Native Atom Types for Knowledge-Based Potentials: Application to Binding Energy Prediction

Brian N. Dominy and Eugene I. Shakhnovich*

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138

Received March 8, 2004

Knowledge-based potentials have been found useful in a variety of biophysical studies of macromolecules. Recently, it has also been shown in self-consistent studies that it is possible to extract quantities consistent with pair potentials from model structural databases. In this study, we attempt to extend the results obtained from these self-consistent studies toward the extraction of realistic pair potentials from the Protein Data Bank (PDB). The new method utilizes a clustering approach to define atom types within the PDB consistent with the optimal effective pairwise potential. The method has been integrated into the SMoG drug design package, resulting in an improved approach for the rapid and accurate estimation of binding affinities from structural information. Using this approach, it is possible to generate simple knowledge-based potentials that correlate ($R = 0.61$) with experimental binding affinities in a database of 118 diverse complexes. Furthermore, predictions performed on a random 1/3 of the database consistently show an average unsigned error of 1.5 log K_i units. It is also possible to generate specialized knowledge-based potentials, targeted to specific protein families. This approach is capable of generating potentials that correlate strongly with experimental binding affinities within these families ($R = 0.8-0.9$). Predictions on 1/3 of these family databases yield average unsigned errors ranging from 1.1 to 1.3 log K_i units. In summary, we describe a physically motivated approach to optimizing knowledge-based potentials for binding energy prediction that can be integrated into a variety of stages within a lead discovery protocol.

Introduction

Knowledge-based potentials are used prodigiously in macromolecular biophysical research. Perhaps their most striking feature is the speed by which the energetics of large systems may be estimated. This partially explains their current use in studying large protein systems but also their applicability to problems in the fields of structure-based drug design and proteomics where screening of extensive structural databases and configurational states may be necessary.

Although their speed is an important feature, tradeoffs between speed and accuracy yield potentials that are not generally applicable to a broad range of systems. Typical approximations include the assumption of pairwise additive potentials as well as coarse-grained distance dependence. These approaches arise naturally as a means of maintaining a reasonable level of precision when determining energy functions from a distribution of atom−atom or residue−residue contacts. While typical, these techniques are certainly not universal and various alternatives have been proposed in an attempt to improve the accuracy of knowledge-based potentials for various applications. The introduction of density estimation to smooth radial distribution functions was found to improve the knowledge-based potential's ability to handle distance dependence.[1] In another study, multibody interactions were treated approximately by manipulating residue−residue interactions based on the secondary structural element with which each residue was associated.[2] Furthermore, knowledge-based potentials have been augmented with mean field solvation energy terms in order to improve their accuracy in predicting protein/small molecule binding affinities.[3] The aim of these and other studies is to create a scoring function that has the speed inherent in knowledge-based potentials but is also accurate enough to provide some practical information.

The field of structure-based drug design is a well-established research area, and a number of methods for scoring the efficacy of putative inhibitors have been developed. Currently, statistical approaches such as quantitative structure−activity relationships (QSAR) and CoMFA analyses play a significant role in computational lead discovery processes.[4−6] These models provide a primary advantage of speed, while also maintaining a reliable level of accuracy in domains strictly limited by their training set. Empirical energy functions, similar to QSAR functionals, fit weights within a linear combination of classical energy terms in order to reproduce binding free energies.[7] Linear interaction energy (LIE) methods are based on the approximation that by evaluating ensemble averages of weighted energy terms in the bound and unbound states one can estimate the binding free energy.[8] Finally, free energy perturbation (FEP) approaches rely on classical potentials and extensive ensemble averages of the bound and unbound states as well as a number of intermediate states in order to rigorously calculate the binding free energy difference typically between closely related ligands in a common active site.[9] Moving from QSAR to LIE and finally to FEP methods, these ap-

* To whom correspondence should be addressed. Tel: 617-495-8733. Fax: 617-384-9228. E-mail: eugene@belok.harvard.edu.

10.1021/jm0498046 CCC: $27.50
Published on Web 08/03/2004

proaches become more grounded in physical principles but simultaneously more computationally intensive.

Knowledge-based potentials, which had been used extensively for protein studies, have recently entered the field of rational drug design as atom-based potentials were developed. Since the introduction of atomic knowledge-based potentials, a variety of methods have emerged to apply this technique to the problem of binding energy prediction.[10−14] The knowledge-based approach is advantageous in its physical derivation as well as the speed of the corresponding energy calculations. The speed has obvious benefits in the context of drug screening and lead discovery protocols where a large number of putative compounds require evaluation. The physical basis of these potentials may also have benefits to the drug design community. Although purely statistical methods such as QSAR and CoMFA are currently very popular in lead discovery protocols, physical methods have the potential for generating robust screening functions that can improve the diversity of leads.

Physically based potentials may play a significant role in future drug design protocols. The significant amounts of data retrieved from combinatorial chemistry techniques, as well as genomics studies, have recently propelled statistical approaches in drug screening to new heights of popularity. However, when a completely novel target is encountered, these techniques can become bogged down. The reason may be a general feature of most statistical techniques, namely, that novel targets require the statistical approach to extrapolate to identify properties of the target. The same argument can be applied not only to novel targets but also to novel inhibitors that bind in modes distinct from those observed to date. The discovery of novel binding modes can have significant implications not only for biological challenges such as the evolution of drug resistance but also legal issues such as avoiding patent violations. As these issues become more relevant to scientific and business interests, physically based techniques can begin to play a more important role in the drug design process. These approaches, when fundamentally based on physical laws, have the ability to move beyond the solution space best treated by statistical approaches.

The physical basis of knowledge-based potentials has been challenged recently in the literature and remains an active area of discussion among theoretical chemists. Valid arguments against the simple physical interpretation of knowledge-based potentials have been made from first principles as well as through simulation.[15,16] Recent theoretical studies, however, have demonstrated the hope for extracting simple pairwise potentials from protein structural databases. These self-consistent model studies have demonstrated that it is possible to accurately extract simple pair potentials from contact statistics in Protein Data Bank (PDB)-like databases.[17,18]

Similar results were also observed in the context of atomic resolution knowledge-based potentials using small molecule growth (SMoG).[19] Results demonstrated that by designing ligands to bind receptors using a predetermined pair potential, a quantity strongly correlated with the true pair potential ($r > 0.8$) could be extracted from contact statistics gathered from these structures.

While quantities strongly correlated with simple pair potentials have been extracted from the self-consistent structural databases described previously, when these approaches are implemented in real systems, the results are less than perfect. Errors in the extracted potentials could result from the same reason that imperfect correlation is observed in the self-consistent studies; however, given the strong correlations observed in these studies, entropy may be a more likely source of the observed errors. The correlation between the binding free energy and the binding potential is only perfect when entropy can be neglected. When evaluating a broad database containing distinct targets and small molecules, entropy may play a key role in defining the relative binding affinity. However, when evaluating homologous targets separately, one would expect that entropic considerations would substantially cancel out. Even in this scenario, the correlation between the predicted binding enthalpy and the experimental binding affinity is weak in some families. We must then turn to the possibility that although the procedure for extracting interaction potentials has been optimized within the self-consistent study, the procedure is not perfectly transferable to the PDB. The potentials themselves are not being extracted with high accuracy.

In this work, we suggest a means for optimizing the potential extraction procedure for real databases such as the PDB. While the demonstrated ability to extract potentials from contact statistics in self-consistent systems is encouraging, an additional step is required to make this approach consistent with real chemical systems. A hidden assumption built into each of the self-consistent studies is that there exists a one-to-one relationship between each pair of particle types and the true interaction potential between these types. This means that for each pair of particle types, there exists one and only one effective interaction between these particles. This assumption is key to the extraction procedure's ability to identify unique pair potentials between these predefined particle types. In real chemical systems, this one-to-one relationship is not built in.

If we are to expect to extract the "true" potential from structural databases, we must have particle types consistent with this potential. In this work, atom types are generated by clustering atomic parameters describing classical interaction energy terms such as partial charge and the Born radius. By grouping atoms according to these parameters, we hope to identify atom types that are consistent with the true effective pair potential. Using a procedure optimized in self-consistent studies to extract potentials highly correlated to the true potential, we introduce native atom types to extract truer potentials from the PDB. The knowledge-based potential generated from these atom types is evaluated by its ability to predict relative binding affinity in a database of protein/ligand crystal structures.

Results

The objective of this work is to identify native atom types or atom types consistent with the true effective pair potential. The first step in generating atom types native to the true potential is to characterize atoms based on properties intrinsic to the true potential. Because we obviously do not know the true

potential *a priori*, we choose properties characteristic of interaction energies from empirical classical potentials such as CHARMM or AMBER. Atoms are then clustered according to these properties to define atom types. These atom types are then used to train a traditional knowledge-based potential based on the approach of Ishchenko et al.[11]

At this stage, atoms are described using two properties. While a number of possible property pairs have been evaluated, the most significant properties appear to be the partial charge and the Born radius. Atom types based on these properties were empirically found to yield knowledge-based potentials capable of producing the highest correlation to experimental binding affinities.

Atoms from a training database are associated with a two-dimensional vector containing the Born radius and partial charge mediated by a weight, which influences the relative contributions of these properties to the atom's description. The protein and ligand atoms are clustered separately into 1−15 clusters. The optimal weights and the optimal number of protein and ligand atom types are not known *a priori*. Therefore, all knowledge-based potentials resulting from the different weights and the different number of atom types are determined.

In this section, we will describe the results obtained from optimizing atom types within the SMoG knowledge-based potential. First, we describe the potentials obtained by choosing atom types that maximize correlation to experimental binding affinities within a diverse database of 118 complexes. We show that by optimizing atom types based solely on a single partial charge descriptor, we are able to significantly improve the correlation between predicted and experimental binding affinities ($R = 0.60$) relative to SMoG 2001 ($R = 0.43$). Furthermore, we show that optimization of atom types based on a combination of the partial charge and the Born radius is also able to demonstrate significant improvement over past results ($R = 0.63$). A more rigorous jackknifing test, using 1/3 or 39 of the 118 complexes, demonstrates the robustness of the optimized knowledge-based potentials. An additional jackknifing test, in which an entire target family is removed from the training set, is performed to examine the practical utility of this approach for examining novel targets.

Next, we examine the binding energy correlations within eight protein target families that comprise the diverse testing set of 118 complexes. We observe that while atom types based on partial charge are capable of demonstrating improved binding energy predictions relative to SMoG 2001 within the diverse database, these atom types are not capable of generally improving the intrafamily correlations. On the contrary, however, atom types based on a combination of partial charge and the Born radius are capable of significantly improving the binding energy correlations observed both within the diverse data set and within the eight protein families.

Finally, we optimize atom types to target families individually, constructing family specific knowledge-based potentials for the purpose of accurate binding energy prediction. A different balance of the Born radius and partial charge properties is optimal in describing the atom types within the different families. Energies obtained from optimized knowledge-based potentials are strongly correlated with experimental binding affinities ($R = 0.8-0.9$). Furthermore, jackknifing results demonstrate that the predictive ability of these family specific potentials is excellent, providing accuracy approaching 1 log K_i unit.

Diverse Database of 118 Complexes. The first set of calculations performed involved optimizing the atom types in order to generate knowledge-based potentials that correlated strongly with the known binding affinities from a diverse set of protein−ligand complexes.

1. Atom Types Based Only on Partial Charge. Atom types based on clustering partial charges would be expected to resemble atom types typically encountered in empirical force fields such as CHARMM or AMBER. Because Gasteiger partial charges are generated through a partial equilibration of initial formal charges, these descriptors retain information relevant to the local bonded connectivity. It is this local inductive polarization effect that is the basis for atom typing in many other potentials and is also the basis for the "chemical intuition" that was used to identify atom types in the previous versions of SMoG.[11,12,19] While some similarity in atom type definitions is expected, this similarity also depends on the number of resulting atom types. As we will describe later, the number of optimal atom type clusters does not resemble the number of atom types described in many other potentials. Regardless of the expected similarity between the partial charge-based atom types and the SMoG 2001 atom types, an improvement was observed in the binding energy prediction.

Using the diverse database, SMoG 2001 demonstrated a correlation of $R = 0.43$ with the experimental binding affinity.[11] This correlation was increased to $R = 0.60$ by optimizing the partial charge-based atom types. In other words, atoms were grouped based on their partial charge and the number of these clusters (or atom types) was chosen based on the optimal correlation of the resulting knowledge-based potential to known binding affinity data. These atom types and the corresponding potential are shown in Table 1. A scatter plot is given showing the predicted vs experimental binding energies using SMoG 2001 and the optimized KB potential (Figure 1). In this test, the optimal knowledge-based potential is chosen based on its ability to correlate strongly with experimental binding energies within the testing set. For this reason, the testing set is actually used to train this potential and therefore does not represent an independent validation. Instead, this analysis offers an indication of whether information extracted from this database can improve our knowledge-based potentials. A more rigorous jackknifing test using independent testing and training databases is described later.

2. Atom Types Based on Partial Charge and the Born Radius. The improvement in the correlation observed using partial charge atom types is also possible using a specific combination of partial charge and Born radius. In Figure 2a, for each increment of the weight from 0 to 1, the optimal correlation associated with some number of protein and ligand atom types is shown. This

Table 1. Optimized Potential and Atom Type Clusters for Diverse Complexes[a]

0.00-3.50 Potential

	L0	L1	L2
P0	0.17	−0.47	0.09
P1	−0.34	−0.27	0.1
P2	0.62	−0.23	0.32

3.50-4.50 Potential

	L0	L1	L2
P0	−0.09	−0.03	0.06
P1	−0.12	0.29	0
P2	0.16	−0.15	−0.12

Three Protein Types

	partial charge		Born radius	
	mean	SD	mean	SD
P0	−0.293	0.041	N/A	N/A
P1	−0.536	0.020	N/A	N/A
P2	0.054	0.122	N/A	N/A

Three Ligand Types

	partial charge		Born radius	
	mean	SD	mean	SD
L0	0.240	0.077	N/A	N/A
L1	−0.335	0.100	N/A	N/A
L2	−0.001	0.067	N/A	N/A

[a] Atom types for the protein and ligand are named P0, P1, and P2 and L0, L1, and L2, respectively. The weighting parameter scaling the properties is $w = 1$ (only partial charge is considered).

describes the maximum correlation possible given atom types described by a combination of partial charge and the Born radius. While the high correlation described previously using simple partial charge-based atom types is observed, another region of significant correlation (average $R = 0.62$) is observed using atom types based on a combination of partial charge and the Born radius. The most frequently observed atom types described by these weighted properties yields a potential with a correlation of $R = 0.63$ to experimental binding affinities. These atom types and the corresponding potential are shown in Table 2. A scatter plot showing the experimental binding affinities vs those obtained from this optimized potential is shown in Figure 2b,c.

These results demonstrate that using these parameter-based atom types, it is possible to extract information pertinent to improving correlations between predicted potential energies and experimental binding affinities in a diverse collection of protein/ligand complexes. Also shown in Figure 2a is a line representing the SMoG 2001 potential using unoptimized atom types. It is clear that the potentials generated from optimized atom types have the possibility of yielding improved agreement with experimental binding energies relative to KB potentials where the atom types are chosen using less systematic means.

3. Jackknifing Tests.

3.1. Leave-1/3-Out Tests. Multiple jackknifing tests were run on this data set to more rigorously examine the predictive ability of the optimized potentials. The jackknife tests were performed by randomly extracting one-third of the database and measuring the ability of the optimal potential determined on the remaining two-thirds of the database to predict binding affinities of the portion left out. In the diverse database, we can see that

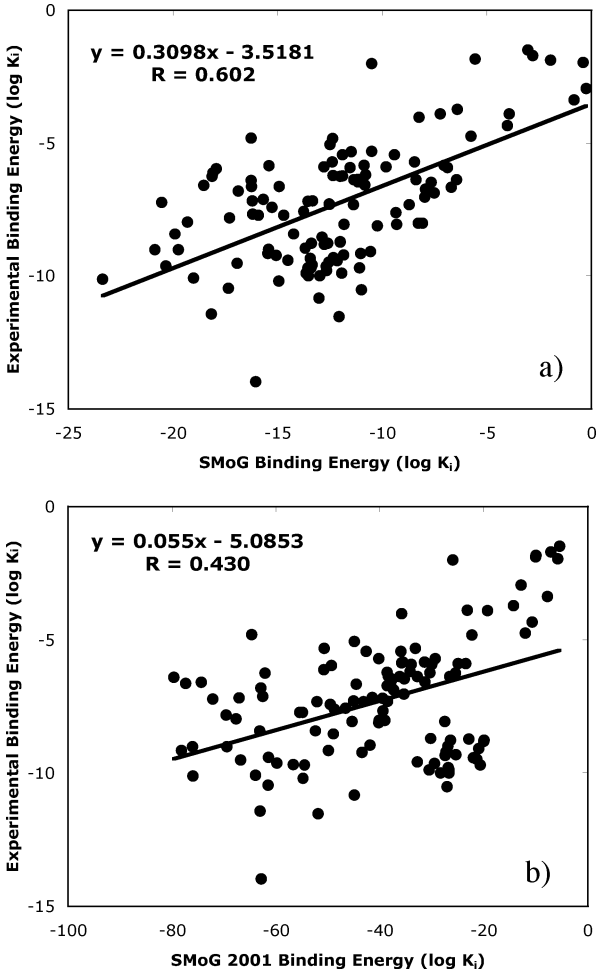


Figure 1. Correlation between experimental binding affinity (log K_i) and predicted energy from (a) SMoG using partial charge-based atom types and (b) SMoG 2001.

the predictive error follows a qualitative pattern closely related to the correlation shown previously (Figure 3a). It is expected if the "left-out" portion of the training set is related to the remaining training set, that a weaker correlation between the KB energy and the experimental binding energy will indicate a weaker predictive ability. As a result, low predictive errors are found using atom types based solely on partial charge, as well as using atom types based on a combination of partial charge and the Born radius.

To draw an independent conclusion, the predictive error is assessed at the weight "w" corresponding to the minimum error of the complexes "left-in" ($w = 0.54$). At this minimum in the error, we see an average predictive error of 1.5 log K_i units or 2.0 kcal/mol at room temperature based on partial charge/Born radius atom types. Using the potential optimized for maximum correlation to the 118 complex data set, the average unsigned error is also 1.5 log K_i units. This compares favorably with SMoG 2001 where the predictive error on this testing set is 1.7 log K_i units or 2.3 kcal/mol at room temperature.

Also shown is a scatter plot demonstrating the experimental vs predicted binding affinity (in log K_i units) for 20 independent "leave-1/3-out" tests on the 118 complexes training set corresponding to the minimum absolute predictive error (Figure 3b). Points within the plot represent the average predicted energy for each

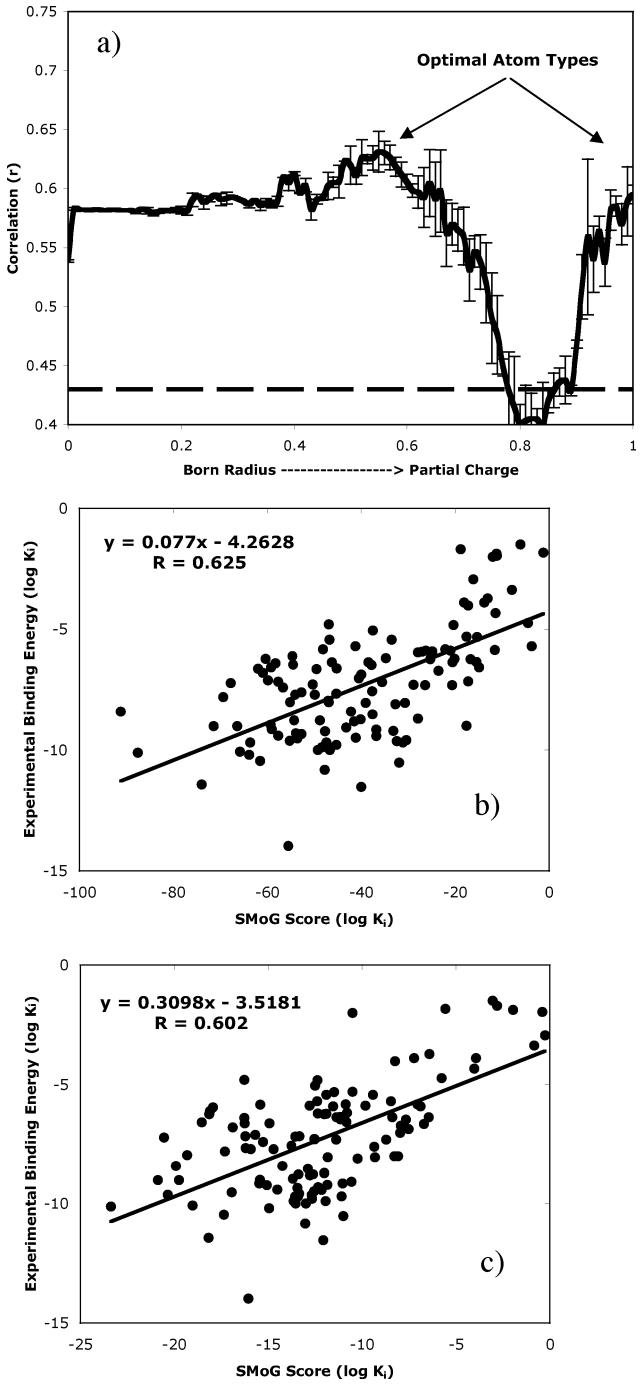


Figure 2. Atoms are described by a combination of the partial charge and the Born radius. The relative contribution of these two descriptors, used to cluster atoms into atom types, is scaled between 0 and 1. Plotted in panel a is the maximum possible correlation using clusters (atom types) based on a particular combination of these parameters. Combinations of descriptors that yield high correlation to the experimental binding affinity in 118 diverse complexes are indicated as optimal atom types. The scatter plots are shown for the two optimal atom types at weights of (b) $w = 0.55$ and (c) $w = 1.0$.

of the 118 complexes generated through multiple jackknife optimizations. The error bars indicate the standard deviation (SD) of the predicted values over the multiple jackknife optimizations. The data show that weaker binders are clearly segregated from stronger binders, while stronger binders are more difficult to distinguish from one another. It also demonstrates that the predicted energies will tend to overestimate the binding energy of very weak binders, while no significant bias appears in tighter binders.

Table 2. Optimized Potential and Atom Type Clusters for Diverse Complexes[a]

0.0−3.50 Potential

	L0	L1	L2	L3	L4	L5	L6
P0	0.97	0.08	−0.36	1.44	−0.62	−1.02	−0.1
P1	0.64	1.3	0.73	−0.54	−0.11	0.79	−1.19
P2	0.08	0.08	0.39	1.03	−0.39	0.62	−0.44
P3	0.11	0.77	0.9	0.34	−0.22	−0.05	−0.96
P4	−0.33	−0.29	−0.07	−0.61	0.06	0.72	−0.56
P5	−0.52	−0.54	−0.26	−0.54	−0.03	0.42	−0.47

3.50−4.50 Potential

	L0	L1	L2	L3	L4	L5	L6
P0	0.25	−0.29	−0.4	−0.94	−0.17	−0.4	0.26
P1	−0.17	0.7	1.45	−1.06	0.32	0.54	−0.96
P2	−0.46	−0.44	0.22	0.12	0.01	−0.19	−0.08
P3	−0.34	0.09	0.77	−0.53	0.17	0.17	−0.52
P4	−0.3	0	0.2	−0.77	0.26	1.28	0.13
P5	−0.26	−0.16	0.18	−0.39	0.02	0.14	−0.32

Six Protein Types

	partial charge		Born radius	
	mean	SD	mean	SD
P0	−0.005	0.098	1.382	0.089
P1	−0.085	0.113	2.202	0.182
P2	−0.002	0.080	1.614	0.064
P3	0.003	0.103	1.839	0.080
P4	−0.295	0.011	0.894	0.070
P5	−0.164	0.028	0.900	0.065

Seven Ligand Types

	partial charge		Born radius	
	mean	SD	mean	SD
L0	0.003	0.094	1.764	0.084
L1	−0.001	0.086	1.537	0.063
L2	0.015	0.078	1.311	0.081
L3	0.010	0.121	2.144	0.203
L4	−0.156	0.040	0.851	0.047
L5	−0.297	0.013	0.851	0.047
L6	-0.164	0.059	1.025	0.064

[a] Atom types for the protein and ligand are named P0, P1, and P2 and L0, L1, and L2, respectively. The weighting parameter scaling the properties is $w = 0.55$.

3.2. Consensus Atom Types from Leave-1/3-Out Tests. Following the jackknife tests, the atom types, upon which each optimized potential is based, are compiled. When the same number of protein and ligand atom types appear frequently in multiple jackknife tests, they are considered "consensus atom types" and may form the basis of a robust and general KB potential. In the current case, we examine consensus atom types arising from the minimum in the left-in error described above. Although not a clear minimum, atom types are also compiled for the interesting case where $w = 1$. These weights ($w = 0.54$ and $w = 1.0$) correspond to atom types based on a combination of partial charge and Born radius and atom types based solely on the partial charge. The weight corresponding to the minimum error ($w = 0.54$) is also closely related to that corresponding to the maximum correlation obtained by optimizing the potential as described previously ($w = 0.55$). Consequently, the consensus atom types obtained through multiple jackknifing tests are virtually identical to the atom types obtained by optimizing to the overall correlation coefficient. The correlation between the result-

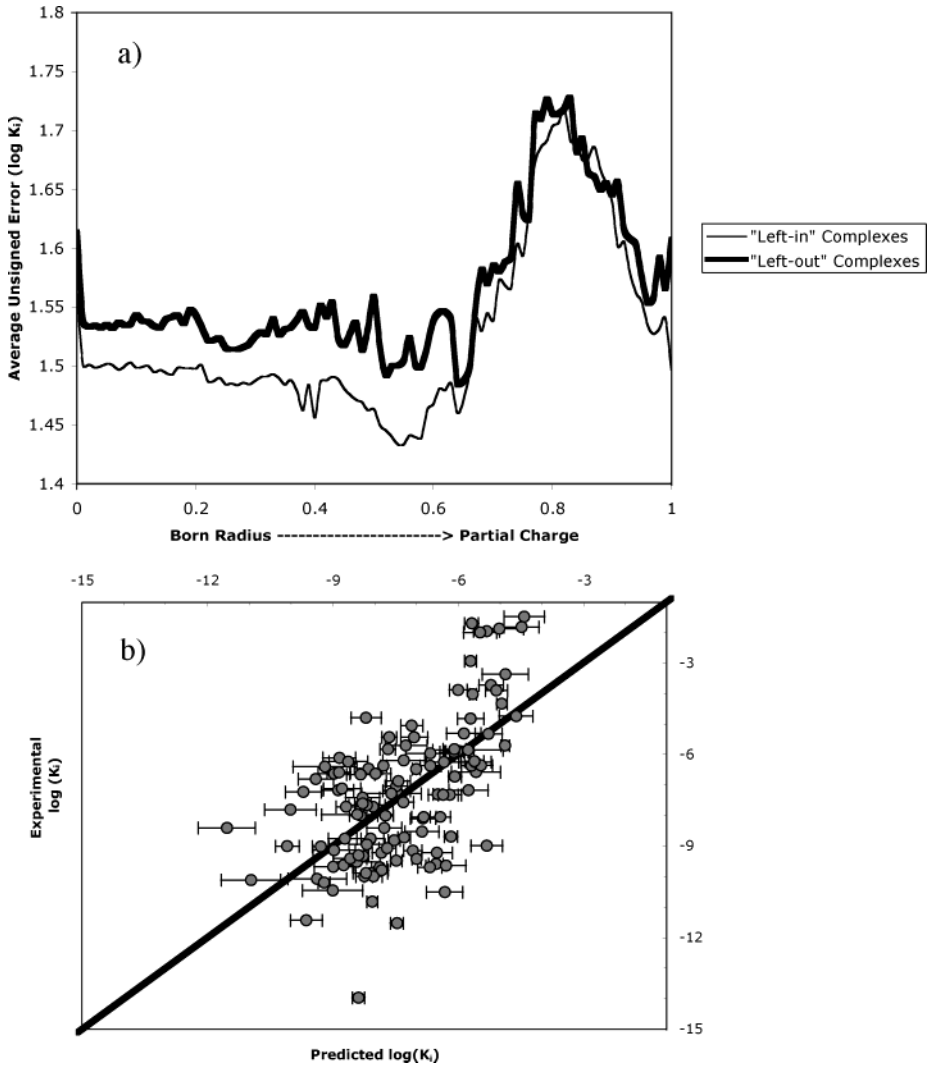


Figure 3. Diverse data set of 118 complexes jackknifed atom type optimizations. In panel a, the thick line shows the average unsigned error between the predicted energy and the experimentally determined binding energy (in the 1/3 of the database removed for testing) as a function of the weighting parameter "*w*" that scales from a Born radius atom description to a partial charge atom description. The error bars show the standard deviation of the unsigned errors accumulated over 20 random jackknife tests. The thin line shows the average unsigned error in the 2/3 of the database used in the optimization. These errors are lower since they are explicitly minimized in the optimization process. In panel b, the average predicted energies vs the experimental binding energies are shown using atom types based solely on partial charge (the minimum average unsigned error shown in panel a) for 118 diverse complexes. The error bars indicate the standard deviation of the predicted energies based on 20 random jackknife tests. The correlation coefficient between the average predicted affinities and the experimental affinities is $R = 0.6$.

ing SMoG energies using these slightly varying potentials is $R > 0.99$. The clusters associated with the protein and ligand consensus atom types based on partial charge and a combination of partial charge and the Born radius, as well as the corresponding potentials, are shown in Tables 3 and 4.

3.3. Leave-Family-Out Test. To more fully test the robustness and predictive ability of the model, a test was performed in which an entire target family was removed from the database of complexes of known binding affinity. The binding affinities of the members of the removed family were then predicted. This tests the scenerio where a novel target is introduced into a lead discovery protocol. In addition, this procedure tests the robustness of the random 1/3 jackknife test that demonstrated a predictive ability of 1.5 log K_i units.

After excluding the aspartic protease family, the metallo protease family and the sugar binding protein family, respectively, results suggest a robust model. As can be seen in Figure 4a−c, the minimum of the absolute error in the left-in complexes is not significantly changed by excluding any of the individual families. In Figure 5, a scatter plot showing the correlation between the experimental activity and the predicted activity is shown for the left-in points, while the left-out complex energies are also shown.

Binding Affinity Correlation within the Eight Protein Families. In addition to evaluating the binding affinities of a diverse structural database of protein/ligand complexes, it is also informative to probe the ability of the potential to predict binding affinities in protein families. By targeting individual families, binding enthalpy may become a better predictor of binding affinity as many entropic effects are attenuated. Improved predictors of binding affinity within protein−drug families can also have a significant role in the drug discovery process. While QSAR approaches have been found to be useful in this context,[4−6] the current

Table 3. Consensus Potential and Atom Type Clusters for Diverse Complexes[a]

0.0−3.50 Potential

	L0
P0	−0.19
P1	−0.12
P2	0.11
P3	0.21

3.50−4.50 Potential

	L0
P0	−0.01
P1	−0.06
P2	0.02
P3	−0.07

Four Protein Types

	partial charge		Born radius	
	mean	SD	mean	SD
P0	−0.293	0.040	N/A	N/A
P1	−0.536	0.020	N/A	N/A
P2	0.165	0.110	N/A	N/A
P3	−0.027	0.037	N/A	N/A

One Ligand Type

	partial charge		Born radius	
	mean	SD	mean	SD
L0	−0.079	0.216	N/A	N/A

[a] The consensus potential and atom type cluster definitions were obtained from multiple jackknifed optimizations against the diverse data set of 118 complexes with a property weight of $w = 1$.

approach to predicting structure-based determinants of binding enthalpy could have broader applicability to novel targets.

The optimal potentials found for each increment of the atomic property weighting parameter w were evaluated for their ability to correlate with binding affinities within eight protein target families comprising the diverse database of 118 complexes. The results demonstrate that while the optimal potential found using partial charge atom types showed a significant improvement in overall correlation within the 118 complexes (relative to SMoG 2001), the average intrafamily correlation coefficient dropped significantly (from 0.55 in SMoG 2001 to 0.38 using partial charge-based atom types). These results can be seen in Figure 6a. A scatter plot showing the intrafamily correlations based on partial charge and Born radius/partial charge atom types is shown in Figure 7.

To determine whether it was at all possible to derive partial charge-based KB potentials that improve the average intrafamily correlation, a further potential optimization was performed. At each increment of the weighting parameter, the optimal knowledge-based potential is chosen such that it provides the highest average intrafamily correlation. Previously, the optimal knowledge-based potential was chosen based on the highest total correlation within the diverse database of 118 complexes. The results indicate that although atom types based on partial charge alone are capable of generating knowledge-based potentials that improve correlation over the entire structural database, these atom types are not capable of improving the intrafamily correlation relative to SMoG 2001 (Figure 6b). Furthermore, atom types based on a combination of the Born radius and the partial charge are capable of significantly improving the intrafamily correlation (Table 5).

Table 4. Consensus Potential and Atom Type Clusters for Diverse Complexes[a]

0.00−3.50 Potential

	L0	L1	L2	L3	L4	L5	L6
P0	−0.48	0.38	−0.53	−0.52	−0.03	−0.51	−0.26
P1	−1.19	0.7	−0.54	1.32	−0.16	0.65	0.76
P2	−0.97	−0.12	0.35	0.78	−0.21	0.11	0.88
P3	−0.44	−0.6	1.04	0.09	−0.39	0.09	0.41
P4	−0.57	0.71	−0.61	−0.29	0.06	−0.31	−0.02
P5	−0.07	-1	1.45	0.07	−0.6	0.97	−0.37

3.50−4.50 Potential

	L0	L1	L2	L3	L4	L5	L6
P0	−0.32	0.13	0.39	−0.16	0.01	−0.26	0.19
P1	−0.97	0.51	−1.07	0.72	0.27	−0.17	1.49
P2	−0.53	0.13	−0.52	0.08	0.16	−0.34	0.78
P3	−0.05	−0.19	0.12	−0.44	-0.01	−0.46	0.22
P4	0.14	1.27	−0.76	0	0.26	−0.3	0.2
P5	0.28	−0.39	0.95	−0.29	−0.16	0.26	−0.4

Six Protein Types

	partial charge		Born radius	
	mean	SD	mean	SD
P0	−0.161	0.029	0.920	0.067
P1	−0.082	0.111	2.250	0.186
P2	0.002	0.102	1.878	0.082
P3	−0.001	0.079	1.649	0.065
P4	−0.290	0.011	0.914	0.072
P5	−0.005	0.097	1.412	0.091

Seven Ligand Types

	partial charge		Born radius	
	mean	SD	mean	SD
L0	−0.156	0.063	1.054	0.066
L1	−0.292	0.012	0.877	0.066
L2	0.010	0.119	2.191	0.207
L3	−0.001	0.085	1.571	0.064
L4	−0.154	0.039	0.872	0.049
L5	0.003	0.092	1.803	0.085
L6	0.014	0.077	1.342	0.082

[a] The consensus potential and atom type cluster definitions were obtained from multiple jackknifed optimizations against the diverse data set of 118 complexes with a property weight of $w = 0.54$.

Family Specific KB Potentials. Finally, we address the more focused approach to optimizing potentials for individual families. This approach attempts to find atom types whose corresponding knowledge-based potentials are accurate predictors of binding affinity within a specific class of protein targets. As addressed earlier, potentials that correlate well with relative binding affinities within individual target families could play a key role in lead discovery protocols. Three families were chosen in order to demonstrate the ability of optimized KB potentials to fit relative binding affinities. These families include the aspartic protease family, the sugar binding protein family, and the metallo protease family.

1. Optimizing the Correlation Coefficients. The analysis described here with regard to optimizing atom types for specific target families is symmetric with the analysis described earlier with regard to the diverse database of 118 complexes. Therefore, the first approach to addressing family specific KB potentials is to choose

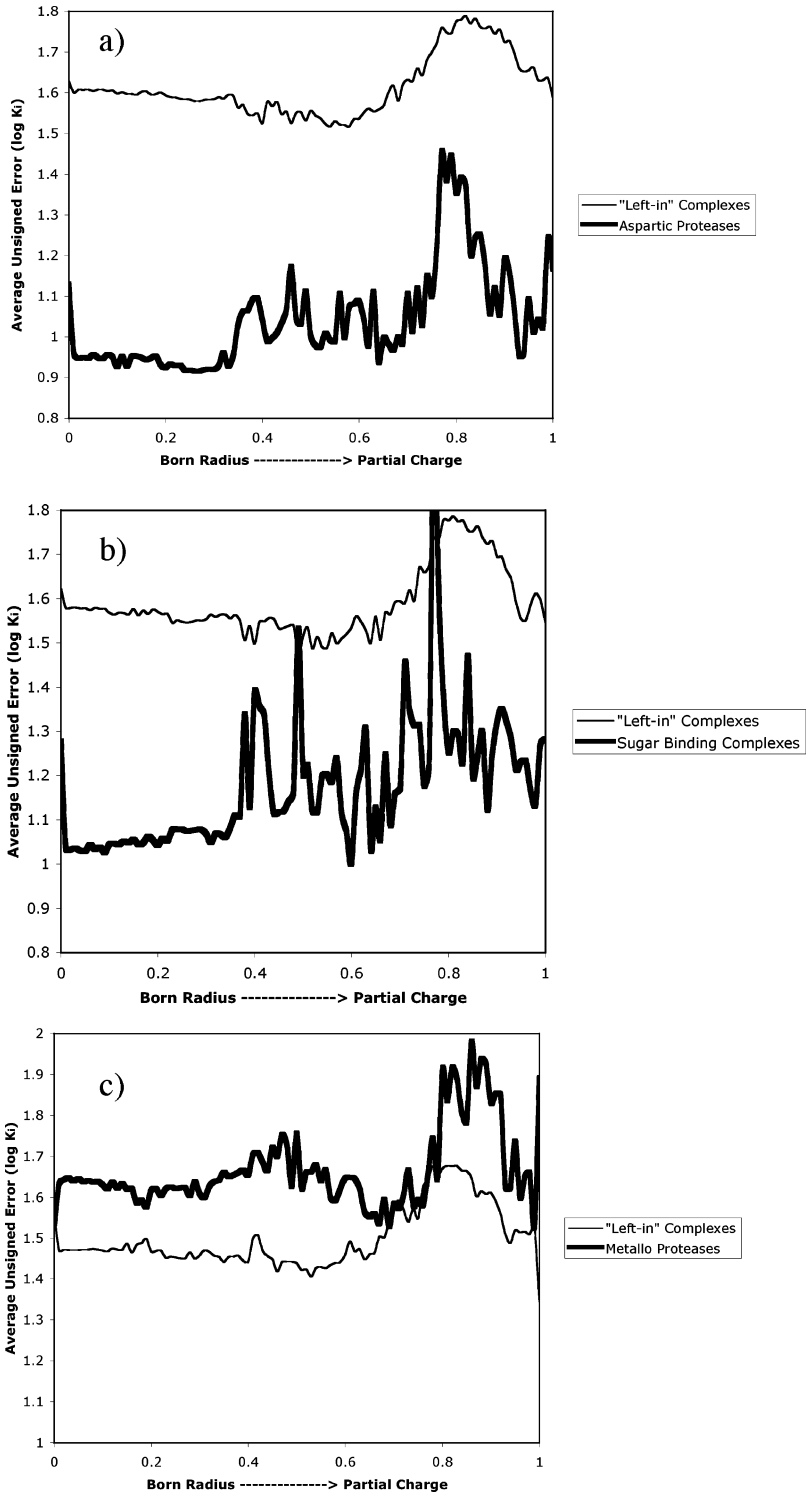


Figure 4. Average unsigned error obtained from jackknifing target families. The abscissa shows the various weighting factors used to balance the partial charge and Born radius properties when clustering to generate atom types. The ordinate shows the average unsigned error in units of log K_i. The minimum error in the "left-in" data set corresponds to a property weight of $w = 0.54$ in each of the three cases: (a) aspartic proteases, (b) sugar binding proteins, and (c) metallo proteases.

atom types that maximize binding affinity correlation within these families. As shown in Figure 8, it is possible to find atom types based on a combination of the Born radius and the partial charge that demonstrate very strong correlations to experimentally determined binding affinities for each of the three families. As in the exploration of the diverse database, one can identify the specific atom types that result in high correlations within each family.

In the case of the aspartic protease target family, potentials are found that correlate with $R = 0.81$ to experimental binding affinities (Figure 9a). Specifically, these potentials are found using a property weight of 0.61. As a reminder, this weight scales the partial charge property by 0.61 and scales the Born radius property by $1.0 - 0.61 = 0.39$. The atom types associated with this optimal correlation are quite unusual. The optimal atom types correspond to six protein atom types

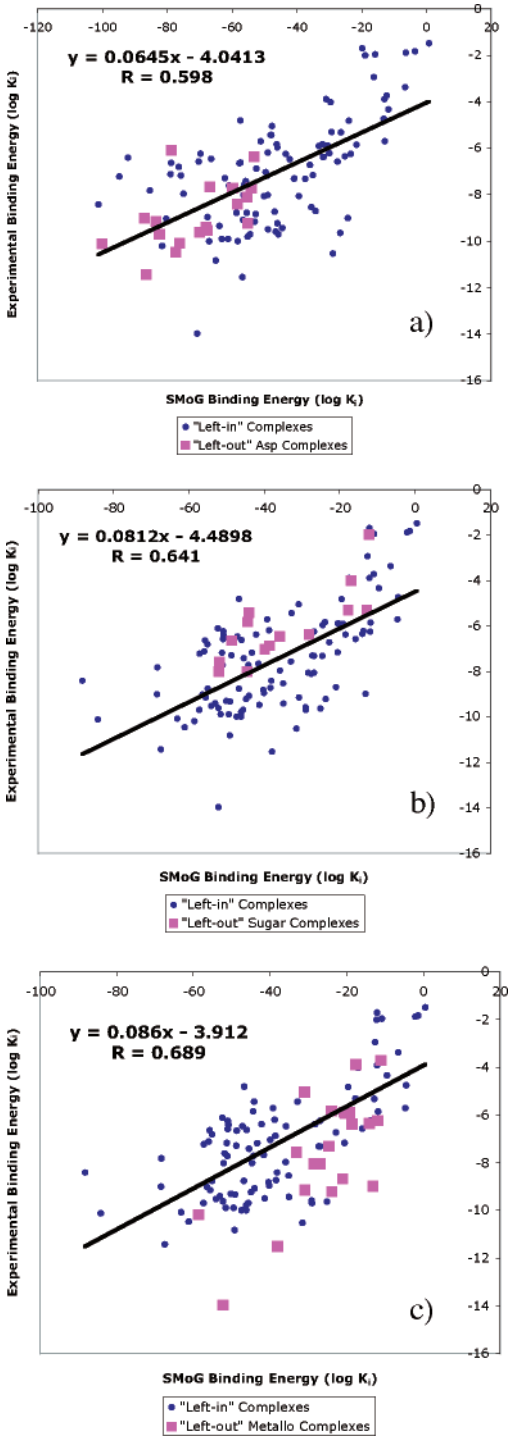


Figure 5. Scatter plots from jackknifing whole families from a diverse set of 118 complexes. The regression line and correlation coefficient are shown corresponding to the left-in complexes. These are representative potentials from four potential optimization runs. The potentials are taken from the property weight w corresponding to the minimum average unsigned error in the left-in complexes, $w = 0.54$ in all three jackknifing tests: (a) aspartic proteases, (b) sugar binding proteins, and (c) metallo proteases. The correlation coefficients, R, are given for the left-in complexes.

and a single ligand atom type (Table 6). This indicates a nonspecific potential. While this nonspecific potential may encourage a model perhaps reflective of a solvation burial penalty, the physical interpretation of this potential is challenging. The complexities of physically interpreting these models are discussed below.

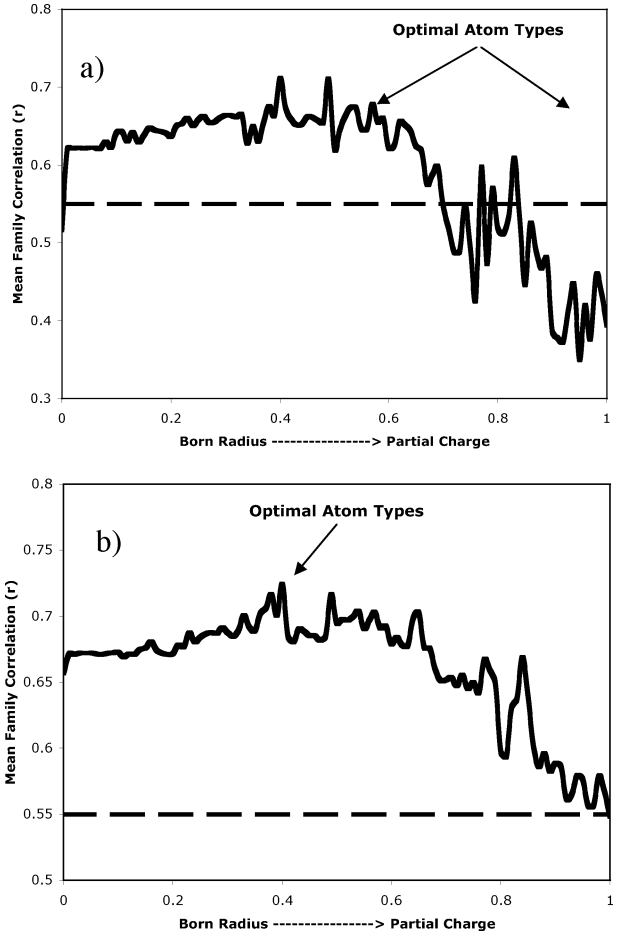


Figure 6. Intrafamily correlation coefficients are shown as a function of the weighting parameter between the partial charge and the Born radius. In panel a, the average intrafamily correlation coefficient is shown based on potentials obtained by optimizing the correlation coefficient of all 118 complexes at each weight. The dotted line indicates the average intrafamily correlation obtained in SMoG 2001. The "optimal atom types" label designates those weights (and therefore potentials) that correspond to the maximum correlation coefficient for all 118 complexes. In panel b, the potentials are optimized at each weight to provide the maximum average intrafamily correlation. Both plots show that using atom types based on a combination of the Born radius and partial charge, it is possible to derive potentials that improve intrafamily correlation with respect to SMoG 2001. The plots also demonstrate that this improvement is not possible using partial charge-based atom types.

Regarding sugar binding proteins, it is found that using a property weight of $w = 0.54$, it is possible to identify potentials that correlate with binding affinity with a coefficient of $R = 0.89$ (Figures 8 and 9b). The corresponding atom types are more complex than in the aspartic protease family, resulting in seven protein atom types and eight ligand atom types (Table 7). A similar situation arises with regard to the metallo protease family. Here, a correlation of $R = 0.90$ was obtained (Figures 8 and 9c) derived from a potential consisting of 15 protein atom types and 11 ligand atom types (Table 8). The property weight corresponding to the maximum correlation was found to be 0.36.

As discussed, these optimized atom types and the corresponding family specific knowledge-based potentials result in high correlations. Scatter plots show the correlation between the SMoG and the experimentally

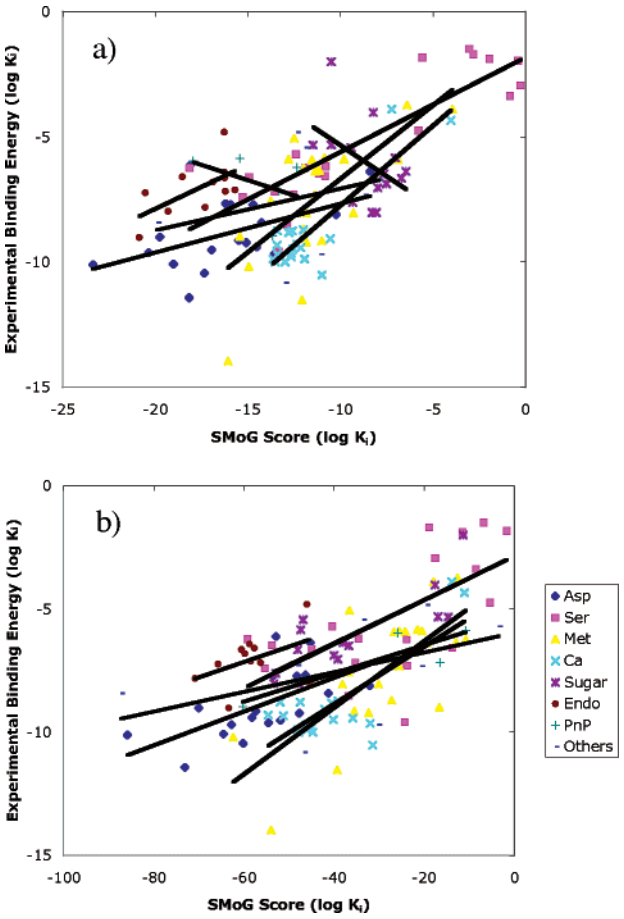


Figure 7. Scatter plot representing the improvement in the intrafamily correlation between experimental log K_i and predicted log K_i starting from (a) atom types described only on partial charge and (b) atom types described by a combination of the partial charge and Born radius.

Table 5. Intrafamily Correlations (R) Obtained through Various Atom Type Optimizations

family/ potential	SMoG 2001[a]	SMoG ($w = 1.0$)[b]	SMoG ($w = 0.55$)[c]	SMoG mean optimized[d]
Asp	0.62	0.49	0.61	0.42
Ser	0.84	0.86	0.68	0.71
Met	0.69	0.68	0.73	0.86
Ca	0.73	0.86	0.77	0.71
Sugar	0.47	−0.46	0.81	0.76
Endo	0.18	0.61	0.50	0.82
PnP	0.05	−0.41	0.86	0.97
others	0.74	0.30	0.48	0.58
average	0.55	0.37	0.68	0.73

[a] SMoG 2001 represents results derived from the algorithm published by Ishchenko et al. (2002). [b] SMoG ($w = 1.0$) represents the potential corresponding to partial charge atom types optimized according to the total correlation coefficient against the full testing set of 118 complexes. [c] SMoG ($w = 0.55$) is the optimized potential obtained using atoms described by a combination of the partial charge and the Born radius. [d] SMoG (mean optimized) is the potential obtained by explicitly optimizing the potential in order to achieve the maximum possible average intrafamily correlation.

determined binding affinities for each target family and are shown in Figure 9.

2. Jackknifing Tests. To get a better idea of the predictive ability of these family optimized potentials, jackknifing tests were performed to assess the expected predictive error. At each weight, instead of choosing the potential that optimizes the total correlation, a potential is chosen that optimizes the correlation within a random

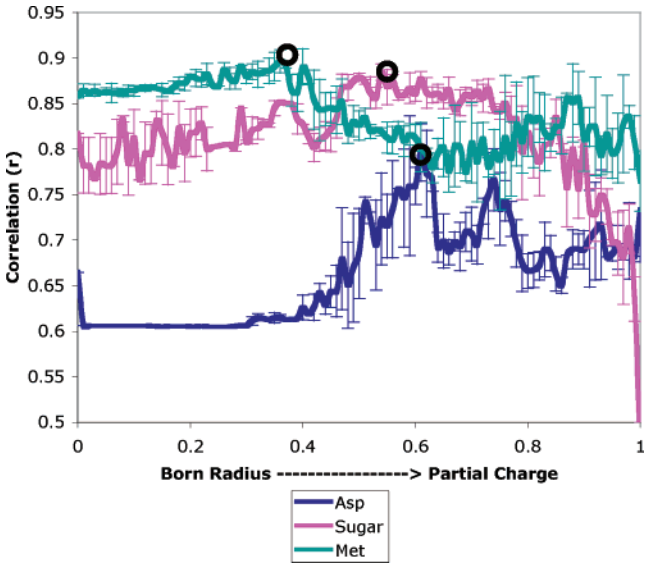


Figure 8. Atoms in this study are described by a combination of the partial charge and the Born radius. The relative contribution of these two descriptors, used to cluster atoms into atom types, is scaled between 0 and 1. Plotted is the maximum possible correlation using clusters (atom types) based on a particular combination of these parameters. Three independent optimizations are performed on the aspartyl protease family, the sugar binding protein family, and the metallo protease family. Four optimizations are performed per family, and some variability in the resulting correlation coefficients is observed due to the stochastic nature of the K-means clustering algorithm used to define the atom types corresponding to each weight. This variability is shown in the error bars representing one standard deviation. Combinations of descriptors that yield high correlation to the experimental binding affinity in these three families are indicated by circles.

2/3 of the database. The potential is evaluated by determining the unsigned error in predicting the remaining 1/3 of the database that was removed for independent testing.

2.1. Aspartic Protease Family. The first example generates a knowledge-based potential capable of accurately predicting the binding free energy within a family of aspartic proteases to an accuracy of approximately 1.1 log K_i units or 1.5 kcal/mol. It is possible to achieve this minimal predictive error in the aspartic protease family using a KB potential based on atom types generated based on a combination of the Born radius and partial charge. In other words, the minimal left-in error is found using a weight $w = 0.61$ (Figure 10a). A scatter plot shows the average predicted energy vs the experimental binding energy for the 18 complexes comprising the aspartic protease family database (Figure 10b). The atom types and the corresponding potential for this weight are shown in Table 6. An interesting feature in the predicted or left-out error is the minimum at $w = 1.0$. Information relevant for distinguishing solvation effects within this family, such as the Born radius, was not required. This does not imply that solvation effects are irrelevant in the energetics of binding within this family. It simply implies that within this group of protein−ligand complexes, solvation effects may not strongly differentiate between binding propensities and may cancel out in binding energy calculations.

2.2. Sugar Binding Protein Family. By investigating sugar binding proteins, however, we see a different

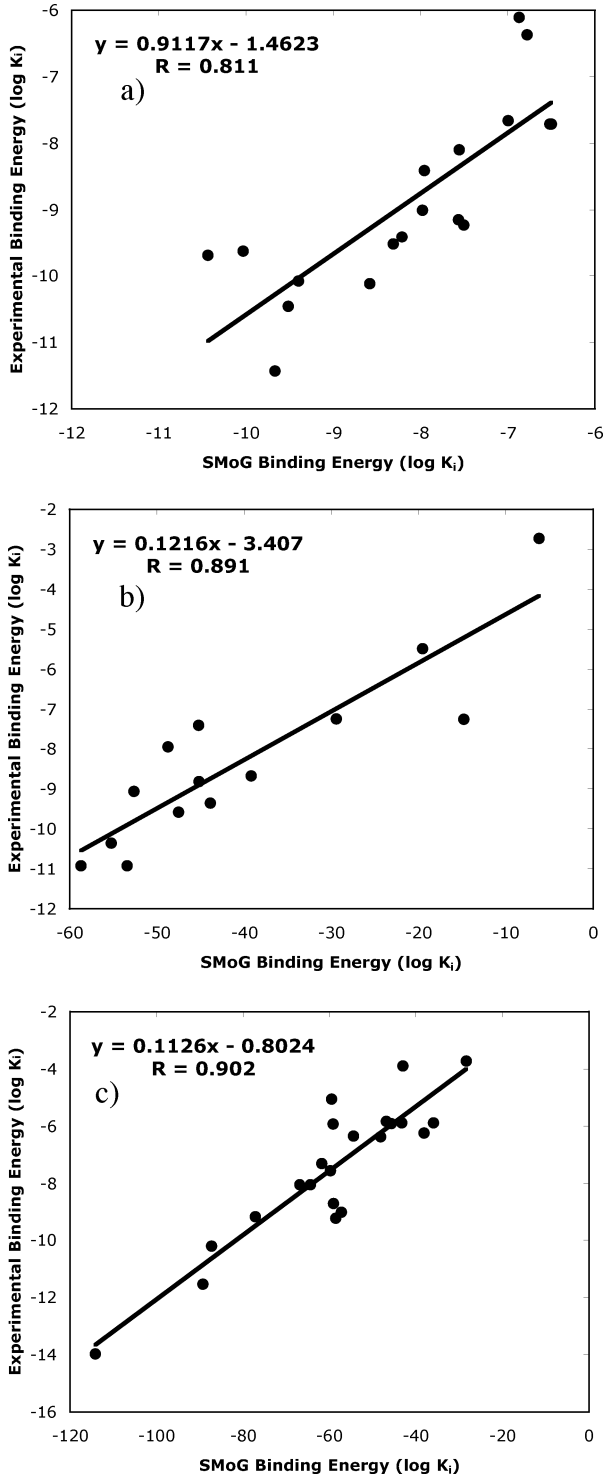


Figure 9. Optimized knowledge-based potentials for each of the individual families. The SMoG energy vs the experimentally determined binding energy are shown for the (a) aspartic protease, (b) sugar binding proteins, and (c) metallo protease families.

picture. Here, solvation effects are quite important for differentiating the binding propensities of the ligands within this family. Electrostatic effects have been suspected to be quite important in binding within this family,[11] and solvation properties appear to be key to achieving small predictive errors (Figure 11a). The optimal weight within the left-in set balances the partial charge, and the Born radius is $w = 0.55$. The atom types and corresponding potential are shown in Table 7. A

Table 6. Optimized Potential and Atom Type Clusters for Aspartic Proteases[a]

0.0−3.50 Potential

	L0
P0	−0.23
P1	−0.21
P2	−0.04
P3	−0.14
P4	0.09
P5	0.03

3.50−4.50 Potential

	L0
P0	−0.04
P1	0.16
P2	0
P3	0.01
P4	−0.09
P5	−0.03

Six Protein Types

	partial charge		Born radius	
	mean	SD	mean	SD
P0	−0.182	0.031	0.785	0.066
P1	0.079	0.075	1.564	0.087
P2	0.009	0.097	1.228	0.084
P3	−0.327	0.012	0.775	0.061
P4	−0.046	0.080	1.425	0.066
P5	−0.119	0.108	1.831	0.172

One Ligand Type

	partial charge		Born radius	
	mean	SD	mean	SD
L0	−0.048	0.132	1.218	0.316

[a] Atom types for the protein and ligand are named P0, P1, and P2... and L0, L1, and L2..., respectively. The scaling parameter for the properties is $w = 0.61$.

scatter plot showing the results of experimental vs predicted binding energies is also shown for the 14 complexes comprising this database (Figure 11b). These results demonstrate that tighter binders appear more accurately predicted than weaker binders. However, the average predictive error is quite low at 1.3 log K_i units or 1.8 kcal/mol.

2.3. Metallo Protease Family. Finally, we investigated a metallo protein data set and found again that atom types based on a combination of the Born radii and partial charge are beneficial in generating predictive KB potentials. Although requiring a different weight describing the balance between the Born radius and the partial charge descriptors, the minimum predictive error was found to be quite small (Figure 12a). The optimal weight resulting in the minimum predictive error is $w = 0.35$. The atom types and corresponding potentials are shown in Table 8. In the case of the metallo proteins, the minimum predictive error was found to be 1.1 log K_i units or 1.5 kcal/mol. As with the sugar binding protein family, the metallo protein family might have been supposed to benefit from some solvation information given the potentially strong electrostatic effects in binding due to bound metals. Again, also shown is a scatter plot demonstrating a rather consistent predictive ability of the minimum error model (Figure 12b). No significant difference is found in the model's ability to predict the affinity of weak or tight binders.

Table 7. Optimized Potential and Atom Type Clusters for Sugar Binding Proteins[a]

0.0−3.50 Potential

	L0	L1	L2	L3	L4	L5	L6	L7
P0	−0.1	0.85	−0.27	0.49	0.44	−0.77	−0.07	0.2
P1	0.22	−0.09	−0.15	1.48	1.14	−1.17	0.41	−0.11
P2	−0.74	0.82	−0.49	0.65	0.66	−1.04	0.35	0.64
P3	−0.93	1.35	−0.6	−0.2	0.08	−0.4	0.02	1.09
P4	0.39	−0.57	−0.16	−0.57	−0.31	−0.71	−0.68	−0.72
P5	0.78	−0.99	−0.11	1.06	0.46	−1.26	0.62	−0.34
P6	−1.14	0.08	−0.74	−0.15	−0.67	−0.15	0.74	1.35

3.50−4.50 Potential

	L0	L1	L2	L3	L4	L5	L6	L7
P0	0.01	-0.08	0.03	−0.09	0.67	−0.34	−0.45	−0.34
P1	0.3	−0.88	0.34	0.66	1.32	−0.76	−0.12	−0.51
P2	0.01	−0.29	0.15	0.23	0.4	−0.12	−0.01	0.02
P3	−0.3	0.85	−0.07	−0.5	0.02	0.08	−0.44	0.2
P4	0.3	−0.57	0.01	−0.15	0.15	−0.28	-0.37	−0.49
P5	0.59	−1.26	0.39	0.73	1.5	−0.92	−0.06	−0.83
P6	−0.43	1.09	−0.19	−0.25	−0.49	0.23	0.34	0.96

	partial charge		Born radius			partial charge		Born radius	
	mean	SD	mean	SD		mean	SD	mean	SD
	Seven Protein Types					Eight Ligand Types			
P0	0.032	0.065	1.734	0.053	L0	−0.292	0.013	0.874	0.063
P1	0.060	0.059	1.939	0.080	L1	0.008	0.129	2.429	0.223
P2	−0.156	0.020	1.891	0.113	L2	−0.153	0.039	0.868	0.047
P3	−0.016	0.073	1.575	0.054	L3	0.009	0.085	1.526	0.058
P4	−0.199	0.062	0.916	0.065	L4	0.008	0.073	1.310	0.077
P5	−0.089	0.108	2.320	0.184	L5	−0.169	0.051	1.037	0.062
P6	−0.003	0.016	1.364	0.080	L6	−0.007	0.089	1.720	0.064
					L7	0.012	0.103	1.992	0.094

0.00−3.50 Potential

	L0	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
P0	0.66	−1.05	0.08	1.13	0.83	0.51	0.36	−0.46	−0.52	−0.73	−0.24
P1	−0.16	−0.82	0.86	−0.84	−1.51	0.73	0.5	1.43	−1.91	1.86	1.63
P2	−0.05	−0.41	0.98	−1.16	−1.89	−0.62	0.9	1.95	−2.2	1.97	3.13
P3	0.59	−0.96	0.22	0.89	0.8	0.23	0.17	−0.01	−0.86	−0.54	−0.02
P4	−1.36	−0.7	−0.84	−1.36	−1.48	−0.15	−0.13	−0.1	−1.66	1.06	−0.13
P5	2.04	−0.68	0.62	2.39	2.21	−0.86	−0.66	−0.06	0.01	−1.32	−0.28
P6	−0.56	−0.95	0.36	−0.51	−0.49	0.77	0.42	1.07	−1.56	0.8	1.27
P7	0.3	−1.04	−0.24	0.7	1.97	0.31	1.28	−0.08	−0.81	−0.35	0.8
P8	−0.26	−0.37	−0.75	−0.09	3.13	−0.99	−0.87	−0.72	−0.05	−0.07	−0.74
P9	−0.11	−0.93	−0.07	0.27	0.31	0.52	0.62	0.03	−1.16	0.01	0.97
P10	−0.62	−0.56	−0.85	−0.72	−0.49	−0.25	0.67	−1	−0.97	0.12	−0.78
P11	−0.67	−0.85	0.19	−0.44	−0.44	1.52	0.28	0.55	−1.49	0.57	1.1
P12	0.83	−0.55	1.44	−0.06	−1.28	−0.71	−1.48	0.8	0.55	−1.42	0.04
P13	1.87	−0.7	0.1	2.44	0.88	−0.21	−0.08	−0.07	−0.06	−1.11	−0.62
P14	0.1	−0.96	−0.32	0.81	1.39	0.27	0.17	−0.13	−0.33	−0.66	−0.1

3.50−4.50 Potential

	L0	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
P0	0.05	−0.43	−0.48	0.43	0.38	0.12	0.35	−0.37	0.08	−0.08	−0.48
P1	−1	−0.4	−0.22	−1.44	1.9	1.82	1.94	0.95	−1.51	2.6	1.57
P2	−1.2	−0.1	−0.34	−1.57	−2.02	1.39	181	1.28	−1.47	2.98	2.39
P3	0.11	−0.17	−0.38	0.27	0.27	−0.27	−0.07	−0.54	0.16	−0.09	−0.54
P4	−0.77	−0.33	−0.54	−0.97	−1.36	0.65	1.47	0.05	−0.89	1.13	0.42
P5	0.5	−0.26	0.32	0.78	1.36	−0.58	−0.51	−0.12	0.43	−0.56	−0.42
P6	−0.82	−0.51	−0.24	−1.02	−1.54	1.72	1.57	0.46	−1.46	1.36	1.42
P7	−0.35	−0.56	−0.58	−0.2	0.54	0.18	0.85	−0.52	−0.08	−0.02	−0.25
P8	−0.03	−0.16	−0.17	0.05	0.44	−0.43	−0.38	−0.28	0.49	−0.15	−0.41
P9	−0.58	−0.62	−0.54	−0.55	−0.36	0.69	1.01	−0.31	−0.66	0.45	0.08
P10	−0.64	−0.42	−0.7	−0.73	−0.77	0.21	0.6	−0.58	−0.45	0.09	−0.31
P11	−0.75	−0.5	−0.53	−0.94	−1.24	1.16	1.55	−0.07	−1.11	0.71	0.5
P12	1.44	0	1.27	1.09	1.67	−0.57	−0.88	0.28	0.95	−0.74	−0.18
P13	0.49	−0.28	−0.16	0.72	1.06	−0.39	0.06	−0.59	0.48	−0.32	−0.58
P14	−0.29	−0.45	−0.72	−0.03	0.39	0.08	0.53	−0.54	0.2	−0.06	−0.39

[a] Atom types for the protein and ligand are named P0, P1, and P2... and L0, L1, and L2..., respectively. The scaling parameter for the properties is $w = 0.54$.

Table 8. Optimized Potential and Atom Type Clusters for Metallo Proteases[a]

15 Protein Types

	partial charge		Born radius	
	mean	SD	mean	SD
P0	−0.002	0.049	2.251	0.036
P1	−0.055	0.074	3.135	0.097
P2	−0.048	0.078	3.568	0.236
P3	−0.106	0.011	2.389	0.080
P4	−0.140	0.051	1.436	0.063
P5	−0.002	0.060	1.969	0.050
P6	0.045	0.043	2.855	0.068
P7	−0.005	0.053	2.372	0.036
P8	−0.138	0.042	1.180	0.040
P9	0.030	0.042	2.497	0.041
P10	−0.126	0.046	1.291	0.033
P11	0.036	0.038	2.656	0.050
P12	−0.009	0.082	1.773	0.070
P13	−0.001	0.052	2.125	0.040
P14	−0.099	0.023	2.575	0.071

11 Ligand Types

	partial charge		Born radius	
	mean	SD	mean	SD
L0	0.006	0.064	2.686	0.075
L1	−0.128	0.041	1.328	0.051
L2	0.001	0.061	2.451	0.057
L3	0.003	0.077	2.981	0.115
L4	0.015	0.096	3.582	0.308
L5	0.005	0.054	1.953	0.052
L6	0.009	0.048	1.758	0.063
L7	−0.005	0.057	2.276	0.947
L8	−0.079	0.063	1.532	0.061
L9	−0.126	0.047	1.170	0.049
L10	0.005	0.056	2.122	0.046

[a] Atom types for the protein and ligand are named P0, P1, and P2..., and L0, L1, and L2..., respectively. The scaling parameter for the properties is $w = 0.36$.

As in the case with the diverse data set, consensus atom types were extracted from multiple jackknifing runs for each of the three families. The resulting atom types were nearly identical ($R > 0.99$) to those determined based on optimizing against the entire data set. The small differences arise primarily due to the stochastic nature of the *K*-means clustering.

Discussion

The objective of this work is to introduce atomic descriptors that capture features of the optimal effective pairwise potential. To elucidate this, we refer to the true multibody potential describing the energetics in real systems. We may imagine that there exists a projection of this multibody potential to a pairwise potential that minimizes the loss of information. It is this optimal effective pairwise potential that we are attempting to extract through the knowledge-based procedure. The basic approach of extracting a true pairwise potential from a structural database has been validated in previous self-consistent studies on artificially created databases.[17−19] One significant difference preventing the results from these self-consistent studies to be applied to real chemical systems is the agreement between the particle type definitions and the effective pairwise potential. By describing the atom types within this study according to descriptors consistent with well-tested, classical force fields, we attempt to improve this agreement and extract potentials consistent with the optimal effective pairwise potential.

The premise of characterizing particles within a knowledge-based potential has been described previously. In one example, residues were described not only according to the traditional classification but also based on the residue's corresponding secondary structural element.[2] In another example, it has even been shown that the grouping of residues by burial can improve fold prediction.[20] In these examples, as well as other KB potentials described in the literature, particles are typed a priori and are not systematically optimized. The systematic optimization of atom types is a unique aspect of the current work and creates a hybrid between traditional knowledge-based potentials and successful empirical scoring functions such as those used in QSAR models.[4−6]

What Are the Implications of the Optimal Atomic Properties Found? Knowledge-based methods and any statistical method for predicting physical quantities may behave in unpredictable ways. When optimizing any incomplete statistical function, including pairwise knowledge-based potentials, the optimization process will combine descriptors in such a way as to produce the optimal agreement with experiment. This optimal agreement may not result from a simple linear combination of well-known physical relationships. In other words, the optimization process is simply a mathematical process and while the optimized function may reproduce a physical property, the components of the optimized function are not typically amenable to a clear physical interpretation. Even in the case of linear interaction models,[21] where the components of the optimized function are weighted physical interaction terms, interpreting the magnitude of the optimized weights as expressing the "importance" of the associated physical interaction is questionable.

It is tempting to link the importance of the partial charge and Born radius descriptors used in this work to the contributions of Coulomb and solvation forces, respectively, in the protein/ligand interactions of various complexes. For the reasons described above, this link is not trivial to establish. In the work presented in this paper, optimal atom types based only on partial charge were capable of producing an improved correlation with binding affinity (relative to SMoG 2001) within a diverse database of 118 complexes. Does this suggest that Coulomb or even electrostatic interactions are sufficient to describe these interactions? No.

To demonstrate the difficulty in ascribing a simply physical interpreation to these optimized knowledge-based potentials, consider an extreme case. By examining the consensus atom types that result from the jackknifed optimization of the 118 complex database, we find four protein atom types (clustered by partial charge) and only one ligand atom type. Obviously, the resulting correlation coefficient of 0.6 is not obtained through specific Coulomb interactions represented within the KB potential. As the model becomes more complete, by representing more aspects of the physics of pairwise intraatomic interactions, a more intuitive physical model may emerge. Until statistical models can very accurately predict the effects of intraatomic interactions (such as binding affinity) in the context of a diverse

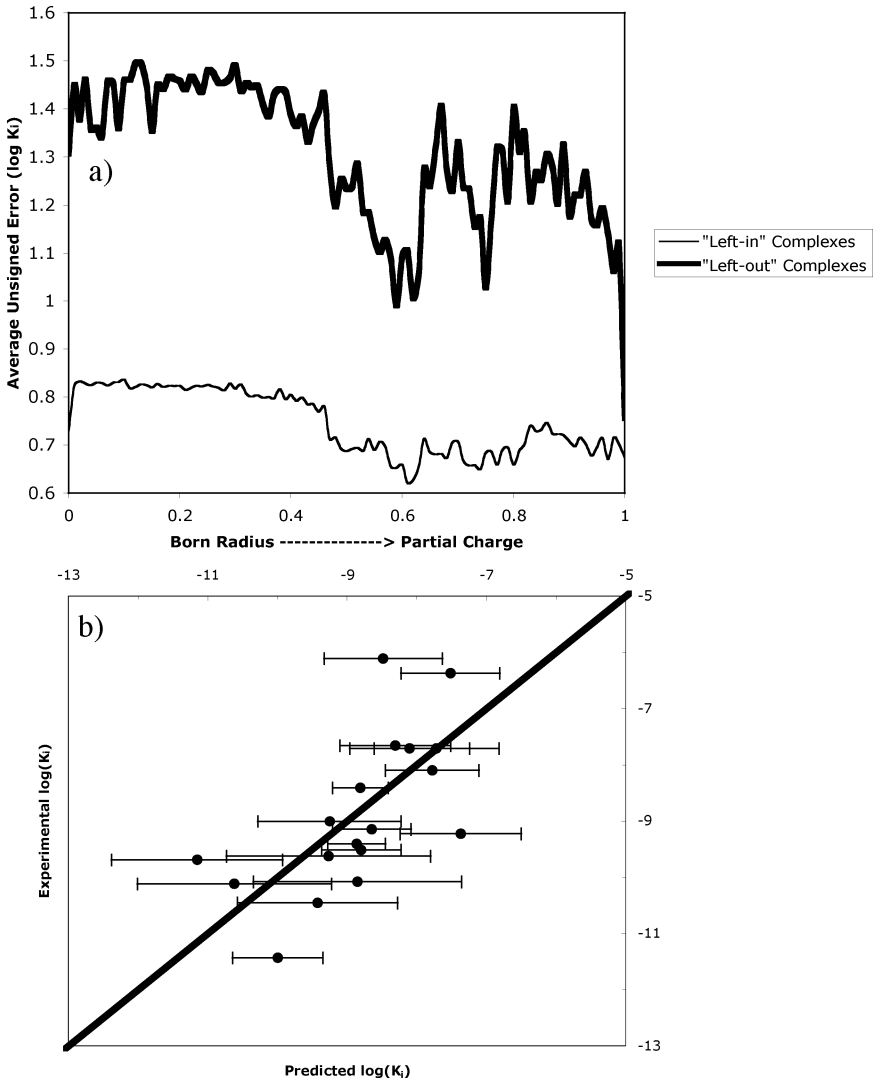


Figure 10. Aspartic protease family jackknifed atom type optimizations. In panel a, the red line shows the average unsigned error between the predicted energy and the experimentally determined binding energy (in the 1/3 of the database removed for testing) as a function of the weighting parameter w that scales from a Born radius atom description to a partial charge atom description. The error bars show the standard deviation of the unsigned errors accumulated over 20 random jackknife tests. The thin line shows the average unsigned error in the 2/3 of the database used in the optimization. These errors are lower since they are explicitly minimized in the optimization process. In panel b, the average predicted energies vs the experimental binding energies are shown using atom types based on a weight of $w = 0.61$ (the minimum average unsigned error shown in panel a) for the left-in aspartic protease complexes. The error bars indicate the standard deviation of the predicted energies based on 20 random jackknife tests. The correlation coefficient between the average predicted energies and the experimentally determined energies is $R = 0.65$.

database of complexes, physical interpretations of these models must be considered only hypotheses.

It should be noted, however, that the parameters used in the current application do appear to contain useful information. In Figure 14, we demonstrate the result in using random numbers to describe atomic properties. Obviously, in this case, it was impossible to generate potentials that resulted in binding energies strongly correlated with experimentally determined affinities. While it is difficult to quantify, the atom types assigned by chemical intuition in SMoG 2001 also resulted in some increase in information and more effective correlations with known binding affinities. By generating atom types through clustering atomic properties such as the Born radius and partial charge, it has been shown that it is possible to further improve the agreement between a simple knowledge-based potential and known binding affinities. In summary, while it is difficult to interpret the physical nature of the interactions, it is clear that the physical properties used to assign atom types is providing useful information to the knowledge-based potential.

Assessing the Predictive Ability of Optimized KB Potentials. Although interpreting the physical significance of the atom types generated in this work may not be sound, it is the model's ability to predict binding affinity that is of practical concern. In this paper, we optimize atom types (and therefore KB potentials) based on three decompositions of the protein/ligand complex database containing associated binding affinities. These include the correlation to binding affinity within the entire database of 118 complexes, the average correlation to binding affinity within each of the eight families comprising the 118 complex database, and the correlation to binding affinity within individual

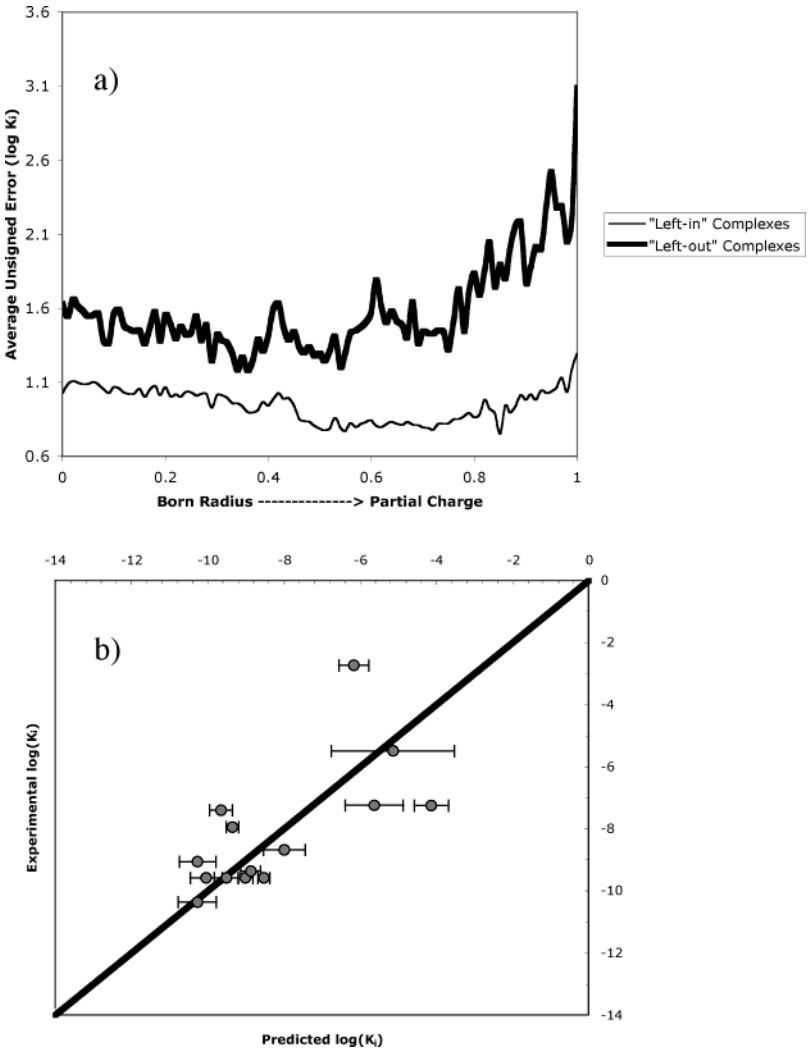


Figure 11. Sugar binding protein family jackknifed atom type optimizations. In panel a, the thick line shows the average unsigned error between the predicted energy and the experimentally determined binding energy (in the 1/3 of the database removed for testing) as a function of the weighting parameter w that scales from a Born radius atom description to a partial charge atom description. The error bars show the standard deviation of the unsigned errors accumulated over 20 random jackknife tests. The thin line shows the average unsigned error in the 2/3 of the database used in the optimization. These errors are lower since they are explicitly minimized in the optimization process. In panel b, the average predicted energies vs the experimental binding energies are shown using atom types based on a weight of $w = 0.55$ (the minimum average unsigned error shown in panel a) for 14 sugar binding complexes. The error bars indicate the standard deviation of the predicted energies based on 20 random jackknife tests. The correlation coefficient between the average predicted energies and the experimentally determined energies is $R = 0.68$.

target families. Each of these approaches contributes to solving practical problems in drug design.

1. Diverse Complex Database. In the first case, we consider the entire diverse database of 118 complexes. While training on a specific target may be preferable in canceling out effects such as binding entropy, which are typically difficult to model, a completely novel protein target may require a more generalized approach. In such cases, where little is known about alternative inhibitors or related targets, it may still be possible to assist lead discovery using more generic potentials. In the work described here, we optimized atom types in order to generate knowledge-based potentials capable of predicting binding affinity within a diverse database of 118 complexes. On the basis of jackknifed results, we found potentials that were capable of predicting binding affinity within 3.5 kT or 2.0 kcal/mol on average. In practical terms, this suggests that within this database, the predictive accuracy is approximately ± 1.5 log K_i units. While certainly not a perfect prediction, this level of accuracy is capable of screening out poor binders. From Figure 3b, it is clear that the predicted log K_i is capable of isolating or screening out inhibitors that bind with micromolar affinity (log $K_i = -6$) or greater.

Furthermore, optimized atom types result in a potential that was found to correlate to binding affinity with a coefficient of $R = 0.63$. This is a significant improvement over the correlation obtained with the SMoG 2001 potential of 0.43. By optimizing the atom types, this simple knowledge-based potential is consistent with other scoring functions in evaluating binding affinity in diverse databases of protein/ligand complexes. An example of this may be seen from the PMF scoring function.[13] Using a ligand volume correction, it is found that the knowledge-based potential is capable of greater correlations with binding affinities in diverse databases.[22] In six distinct diverse databases containing between 61 and 170 complexes, it was found that binding correlations resulted in coefficients ranging from 0.52 to 0.73. This is consistent with the observed

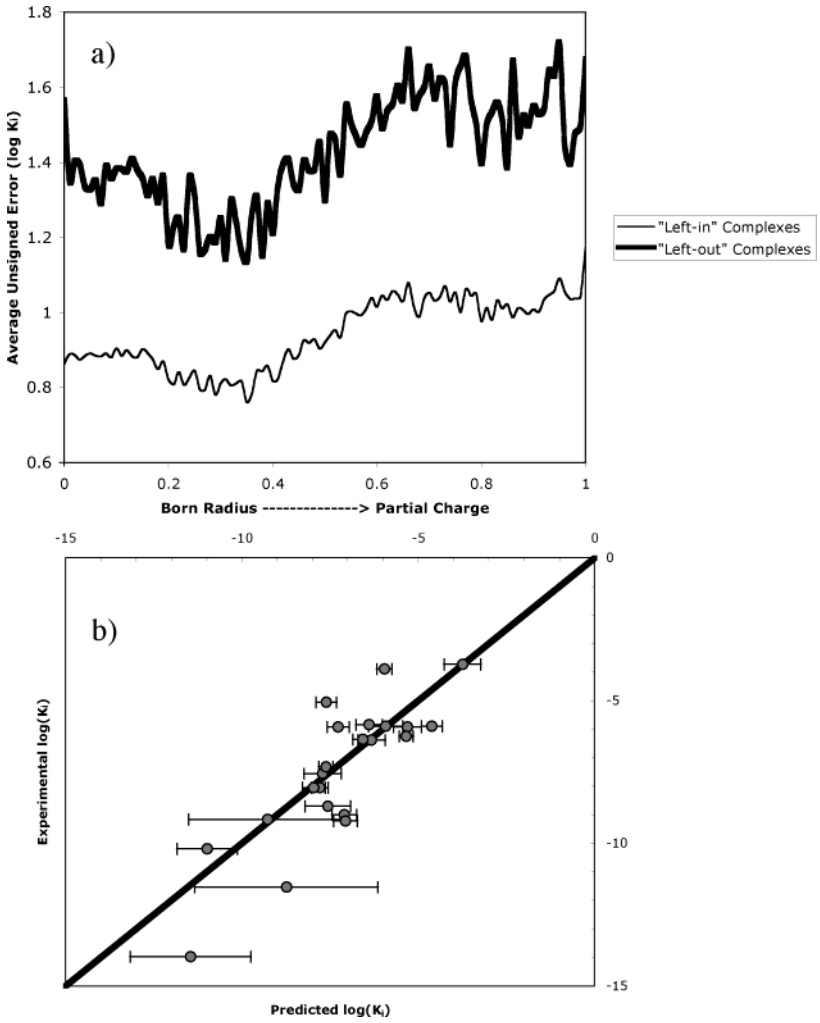


Figure 12. Metallo protein family jackknifed atom type optimizations. In panel a, the thick line shows the average unsigned error between the predicted energy and the experimentally determined binding energy (in the 1/3 of the database removed for testing) as a function of the weighting parameter w that scales from a Born radius atom description to a partial charge atom description. The error bars show the standard deviation of the unsigned errors accumulated over 20 random jackknife tests. The thin line shows the average unsigned error in the 2/3 of the database used in the optimization. These errors are lower since they are explicitly minimized in the optimization process. In panel b, the average predicted energies vs the experimental binding energies are shown using atom types based on a weight of $w = 0.35$ (the minimum average unsigned error shown in panel a) for 22 metallo protein complexes. The error bars indicate the standard deviation of the predicted energies based on 20 random jackknife tests. The correlation coefficient between the average predicted energies and the experimentally determined energies is $R = 0.84$.

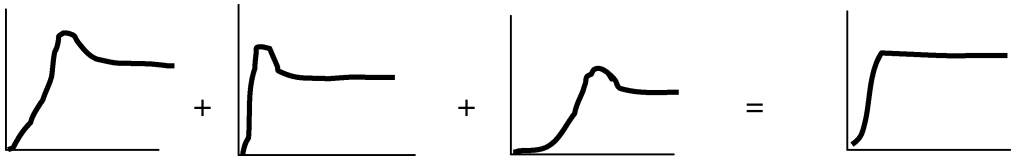


Figure 13. This cartoon illustrates the effect of mixing atom types on simple radial distribution functions. These functions describe the local density of contacts made between particles exhibiting varying potential attraction. As the contact probability functions corresponding to "native" atom types are mixed in arbitrary atom typing, information is lost resulting in probability functions that primarily reflect information regarding bulk density within the system.

correlation of 0.63 observed using consensus atom types derived from a diverse database of 118 complexes.

While the optimized KB potentials described here provide results consistent with other simple KB potentials, these potentials are in strict competition with the equally fast QSAR-like models. It should be mentioned that it is possible to derive QSAR-like models, or empirical scoring functions, that are capable of improved correlations and lower predictive errors. A recent example of such a potential demonstrated a predictive error of 3.4 kT or ~2.0 kcal/mol and a correlation coefficient of $R = 0.76$ for a larger data set of 200 protein−ligand complexes.[7] While QSAR models are currently equally fast and accurate relative to KB potentials, one advantage of KB potentials may be their physical basis. Although still an active area of debate,[15,16,23−25] the exploration of the physical nature of KB potentials may eventually lead to even more robust and accurate models of macromolecular energetics.

2. Analysis of Intrafamily Correlation. Next, we evaluated the intrafamily correlation coefficients within

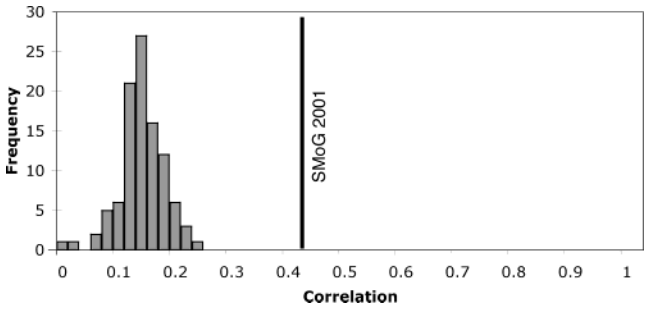


Figure 14. Assigning random atom types to the structural databases used within SMoG results in a distribution of low correlations with experimental binding affinities resulting from the associated knowledge-based potentials. Using SMoG 2001 atom types, the correlations improve. By further optimizing the atom types, the aim of this work is to further improve the correlation between predicted and experimental binding energies.

the diverse database of 118 complexes. Intrafamily correlation coefficients are advantageous with respect to the correlation coefficient of the entire diverse database, in that the families are more homogeneous potentially resulting in the cancellation of entropic and other specific energy terms. Additionally, binding constants obtained from diverse complex databases are difficult to assess when comparing distinct target proteins. The reason is that binding constants are sensitive to the conditions under which the experiment is run. These conditions typically vary from one target to another and therefore bring into question the inherent limits of one's ability to predict relative binding affinities between disparate protein targets. Our analysis has shown that using atom types based on both the Born radius and the partial charge, it is possible to derive knowledge-based potentials that can be generally applied to the determination of relative binding affinity within a variety of target families. The average correlation coefficient across eight target families was found to be $R = 0.68$, and a significant improvement over the average of $R = 0.55$ was found in SMoG 2001. This has obvious applications in the lead discovery process where one would like to either assess multiple lead compounds to a given target or engage in an automated lead discovery protocol. This has particular use in the beginning stages of lead discovery when examining a new target for which no structures of related complexes are available.

Finally, we consider an approach to optimize KB potentials for individual families. In this scenario, we optimized three sets of atom types in order to produce the optimal correlation within three distinct protein families. The families chosen were the aspartic proteases, the sugar binding proteins, and the metallo proteins. As in the optimization performed on the average intrafamily correlation, optimization on single families further reduces the information content of the structural database that must be extracted to generate a predictive potential. The results are individual knowledge-based potentials tailored to individual families that are capable of predicting binding affinity with high accuracy.

On the basis of jackknifing results, we find the that optimized potentials are capable of predictive errors between 1.1 and 1.3 log K_i units. As expected, this accuracy surpasses the 1.5 log K_i unit accuracy found within the diverse database. First, we examined the aspartic protease family. In this example, we found an average unsigned error of only 1.1 log K_i units corresponding to a predictive accuracy of 1.5 kcal/mol. The minimum unsigned error corresponds to a property weight $w = 0.61$, meaning that atom types are described by a combination of the partial charge (scaled by 0.61) and the Born radius (scaled by $1.0 - 0.61 = 0.39$). The resulting potential is nonspecific resulting in six protein atom types and a single ligand atom type. Given the KB potential's nonspecific interactions with the ligand atoms, a van der Waals or self-polarization solvation effect may be described within the potential. Using these optimized atom types, the resulting KB potential correlates strongly with the entire aspartic protease family database with a coefficient of $R = 0.80$. This compares favorably with the SMoG 2001 potential, which demonstrates a correlation of $R = 0.62$ within the aspartic protease family.

In addition, the metallo protease family and the sugar binding protein family were used to generate optimized atom types. In the 22 member metallo protease family, the corresponding optimized KB potential resulted in an average unsigned error of 1.1 log K_i units or a predictive accuracy of 1.5 kcal/mol. On the basis of the optimized atom types determined from multiple jackknifing runs, the corresponding KB potential results in a correlation coefficient of $R = 0.88$. This is comparable to the correlation coefficient of $R = 0.76$ found in a family of 15 metallo proteases analyzed in PMF.[22]

3. Atom Types Optimized for Specific Target Families. Finally, we examined potentials optimized on the sugar binding protein family. In the 14 member sugar binding protein family, the corresponding optimized KB potential resulted in an average unsigned error of 1.3 log K_i units or a predictive accuracy of 1.8 kcal/mol. What is interesting about this optimization is the strong error encountered for sugar binding proteins when the atomic descriptors are dominated by partial charge. In the aspartic protease family, the lowest predictive error was found using atom types based only on partial charges. It is clear from Figure 8a that sugar binding proteins prefer a distinctly different atom typing and corresponding KB potential. The difference observed in optimal atom types in all three of the families shown here demonstrates that a single generic KB potential to describe all complexes optimally remains elusive. The small predictive errors in all of the families suggest that optimized KB potentials for individual families may have a role in the lead discovery process.

Summary. The development of robust KB potentials relies upon the further development of the underlying physical theory. This is true even, and maybe even particularly, when these explorations reveal fundamental limitations. Are KB potentials simply scoring functions, or can they be linked to a real physical quantity? Theoretical explorations suggest the former, while empirical studies on model systems suggest the latter.

Whatever the underlying theory, KB potentials describe energetics through an examination of local density surrounding a set of predefined particle types. If these particle types do not map to the true or optimal

effective pairwise potential, crucial information contained within the structural database cannot be extracted. The atom typing procedure described in this paper attempts to move closer to this correct mapping.

Conclusions

While it has been shown that pair potentials may be extracted from contact statistics within "toy" structural databases, these approaches have not always met with the greatest success when applied to the PDB. To take advantage of the results obtained in these model studies, we introduce a new method for identifying particles consistent with the true potential. We show that in extreme cases atom types unrelated to the true potential are incapable of extracting any useful information from the structural database. The method that we propose for constructing native atom types involves clustering atoms in the training database according to parameters that influence pair potentials in classical empirical force fields.

Further exploration of this model is planned, including the explicit incorporation of more physiochemical properties into the atom type descriptions. It has been mentioned that the Born radius has been used in certain models to approximate hydrophobic effects and could also be considered to capture some of this information within our model as well. However, it may be possible to include the hydrophobic and other effects more directly by including in the atom descriptors properties coupled to the hydrophobic and other effects. These higher dimensional descriptors could include parameters describing Lennard−Jones attraction as well as field effect polarization.

In the current work, we have shown that knowledge-based potentials constructed on the basis of atom types grouped by the Born radius and partial charge can improve agreement with experimentally determined binding affinities. Specifically, the incorporation of native atom types into the SMoG program for *ab initio* drug design shows an improvement in the knowledge-based potential's ability to predict relative binding affinity within a diverse set of protein/ligand complexes. Furthermore, inclusion of the Born radius as an atomic descriptor yields potentials capable of accurately describing binding affinity within a variety of protein families. This improvement in the intrafamily correlation was not possible using atom types based only on local partial charge information. Finally, in optimizing potentials for individual families, we found that it was possible to very accurately predict binding free energies within these families. These models have strong implications for rapid drug screening or possibly *ab initio* design methodologies.

Theory and Methods

Theory. The purpose of this work is to extend the lessons learned from self-consistent studies on knowledge-based potentials and develop an algorithm consistent with these past studies and also applicable to real chemical systems. To accomplish this, it is helpful to first outline the results from these self-consistent studies as well as the effect of inconsistent atom types.

It has been described that contact statistics extracted from dense thermodynamic ensembles reveal information relevant to the potential of mean force between a pair of particles and not the simple pair potential.[26,27] It is clear, however, from self-consistent studies that it is possible to extract quantities strongly correlated with the true pair potential from contact statistics. A more rigorous statistical mechanical explanation of this observation is a subject that will be explored more completely in a future paper. For the purposes of this work, it will suffice simply to recognize the empirical observation that quantities related to pair potentials can be extracted in model systems.

$$U_{ij}^{\text{true}} = a \cdot \ln[p_{ij}(r)] + b$$

where U_{ij}^{true} refers to the true pair potential energy between particles of type i and j and $p_{ij}(r)$ refers to the fraction of observations where particles of type i and j are at a distance of r.

It should be noted that in these self-consistent studies the potential is explicitly defined as a pairwise potential with respect to a set of predefined particle types. We describe these particle types as native to the predefined pairwise potential. This brings up two key challenges in transferring the model system results to real chemical systems. First, in a real chemical system, the true potential is not pairwise. However, we assume that there exists a pairwise potential and a corresponding set of particle types that describes the true multibody potential with a minimal loss of information. This effective pair potential may be described as the optimal projection of the true multibody potential. To extract this effective pairwise potential from a real chemical system, the native particle types must be defined in order to determine the relevant contact statistics. This brings us to the second key challenge, namely, how to define the particles themselves.

As already described, a correspondence between the particle type definitions and a true pairwise potential is built into the self-consistent studies; however, these "native" particle types are not known in real chemical systems. In real chemical systems, the environment surrounding particles or atoms i and j may influence the effective potential felt between the atoms. One of the biggest influences may be electrostatic solvation effects. If atoms i and j represent atoms of opposite charge, they will certainly feel an attractive potential. However, the magnitude of this potential will be mediated by the degree of burial or separation from the aqueous solvent. This is one example of how under an incompatible set of atom types, the extracted energy may be a function of a mixture of pair frequencies based on the true or native atom types.

$$U_{mn} = a \cdot \ln[p_{ij}(r) + p_{kl}(r)] + b$$

where m and n represent non-native atom types typically chosen in knowledge-based applications and i, j, k, and l represent native atom types consistent with the optimal pair potential. As described previously, the optimal pair potential is pair potential, which is a projection of the actual multibody potential, that minimizes the loss of information.

By investigating an extreme case, we can further illustrate the consequences of mistyping atoms. The extreme case would involve atom types that are com-

pletely uncorrelated with the native atom types. In this situation, the pair energies are a function of a random mixture of the pair frequencies based on the native atom types. In this scenario, the contact frequencies between different atom types will reflect the bulk density of the system (not the local densities as do frequencies based on native atom types) and will be the same for all pairs of atom types. In the limit of an infinite training database, interaction energies between the particles will be identical and will therefore contain no information relevant to atom−atom potentials.

Regardless of the specific reference state or other details of the potential, the extracted energies are reflective of bulk density and not local density and therefore have lost a significant amount of information contained within the true potential. This can be illustrated by looking at the effect of averaging a variety of radial distribution functions for fluids (Figure 13). In this scenario, the averaged radial distribution function contains information on the excluded volume of atoms and the bulk density of the system; however, the crucial local density information has been averaged out. In the case of binding energy prediction, it is obvious that energies obtained using random atom types should not correlate with the true potential or the experimental binding affinities.

To test this assumption, we have assigned atom types randomly within the SMoG training and testing structural databases. A knowledge-based potential was generated using the same approach described in a previous paper.[11] Using this procedure, contact frequencies between protein and ligand atom types are obtained from a training database of protein−ligand complexes corresponding to two distance bins of 0.0−3.5 Å and 3.5−4.5 Å. By evaluating the contact statistics and the relative concentrations of the various atom types, one can obtain an association energy for each of these two bins. The total interaction energy is the sum of the energies obtained from these two distances. The energy expression is shown below.

$$E_{ij} = |\ln\left(\frac{N_{ij}}{N_{\text{total}}|X_i^{0.9}|X_j^{0.9}}\right)$$

where N_{ij} is the number of contacts found between protein atom type i and ligand atom type j, N_{total} is the total number of contacts within the structural training database, and X_i is the mole fraction of atom type i.

The number of atom types was varied between 1 and 15 for both the protein and the ligand and the combination producing the largest correlation with experimental binding affinities was selected. This procedure was repeated multiple times producing a variety of randomly assigned atom types and corresponding optimal knowledge-based potentials. The distribution of correlation coefficients obtained for these knowledge-based potentials was centered close to 0 (Figure 14). As expected, random atom types do not yield potentials that correlate with the experimental binding affinity.

While random atom types are ineffective at producing useful knowledge-based potentials, it is clear that atom types based on "chemical intuition" such as invoked in SMoG 2001 are more effective. One reason for this is that these atom types are more compatible with the native atom types corresponding to the optimal pair potential.

While in theory there may exist an optimal set of atom types for knowledge-based potentials, in practice it is unclear how to identify this ideal set. However, on the basis of the framework described in this section, an algorithm has been developed for optimizing atom types within knowledge-based potentials. In the following section, we describe an algorithm for optimizing atom types within the SMoG potential for binding energy prediction. These atom types are based on physical atomic properties known to be significant in determining effective pair potentials. Furthermore, a small number of fitting parameters are used within the model to account for components that are unknown a priori. This results in a model that is a hybrid between the successful, but narrowly applicable, QSAR approaches and the robust, but less accurate, knowledge-based potentials.

Methods. The foundation of the work in this paper is the automated drug design program SMoG developed in this laboratory.[12] Significant changes have been made to the program in order to incorporate native atom types. The OpenBabel molecular data structure library[28] has been linked into SMoG, allowing for rapid calculation of atomic properties including Gasteiger partial charges[29] used to characterize atoms in this study.

Two structural databases are involved in the development of the SMoG knowledge-based potential: a training database used to construct the knowledge-based potential based on contact statistics and a testing database containing associated binding affinities used to optimize the atom types and evaluate the resulting potential. The training database consists of 250 protein/ligand crystal complexes. Previous studies of the SMoG potential have shown that the potential converges using a relatively small training database. Furthermore, optimized correlation coefficients obtained from the potentials used in this study are invariant to the use of a training database of 690 complexes. The PDB codes identifying these structures are as follows: 148l, 181l, 182l, 183l, 184l, 185l, 186l, 187l, 188l, 1aam, 1aaq, 1aaw, 1abi, 1abj, 1abo, 1abr, 1acj, 1ack, 1acl, 1acy, 1add, 1ads, 1aec, 1agp, 1aha, 1ahb, 1aht, 1aia, 1aib, 1aic, 1akb, 1akc, 1ama, 1amq, 1amr, 1ams, 1apg, 1apm, 1apt, 1apu, 1apv, 1apw, 1arc, 1ars, 1asa, 1asb, 1asc, 1asd, 1ase, 1asf, 1asg, 1asl, 1asm, 1asn, 1at1, 1atn, 1avd, 1aya, 1ayb, 1ayc, 1azm, 1bac, 1baf, 1bbr, 1bcr, 1bcs, 1bdm, 1bib, 1blc, 1blh, 1bll, 1bma, 1bmd, 1bra, 1btc, 1byc, 1byd, 1bzm, 1cbr, 1cbs, 1ccg, 1cde, 1chb, 1cka, 1ckb, 1cla, 1cme, 1cnb, 1cne, 1cnf, 1com, 1cp4, 1cpd, 1cpf, 1cpi, 1cqh, 1crb, 1cts, 1ctt, 1ctu, 1cwa, 1cwb, 1cwc, 1cxa, 1cyn, 1dbb, 1dbj, 1dbk, 1dbm, 1dcc, 1dgd, 1dge, 1dhf, 1dhi, 1dhj, 1dhr, 1die, 1dit, 1dld, 1dlr, 1dls, 1doc, 1dod, 1dog, 1dpp, 1dr1, 1dr2, 1dr3, 1dr6, 1dr7, 1drf, 1drh, 1dtp, 1dwb, 1dwc, 1dwd, 1dwe, 1dyh, 1dyi, 1dyj, 1ead, 1eae, 1eaf, 1eap, 1eas, 1eat, 1eau, 1ebd, 1eed, 1eft, 1ela, 1elb, 1elc, 1eld, 1ele, 1elf, 1elg, 1els, 1emy, 1ent, 1epb, 1epl, 1epm, 1epn, 1epo, 1epp, 1epq, 1epr, 1esb, 1eta, 1etb, 1etr, 1ets, 1ett, 1etu, 1fbc, 1fbf, 1fbg, 1fbp, 1fc2, 1fkb, 1fkg, 1fki, 1fmp, 1fnd, 1fpb, 1fpt, 1frg, 1frt, 1fut, 1gbb, 1gbc, 1gbd, 1gbf, 1gbh, 1gbi, 1gbk, 1gbl, 1gbm, 1gec, 1ger, 1get, 1gfi, 1ggi, 1ghb, 1gil, 1gky, 1gla,

1glb, 1glc, 1gld, 1gle, 1glp, 1glq, 1gmh, 1gne, 1gnp, 1gnq, 1gnr, 1gpy, 1gra, 1gre, 1grf, 1grg, 1gsq, 1gst, 1hag, 1hah, 1hai, 1hbt, 1hbv, 1hcs, 1hdc, 1hdt, 1hef, 1heg, 1hew, 1hgt, 1hhg, 1hhh, 1hhi, 1hhj, 1hhk, 1hih, 1hii, 1him, 1hin, 1hlp, 1hlt, and 1hmr.

The testing database consists of 118 protein/ligand crystal complexes for which experimental binding affinities are known. The 118 testing complexes are further characterized as eight target families including: serine proteases (20 complexes), aspartic proteases (18 complexes), metallo proteins (22 complexes), carbonic anhydrase (19 complexes), sugar binding proteins (14 complexes), endothiapepsin (11 complexes), purine nucleoside phosphatase (five complexes), and other proteins (nine complexes). The PDB codes and binding affinities describing the testing set are given in Ishchenko et al.[11] The structure 4dfr contained within the other proteins family was excluded from this analysis.

The properties chosen to describe the atoms with this model, providing the basis for atom typing, are the Born radius and the Gasteiger partial charge. The Born radius is an empirical parameter used in a variety of models to primarily describe electrostatic solvation effects. It effectively describes the average distance between the center of a given atom and the solvent boundary. This measure of "buriedness" has also been used in certain applications to describe hydrophobic solvation as well as electrostatic effects. The partial charge is meant to capture the electrostatic distinctiveness of atoms as characterized by its direct interaction with other atoms through Coulomb forces. By grouping atoms by partial charge and Born radius, we hope to group the real interactions between these atoms in protein−ligand complexes. While it appears that we are focusing only on electrostatic interactions when describing atoms and atom types, this is not necessarily the case. The explicit grouping of atoms by the partial charge and Born radius may also capture some information relevant to hydrophobic or van der Waals interactions. We also note that the quantitative accuracy of these empirical descriptors is not necessary in our model. These descriptors are used solely for clustering atoms together with similar physiochemical properties. The interaction energies themselves are determined through a traditional knowledge-based approach.

The development of the native atom type knowledge-based potentials can be described as follows. First, each atom within the training database is assigned a partial charge and Born radius. The partial charges are determined within the OpenBabel library using the Gasteiger method. Born radii are determined using the approach described by Hawkins and Truhlar.[30] Each atom is then described by these two parameters. The partial charge of each atom is then multiplied by a weight w while the Born radius is multiplied by a weight $1 - w$. This determines the relative importance of the two parameters as w is incremented from 0 to 1.

These points in a two-dimensional descriptor space are clustered using a *K*-means algorithm.[31] N clusters are generated by first choosing N atoms randomly from the training database. The *K*-means algorithm then generates clusters by iteratively grouping the remaining points to the closest cluster center. This is repeated multiple times by choosing N new random atoms, and the resulting clusters with the minimum total variance are chosen as the optimal clustering of the training set into N clusters. One to 15 clusters are generated for protein and ligand atoms separately resulting in a compromise between the information content in the clusters (which rises with the number of clusters) and the precision of the resulting knowledge-based potential (which decreases as the number of clusters or atom types increases). This results in a total of $15 \times 15 = 225$ knowledge-based potentials for each choice of weighting parameter w, constructed using the procedure described by Ishchenko et al.[11] The number of protein and ligand atom types was limited to a maximum of 15 in order to preserve a reasonable level of precision in the resulting statistical potentials.

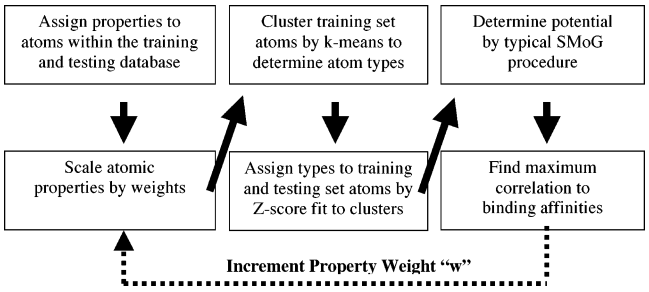
The resulting 225 potentials are used to evaluate the interaction energies between the protein and the ligand atoms within the testing structural database. The Pearson correlation between these predicted binding affinities and the experimental binding affinities (log K_i) is used to evaluate the success of the given SMoG potential. For each weight, the optimal potential was identified as the one potential of the 225 that produced the largest Pearson correlation with the experimental binding data. The optimal correlation as a function of the weighting parameter w is shown for a variety of testing sets.

K-means clustering is a fast and simple clustering algorithm requiring only $O(n)$ time to complete where n is the number of elements or atoms that are being clustered. One sacrifice that must be made when using *K*-means is determinism. The *K*-means algorithm is inherently stochastic and therefore can yield different local clustering solutions when operating on the same dataset. While the speed and simplicity of *K*-means is desirable, the randomness of the resulting clusters is not. Therefore, to improve the robustness of the *K*-means clustering, we incorporate a small alteration to the algorithm. Each time the n atoms within the training database are clustered by *K*-means, the algorithm is actually run 10 times using different random cluster seeds. At the end of each of the 10 independent clustering runs, the sum of the variances of each cluster is determined. The variance of each cluster is defined as the sum of the variances of each cluster dimension. The run producing the minimum total variance is chosen as the optimal *K*-means clustering of the n training set atoms. This simple iterative approach improves the ability of the clustering algorithm to locate a global optimum rather than settling for the first local solution.

While clustering using *K*-means is fast enough to be practical for generating optimized KB potentials, future applications of the potential evaluating new testing complexes or even growing new inhibitors require a faster means of assigning atom types. To avoid the time-consuming clustering algorithm while assigning atom types to the testing database, atoms are assigned to a cluster based on the minimum *Z*-score between the new atom and each cluster. The *Z*-score is the number of SDs between the two-dimensional property vector for the atom in question and each cluster center is computed. The cluster resulting in the minimum number of SDs

is assumed to describe the atom type for the atom in question.

The minimum *Z*-score approach has been tested on the training database (for which clustering is explicitly performed), and it successfully predicts the cluster assigned to a given atom with over 99% accuracy. This is due to the fact that the clusters are relatively tight (have small variances) and do not significantly overlap with one another. To maintain consistency between the testing database and the training database, atom types are assigned to the atoms within both databases using the minimum *Z*-score procedure. A flowchart describing this procedure is given below.



Note Added in Proof. The *Z*-score distance function utilized in this study to assign atom types was given as $d = |\sum_i^{N\text{-dim}} Zscore_i|$. An alternative "taxi-cab" distance function, $d = \sum_i^{N\text{-dim}} |Zscore_i|$, was also evaluated and similar results were obtained. These results are shown in the supplementary information available at http://www-shakh.harvard.edu/publications.html.

References

(1) Rojnuckarin, A.; Subramaniam, S. Knowledge-based interaction potentials for proteins. *Proteins* **1999**, *36*, 54−67.
(2) Zhang, C.; Kim, S. H. Environment-dependent residue contact energies for proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2550−2555.
(3) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(4) Katritzky, A. R.; Fara, D. C.; Petrukhin, R. O.; Tatham, D. B.; Maran, U.; et al. The present utility and future potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors. *Curr. Top. Med. Chem.* **2002**, *2*, 1333−1356.
(5) Kellogg, G. E.; Semus, S. F. 3D QSAR in modern drug design. *Exs* **2003**, 223−241.
(6) Selassie, C. D.; Mekapati, S. B.; Verma, R. P. QSAR: Then and now. *Curr. Top. Med. Chem.* **2002**, *2*, 1357−1379.
(7) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.
(8) Hansson, T.; Marelius, J.; Aqvist, J. Ligand binding affinity prediction by linear interaction energy methods. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 27−35.
(9) Beveridge, D. L.; DiCapua, F. M. Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431−492.
(10) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP−Potential of mean force describing protein−ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.
(11) Ishchenko, A. V.; Shakhnovich, E. I. Small molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein−ligand interactions. *J. Med. Chem.* **2002**, *45*, 2770−2780.
(12) DeWitte, R. S.; Shakhnovich, E. I. SMoG: De novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.
(13) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(14) Gohlke, H.; Klebe, G. Statistical potentials and scoring functions applied to protein−ligand binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231−235.
(15) Ben-Naim, A. Statistical potentials extracted from protein structures: Are these meaningful potentials? *J. Chem. Phys.* **1997**, *107*, 3698−3706.
(16) Thomas, P. D.; Dill, K. A. Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* **1996**, *257*, 457−469.
(17) Mirny, L. A.; Shakhnovich, E. I. How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* **1996**, *264*, 1164−1179.
(18) Zhang, L.; Skolnick, J. How do potentials derived from structural databases relate to "true" potentials? *Protein Sci.* **1998**, *7*, 112−122.
(19) Shimada, J.; Ishchenko, A. V.; Shakhnovich, E. I. Analysis of knowledge-based protein−ligand potentials using a self-consistent method. *Protein Sci.* **2000**, *9*, 765−775.
(20) Bahar, I.; Jernigan, R. L. Interresidue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.* **1997**, *266*, 195−214.
(21) Wang, W.; Wang, J.; Kollman, P. A. What determines the van der Waals coefficient beta in the LIE (linear interaction energy) method to estimate binding free energies using molecular dynamics simulations? *Proteins* **1999**, *34*, 395−402.
(22) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418−425.
(23) Reith, D.; Huber, T.; Muller-Plathe, F.; Torda, A. E. Free energy approximations in simple lattice proteins. *J. Chem. Phys.* **2001**, *114*, 4998−5005.
(24) Sippl, M. J.; Ortner, M.; Jaritz, M.; Lackner, P.; Flockner, H. Helmholtz free energies of atom pair interactions in proteins. *Fold Des.* **1996**, *1*, 289−298.
(25) Finkelstein, A. V.; Badretdinov, A.; Gutin, A. M. Why do protein architectures have Boltzmann-like statistics? *Proteins* **1995**, *23*, 142−150.
(26) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Sausalito, CA, 2000; Vol. xii, 641 p.
(27) Sippl, M. J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **1990**, *213*, 859−883.
(28) OpenBabel, 2001.
(29) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity−A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3288.
(30) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **1995**, *246*, 122−129.
(31) Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data*; Prentice Hall: Englewood Cliffs, NJ, 1988; Vol. xiv, 320 p.

JM0498046