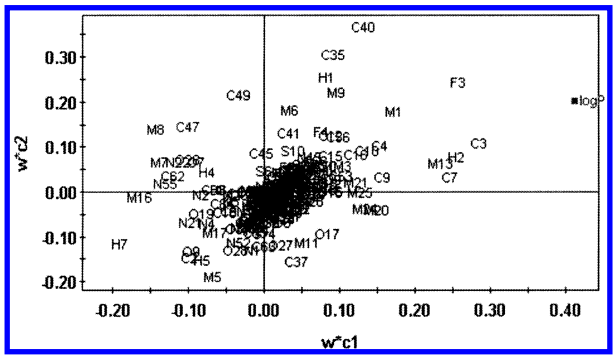

Figure 5. PLS weight plot of w*c1/w*c2 in the nine-component logP model.

all of the structural information affecting the target property. Thus, an exact model can only be obtained for a carefully selected small data set.

The structures of the 100 most poorly predicted compounds of this study, and those of the ClogP analysis, were compared, and about half of them were overlapped. The worst prediction from this study was the compound shown in Figure 4, where the experimental logP of the compound was −0.21, while the predicted value was 2.68. On the contrary, ClogP produced a perfect prediction for the compound, −2.06, by arbitrarily defining the contribution of the amine double-bonded to an aromatic ring, which occurred only once in the whole data set. The fragmental method has the freedom of deriving the contribution of a fragment without experimental value from the logP value of the whole molecule. In PLS regression, the influence of each original variable on the model is not independent, since the influence is determined by the importance of the variable in the principal components and the relative importance of each component.

A predictive model is more powerful if it can provide guidance regarding the structural modifications that will improve a property. PLS regression is able to disclose the relative importance of each original variable. In the logP model, the first two components accounted for, respectively, 64.0% and 14.7% of the overall correlation. Figure 5 plots the weights of both X and Y variables of the first two components, illustrating the correlation structure between X and Y. Aromatic carbons, total number of aromatic rings, and molecular weight are among the factors that are positively correlated to logP; alkylamines, the alkyl hydroxy group, and flexibility of the molecule are the factors negatively correlated to logP. Introducing the atom types positively correlated to logP will increase the molecular logP, and introducing negatively correlated atom types should decrease logP.

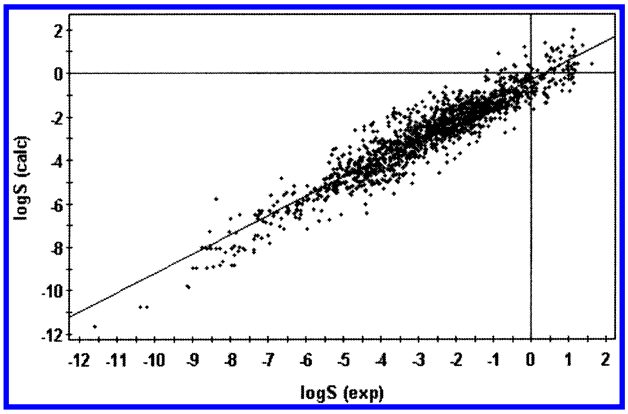

Figure 6. Relationship between experimental and calculated logS for the 1478-compound training set (7-component, $r^2 = 0.890$, $q^2 = 0.807$, RMSEE = 0.654).

Table 2. Observed and Predicted Aqueous Solubilities of 21-Compound Test Set

no.	name of compound	cal logS	exp logS	resid
1	2,2′,4,5,5′-PCB	−7.320	−7.890	−0.570
2	benzocaine	−1.990	−2.320	−0.330
3	acetylsalicylic_acid	−1.695	−1.720	−0.025
4	theophylline	−1.194	−1.390	−0.196
5	antipyrine	−0.991	−0.560	0.431
6	atrazine	−4.118	−3.850	0.268
7	phenobarbital	−2.463	−2.340	0.123
8	diuron	−3.288	−3.800	−0.512
9	nitrofurantoin	−3.007	−3.470	−0.463
10	phenytoin	−3.792	−3.990	−0.198
11	diazepam	−3.398	−3.760	−0.362
12	testosterone	−4.399	−4.090	0.309
13	lindane	−3.838	−4.600	−0.762
14	parathion	−4.467	−4.660	−0.193
15	diazinon	−4.387	−3.640	0.747
16	phenolphthalein	−3.836	−2.900	0.936
17	malathion	−2.326	−3.370	−1.044
18	chlorpyriphos	−6.157	−5.490	0.667
19	prostaglandin_E2	−4.315	−2.470	1.845
20	p,p′-DDT	−7.235	−7.150	0.085
21	chlordane	−7.180	−6.860	0.320

LogS Prediction. Continuing with the use of the same atom types as molecular descriptors, a logS model was derived on the basis of a data set of 1478 structurally diverse compounds. The seven-component PLS-model estimated the logS of the training set with a r^2 of 0.890, q^2 of 0.807, and RMSEE of 0.654 (Figure 6). Without considering the effects of polymorphism, self-assembly of solute, or symmetric packing, the seven-variable linear model gave a standard error comparable to experimental measurements. Many factors affect the process of dissolving a compound in aqueous solution, which makes aqueous solubility difficult to be measured precisely. It was reported that a standard error of 0.6 log units should be expected for experimental logS.[39] Although there is still room for further improvement, the seven-variable linear model was good enough to illustrate that the same atom types and correction factors optimized for logP prediction were also nearly optimal for logS prediction.

The general applicability for the prediction tool was tested by using the test set of 21 commonly used compounds of pharmaceutical and environmental interest, designed by Yalkowsky.[22] The results are presented in Table 2 and Figure 7. The present model gave a RMSEP of 0.639, better than the models developed by Klopman and Kühne and compa-

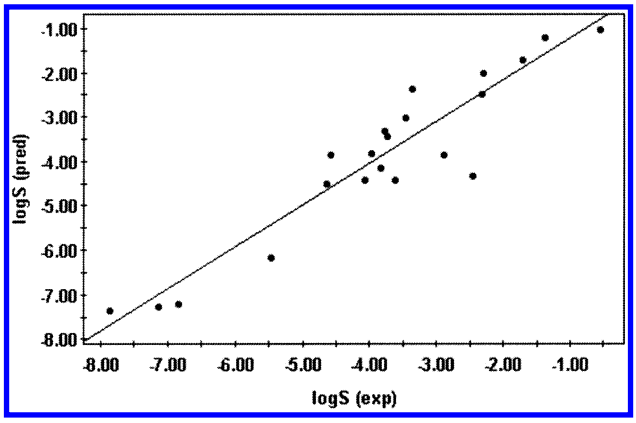

Figure 7. Correlation of predicted and experimental logS of the 21-compound test set.

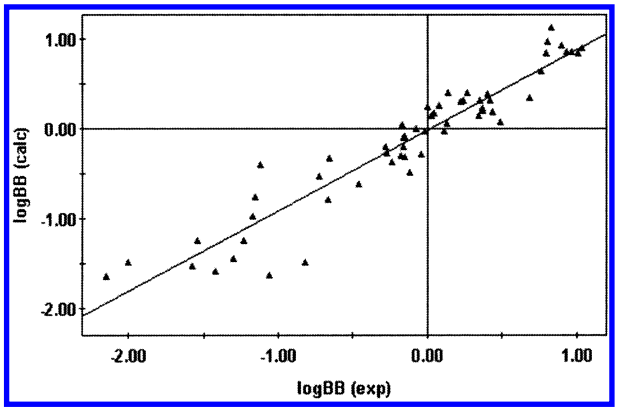

Figure 8. Relationship between experimental and calculated logBB for the 57-compound training set (3-component, $r^2 = 0.897$, $q^2 = 0.504$, RMSEE = 0.259).

rable to Huuskonen's 30−12−1 ANN model and Tetko's 33−4−1 ANN model. However, this model was much simpler with much fewer variables. One obvious outlier was prostaglandin E2, excluding the outlier reduced RMSEP to 0.509.

LogBB Prediction. A three-component model was built from the same atom type descriptors, and it estimated the data set of 57 compounds with a r^2 of 0.897, q^2 of 0.504, and RMSEE of 0.259 (Figure 8). The relatively lower q^2 resulted from the small size of the data set. Totally, 94 different atom types were identified for the 57 compounds, and half of these atom types occurred only once or twice through the whole data set. When the compounds containing these atom types were left out in cross-validation, the contribution of these atom types could not be predicted accurately, since they did not appear in the training set. After changing the cross-validation scheme from seven rounds to leave-one-out, the q^2 value increased to 0.552.

When the model built from the "Abraham data set" was used to predict the 13-compound test set, the RMSEP was 0.671, worse than Clark's or Abraham's models. The major reason was that the 13 compounds in the test set were not well covered by the chemical space defined by the training set. The 57-compound "Abraham data set" contained 22 compounds from the "Young data set" and 35 of their own compounds. The "Young data set" is mostly composed of three structural analogues, cimetidines, guanidinothiazoles, and ranitidines,[27] while the other compounds were largely small organic compounds. So, the 57-compound "Abraham data set" is not structurally diverse. Abraham's descriptors and PSA used by Clark were less sensitive to the subtle structural variation, thus their models performed very well in this case. After swapping 4 pairs of compounds in the training (compounds 16, 20, 26, and 28, as numbered in ref 11) and test data sets (compounds Y-G20, SKF 101468, 32, and 35) to improve the chemical space coverage of the training set, r^2 and RMSEE of the new model improved to 0.910 and 0.502, and RMSEP of the new test set of 13 compounds was significantly reduced to 0.326. By using the method of this study, a better predictive model should be derived, once a larger data set of high quality becomes available.

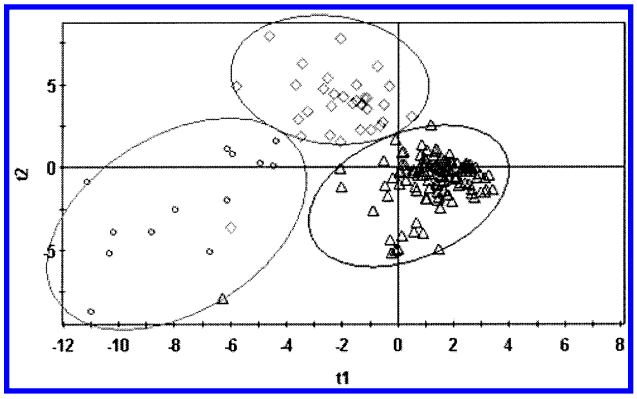

Figure 9. PLS−DA t1/t2 score plot, showing the separation of 3 classes of HIA, good (△), medium (◇), and poor (○).

HIA Prediction. The HIA values of 169 drugs were expressed as the percentage of the drug absorbed by the intestine. It is desirable if the percentage absorption can be predicted precisely, but in many cases, a qualitative classification might be sufficient. The 169 drugs in the data set were classified into three classes: class 1: absorption > 80%, class 2: 80% ≥ absorption > 20%, and class 3: absorption ≤ 20%. The five-component PLS-DA model separated the compounds with a r^2 of 0.921 and q^2 of 0.787.

The first three components carried 42.3%, 26.6%, and 14.1% of total variance, respectively. In the 2D scoring plot, two drugs, digoxin and cymarin, were misclassified (Figure 9). Digoxin (Figure 10), with measured HIA of 81% and misclassified as poorly absorbed, has 3 violations against the rule-of-five—its molecular weight is 780, and it has 14 hydrogen bond acceptors and 6 hydrogen bond donors. The molecular weight of cymarin, another drug misclassified as poorly absorbed, is also above 500. In the 3D scoring plot (Figure 11) all 169 drugs were well separated. Abraham's model misclassified 29 drugs out of 169. In the case of a virtual screen, only poorly absorbed compounds would need to be identified and removed. A three-component PLS-DA model separated the compounds of less than 20% absorption, with $r^2 = 0.939$ and $q^2 = 0.861$ (Figure 12).

CONCLUSIONS

A universal molecular descriptor system has been established from a trained atom classification tree. Atom types and correction factors are rapidly computed from SMILES representations of molecules, at a rate of about 10 000 structures per minute on an SGI Octane. PLS models have been built on the basis of a single molecular descriptor system to accurately predict logP, logS, logBB, and HIA.

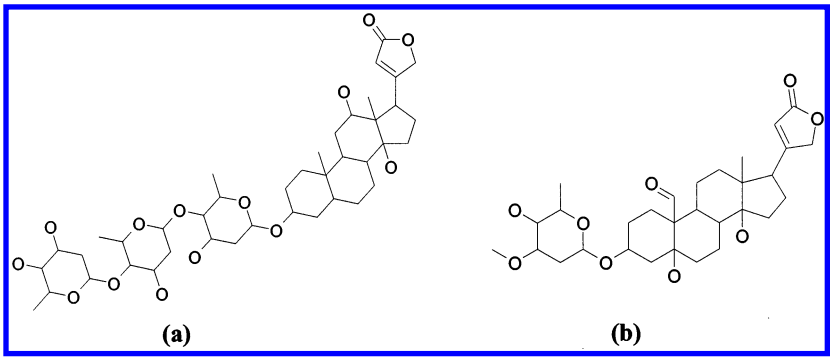

Figure 10. Structures of digoxin (a) and cymarin (b).

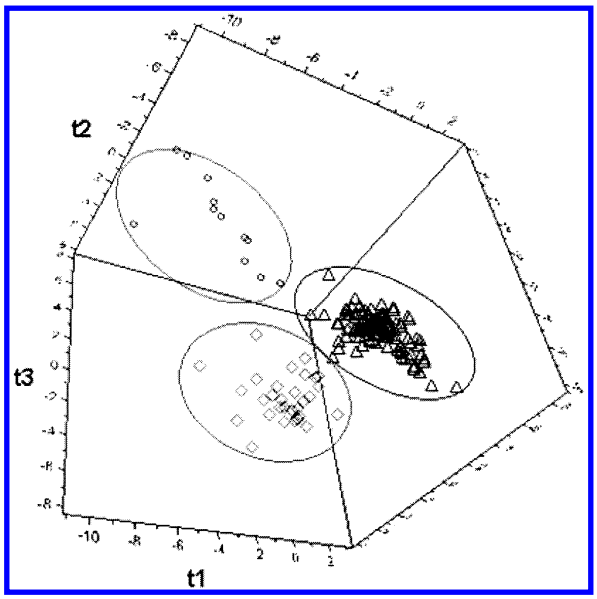

Figure 11. PLS−DA t1/t2/t3 score plot. The three classes of percentage HIA (good: △, medium: ◇, poor: ○) are clearly discriminated.

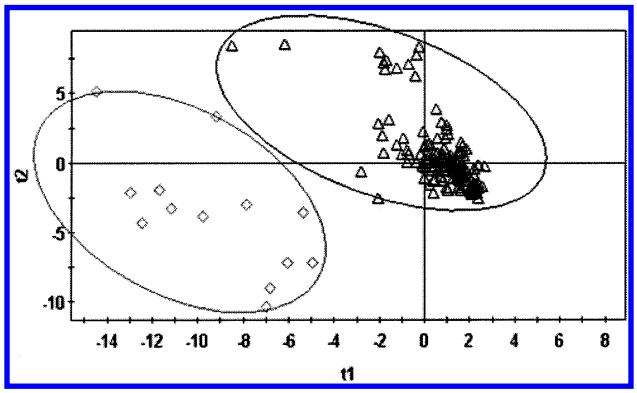

Figure 12. PLS−DA t1/t2 score plot, showing the clustering of 2 classes of HIA (good: △, HIA% ≥ 20%; poor: ◇, HIA% < 20%).

By analyzing the atomic contributions to different target properties, specific atom types affecting these properties can be identified. This molecular descriptor system can also be applied to generate QSAR models, offering a tool to optimize potency and ADME properties simultaneously.

ACKNOWLEDGMENT

Drs. Sung-Sau So and David Fry are particularly acknowledged for critical reading of the manuscript. Drs. Jarmo Huuskonen and Michael H. Abraham are thanked for supplying the data sets of solubility, logBB and HIA.

REFERENCES AND NOTES

(1) Butina, D.; Segall, M. D.; Frankcombe, K. Predicting ADME properties in silico: methods and models. *Drug Discov. Today* **2002**, *7*, S83−S88.
(2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeablity in drug discovery. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3−25.
(3) Sinko, P. J. Drug selection in early development: screening for acceptable pharmaco-kinetic properties using combined in vitro and computational approaches. *Curr. Opin. Drug Discov. Dev.* **1999**, *2*, 42−48.
(4) Ekins, S.; Rose, J. In silico ADME/Tox: the state of the art. *J. Mol. Graph. Model*. **2002**, *20*, 305−309.
(5) Selick, H. E., et al. The emerging importance of predictive ADME simulation in drug discovery. *Drug Discov. Today* **2002**, *7*, 109−116.
(6) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.
(7) 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. Rev.* **2001**, *46*, 3−26.
(8) Clark, D. E.; Grootenhuis, P. D. J. Progress in computational methods for the prediction of ADMET properties. *Curr. Opin. Drug Discov. Dev.* **2002**, *5*, 382−390.
(9) Ekins, S.; de Groot, M. J.; Jones, J. P. Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome p450 active sites. *Drug Metab. Dispos*. **2001**, *7*, 936−944.
(10) Sanderson, D. M.; Earnshaw, C. G. Computer prediction of possible toxic action from chemical structure: The DEREK System. **1991**, *10*, 261−273.
(11) Liu, R.; Sun, H.; So, S. S. Development of quantitative structure−property relationship models for early ADME evaluation in drug discovery. 2. Blood-brain barrier penetration. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1623−1632.
(12) Wild, D. J.; Blankley, C. J. Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 155−162.
(13) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(14) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, Y.; Matsushita, Y. Simple method of calculating octanol/water partition coefficient. *Chem. Pharm. Bull.* **1994**, *42*, 976−978.
(15) CLOGP, Daylight Chemical Information Systems, Santa Fe, NM, http://www.daylight.com.
(16) Petrauskas, A. A.; Kolovanov, E. A. ACD/LogP method description. *Persp. Drug Discuss. Design* **2000**, *19*, 99−116.
(17) Meylan, W. M.; Howard, P. H. Atom/fragment contribution method for estimating octanol/water partition coefficients. *J. Pharm. Sci*. **1995**, *84*, 83−92.
(18) Mannhold, R.; Dross, K. Calculation procedures for molecular lipophilicity: a comparative study. *Quant. Struct.-Act. Relat*. **1996**, *15*, 403−409.
(19) Wold, S.; Johansson, E.; Cocchi, M. PSL − partial least-squares projections to latent structures. *3D QSAR Drug Des*. **1993**, 523−550.
(20) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *J. Pharm. Sci*. **2001**, *90*, 234−252.
(21) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of aqueous solubility of organic molecules by the group contribution approach.

Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.
(22) Yalkowski, S. H.; Banerjee, S. In *Aqueous solubility, method for estimation for organic compounds*; Marcel Dekker: New York, 1992; pp 128−148.
(23) Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 773−777.
(24) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1488−1493.
(25) Liu, R.; So, S. S. Development of quantitative structure−property relationship models for early ADME evaluation in drug discovery. 1. Aqueous solubility. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1633−1639.
(26) Abraham, M. H.; Chadha, H. S.; Mitchell, R. C. Hydrogen bonding. 33. Factors that influence the distribution of solutes between blood and brain. *J. Pharm. Sci*. **1994**, *83*, 1257−1268.
(27) Young, R. C., et al. Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists. *J. Med. Chem*. **1988**, *31*, 656−671.
(28) Kelder, J., et al. Polar surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* **1999**, *16*, 1514−1519.
(29) Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *J. Pharm. Sci*. **1999**, *88*, 815−821.
(30) Norinder, U.; Haeberlein, M. Computational appraoches to the prediction of the blood-brain distribution. *Adv. Drug Deliv. Rev*. **2002**, *54*, 291−313.
(31) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm. Res*. **1997**, *14*, 568−571.
(32) Zhao, Y. H., et al. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure−activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci*. **2001**, *90*, 749−784.
(33) Daylight toolkits, Daylight Chemical Information Systems, Santa Fe, NM. http://www.daylight.com.
(34) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis Principles and Applications*; Umetrics Academy: Kinnelon, NJ, 2001.
(35) SIMCA-P 10.0, Umetric Inc., Kinnelon, NJ. http://www.umetrics.com.
(36) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.
(37) Cerius 2, Accelrys, Inc. San Diego, CA. http://www.accelrys.com/cerius2.
(38) MOE, Chemical Computing Group, Inc. Montreal, Canada. http://www.chemcomp.com.
(39) Jorgenson, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev.* **2002**, *54*, 355−366.

CI030304F