JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

One Size Does Not Fit All: The Limits of Structure-Based Models in Drug Discovery

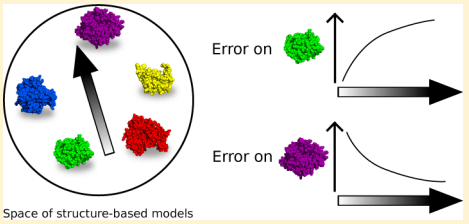
Gregory A. Ross,[†] Garrett M. Morris,[‡,§] and Philip C. Biggin*[,†]

[†]Structural Bioinformatics and Computational Biochemistry, Department of Biochemistry, University of Oxford, South Parks Road, Oxford, Oxfordshire OX1 3QU, United Kingdom

[‡]InhibOx, Ltd., Oxford Centre For Innovation, New Road, Oxford, Oxfordshire OX1 1BY, United Kingdom

Supporting Information

ABSTRACT: A major goal in computational chemistry has been to discover the set of rules that can accurately predict the binding affinity of any protein-drug complex, using only a single snapshot of its three-dimensional structure. Despite the continual development of structure-based models, predictive accuracy remains low, and the fundamental factors that inhibit the inference of all-encompassing rules have yet to be fully explored. Using statistical learning theory and information theory, here we prove that even the very best generalized structure-based model is inherently limited in its accuracy, and protein-specific models are always likely to be better. Our results refute the prevailing assumption that large data sets and advanced machine learning techniques will yield accurate, universally applicable models. We anticipate that the results will aid the development of more robust virtual screening strategies and scoring function error estimations.



INTRODUCTION

The accurate prediction of protein−ligand affinity remains one of the great challenges in computational chemistry.[1] A fast and generally applicable method would greatly benefit the pharmaceutical industry by speeding up the discovery of new drugs and reducing reliance on expensive wet lab experiments.

In principle, one can calculate the binding affinity of any protein−ligand complex using molecular simulations and rigorous statistical mechanics techniques.[2,3] These methods have been successfully applied in lead-optimization, especially when calculating the relative affinities of congeneric ligands.[4,5] Despite their general applicability, these techniques remain too time-consuming to assay the vast chemical libraries used in pharmaceutical research. As a result, very rapid models, called scoring functions, are typically used in virtual screens. Scoring functions predict affinity using only a single snapshot of a molecular complex, so that the objective of scoring function research is to elucidate any emergent laws that operate above the physical laws of atomic motion. They are designed to be applicable to all proteins and ligands, in contrast to quantitative structure−activity relationship (QSAR) models. Unlike a QSAR model, the ideal scoring function would not need to be optimized for any particular protein or a set of ligands and could accurately predict hits early in a drug discovery project.

Despite decades of development, however, the performance of scoring functions varies greatly between different protein systems, and predictions often correlate poorly with experimental data.[6−9] The cause of scoring function error remains as yet uncertain and has been attributed to a number of different factors. Physical arguments, for instance, have highlighted the poor treatment of explicit solvent effects and protein−ligand flexibility.[10−12] On the other hand, recent efforts have sought to improve scoring functions by empirical means, employing advanced machine learning techniques, many protein−ligand interaction descriptors, and large training sets.[13−16] However, it is still not clear whether such approaches significantly improve accuracy.[8] A technique known as 'consensus scoring' has been been shown to reliably lower error by combining the predictions of different scoring functions, although accuracy still remains far below that of experiment.[7,17−19]

Given the potential impact on drug discovery and the substantial effort in scoring function development, it is vital to understand the fundamental uncertainties in these rapid affinity models. Recently, Faver et al. investigated the systematic and random errors associated with the interaction energies of protein−ligand complexes.[20] Using a fragment based approach,[21] they found that the random error in electrostatic interaction calculations rises with the system size. Protein targeted scoring functions have also been observed to be more accurate than generalized models,[22] challenging the utility of a universal scoring function, yet questions still remain with regards to the limits of scoring function accuracy. For instance, is it possible for a particular set of descriptors and functional form of a model to achieve a negligible error? Also, can a universally applicable scoring function ever be better than a targeted one? Without a formal analysis of the structure-based modeling process, questions such as these cannot be answered fully.

In this work, we investigate the inherent uncertainties in empirical structure-based models with a rigorous mathematical analysis utilizing statistical learning theory and information

Received: May 22, 2013
Published: August 5, 2013

   dx.doi.org/10.1021/ct4004228 | J. Chem. Theory Comput. 2013, 9, 4266−4274

theory. As statistical learning theory is implicitly applied whenever one trains a model using regression or classification, we establish the statistical relationship between a structural snapshot of a protein−ligand complex and its affinity. By doing so, we find that a critical assumption of statistical learning theory is often violated when training and applying empirical structure-based models. An implication of this is that a scoring function that performs optimally on one set of protein−ligand complexes necessarily performs poorly on another set. Also, we prove that even the very best universal structure-based model is significantly limited in its accuracy, and protein-specific models are always likely to be better. Throughout, we use an information theoretic measure to quantify scoring function error to ensure that our analysis is independent of how the protein−ligand interactions are modeled, so that our results apply to any set of descriptors and regression method. We verify our theoretical predictions with our own scoring models.

THEORY

To understand empirical scoring function error we must first appeal to the theory that underpins regression analysis: statistical learning theory.[23] We denote the structural features of a protein−ligand complex as x and the binding affinity as y. Statistical learning theory formalizes the process of elucidating the functional relationship between x and y by assuming there is a probabilistic process that generates the data used to train and test a model. We denote the probability distribution function (PDF) over x and y as $p(x,y)$. The functional relationship between structure and affinity is encoded in the conditional PDF $p(y|x)$, as with this PDF, one can find the most likely binding affinity for a given structural description of a protein−ligand complex. A further, critical assumption in statistical learning theory is that the data used to train a model are generated by the same probabilistic process as the data found in the predictive setting.

The quantitative performance of a model is assessed using a loss function. A popular measure, for instance, is the mean squared error between the affinity predictions of a model and their experimental values. For our loss function, we utilize the cross (Shannon) entropy between the true and modeled probability distributions. The cross entropy, denoted $C(Y|X)$, is given by

$$C(Y|X) = -\int\int p(x, y) \ln q(y|x)\, dx\, dy \tag{1}$$

and is a quantification of the uncertainty in the affinity given a set of structural descriptors sampled from $p(x,y)$, and a model, denoted $q(y|x)$. With this loss function, the task in scoring function development is to find the model that lowers the uncertainty as much as possible. Cross entropy is a very general loss function,[24,25] and while commonly used in classification,[23] judicious selection of specific forms of $q(y|x)$ allows for more familiar measures of regression error. Indeed, if $q(y|x)$ encodes for a scoring function that minimizes $C(Y|X)$, then the same scoring function also minimizes the mean squared error (see Supporting Information Section 2.1). In a well-known information theoretic result,[26] cross entropy expands to

$$C(Y|X) = h(Y|X) + D(p(y|x)||q(y|x)) \tag{2}$$

where $h(Y|X)$ is the conditional entropy of the true distribution, given by

$$h(Y|X) = -\int\int p(x, y) \ln p(y|x)\, dx\, dy \tag{3}$$

and $D(p(y|x)||q(y|x))$ is the Kullback−Leibler divergence or relative entropy between the true and modeled distributions. It is given by

$$D(p(y|x)||q(y|x)) = \int\int p(x, y) \ln \frac{p(y|x)}{q(y|x)}\, dx\, dy \tag{4}$$

Relative entropy is a convex function for the two conditional distributions; it is zero only when the distributions are the same, and positive otherwise. Equation 2 can be understood as representing the minimum uncertainty, or error, of the binding affinity given the structure, plus the uncertainty due to our model; choosing the wrong model only increases our uncertainty of the system. In our information theoretic perspective, $h(Y|X)$ is the minimum achievable error, which is irreducible for a given set of descriptors, and $D(p(y|x)||q(y|x))$ is the bias for assuming the structure-affinity relationship is encoded in $q(y|x)$ when in reality it is encoded in $p(y|x)$. This bias is a type of random error, as opposed to a systematic error. Any bias ensures that the minimum error is not achieved for the wrong model, and borrowing a term from financial decision theory[27] and signal processing,[28] we shall refer to the positive deviation from the minimum error as *regret*.

Many factors can cloud the reasons for an empirical model's inaccuracy. In structure-based affinity models one has to choose the representation of a complex - a possible description might be the surface complementarity between a drug bound to a protein - and the functional form of the model. Both choices may introduce errors that are difficult to disentangle from the fundamental uncertainties arising from rapid affinity prediction. The benefit of using $C(Y|X)$ to quantify a error is that all the confounding factors that hinder a scoring function's accuracy unravel, and we need only consider the relationship between a protein−ligand's affinity and a complete structure of the complex obtained from solution at equilibrium. First, the minimum error, $h(Y|X)$, is a property of the data and its underlying PDF only, and thus is independent of the functional form of a scoring model. Second, the mere process of selecting a subset of structural features of a protein−ligand complex to use in a model means some data about the complex is discarded. Discarding data can either maintain or increase $h(Y|X)$,[26] so to analyze the minimum achievable error of a typical scoring function, we use x to denote snap-shots of entire complexes of structures in solvent, thereby avoiding any reduction in information. Third, we use conditional PDFs and their corresponding optimal models so as not to consider the variance that occurs in when fitting a scoring function in practice.[23] Thus, we focus on an idealization of a structure-based model, the error of which is a lower bound of what can be achieved in practice.

METHODS

Analysis. Our analysis begins by placing the training and testing of an empirical structure-based scoring function in a statistical framework. Using this framework, we derive an upper bound for the regret of a model that is trained and applied to protein−ligand complexes sampled from different distributions, a common occurrence in virtual screening. We also analyze and compare the bias and minimum error of a universal structure-based model to those of protein-targeted models. Regret due to scoring function optimization is also considered.

Scoring Function Development. We verified our analytic predictions by fitting and testing a scoring function that is archetypal of popular empirical models.[29−32] Described in more detail in the Supporting Information, it comprises of five protein−ligand interaction terms including shape complementarity and electrostatic energy, and the number of rotatable bonds in the ligand. As is typical, the functional form of the scoring function is linear. We use elastic net regularization as our fitting procedure.[33] This method improves model robustness by controlling for the L_1 and L_2 norms of the regression coefficients[23] at the cost of having two free parameters that can be optimized on a validation data set or by using cross-validation. We tested L_1 fractional penalty between 0 and 1 in 0.1 increments, and following Zou and Hastie,[33] we tested L_2 penalties of 0, 0.01, 0.1, 1, 10, and 100. Regression was performed using the statistical programing language R[34] and the 'elasticnet' package.

As our analysis is an idealization of the structure-based modeling process, we trained and tested our canonical scoring function using data sets of protein−ligand complexes with known binding affinities and whose structures had been resolved by X-ray crystallography. We used 10 data sets in total which were constructed using the 24th September 2010 version of the CSAR high quality data set of protein−ligand complexes[9] and the 2011 version of the PDBbind data set.[7] The CSAR data set was used to construct one training set (207 complexes) and one test set (83 complexes) that both consist of a diverse range of protein−ligand complexes. The CSAR data set was provided with a list that contained all of the complexes assembled into groups of 90% protein sequence identity. One complex from each group was selected at random to form the training set. The complexes that did not share 90% sequence identity with any other protein were also selected for the training set. From the complexes that remained, another random selection was made to form the diverse test set, so it too was composed of proteins with less than 90% sequence identity. This test set is labeled as data set A in Table 1 and in the Results section. The PDBbind data set was used to assemble 8 single-protein test sets which are labeled as B−I and are listed in Table S1 of the Supporting Information. Additional complexes were added to the HIV1 protease (labeled A), trypsin (C), carbonic anhydrase II (E), and thrombin (E) data sets, in order to match the single-protein data sets in the 2009 study by Cheng et al.[7] In that study, the predictive performance of thirty-three variants of popular scoring functions (including GOLD, GlideScore and X-Score) was evaluated. This allowed us to test how representative our scoring function is of these popular models. No single complex appeared in more than one data set.

Table 1. Data Sets Used for Training and Testing of the Scoring Functions in Order of Data Set Size

data set	label	# of complexes
diverse training set	-	207
diverse test set	A	83
HIV1 protease	B	108
trypsin	C	66
factor Xa	D	43
carbonic anhydrase II	E	40
PTP	F	38
thrombin	G	36
OppA	H	32
urokinase	I	31

We created a *diverse* scoring function by fitting our canonical model on the training set of diverse protein−ligand complexes optimizing the free parameters to data set A. Our model was also fitted to data sets B−I and optimized using leave-one-out cross-validation to create 8 *single-protein* scoring functions. All of these scoring functions were then applied across the data sets listed in Table 1, and their accuracy was measured using the mean absolute error and the Spearman rank coefficients of the affinity predictions against the experimental values.

RESULTS

Protein−Ligand Structures Have Unique Probability Distributions. We now prove, by construction, that the PDF of a protein−ligand's structure x, and affinity y, is in general different for each complexes. This means that the fundamental condition in regression analysis for the training and test set to be sampled from the same probability distribution is often violated for structure-based scoring functions.

In rigorous free energy calculations,[2,3] a control parameter, denoted λ, is often used to define a molecular system or set of constraints on that system. In ligand binding free energy calculations, the control parameter associated with the bound state, denoted λ_b, is switched over the course of possibly many simulation windows to the value associated with the unbound state, denoted λ_u. As free energy is a state function, the binding affinity y depends only on λ_b and λ_u. Experimental error gives rise to uncertainty in the value of the 'true' free energy difference between the two states. We represent this intrinsic uncertainty in the free energy between the states via the PDF $s(y|\lambda_b,\lambda_u)$.

In contrast to rigorous methods, scoring functions use a single snapshot of the protein−ligand complex to predict affinity. As sampled from the bound state, any such snapshot is dependent only on λ_b. For instance, if the snapshot is sampled from the equilibrium distribution of the complex in solvent, then from statistical mechanics, the PDF of a structure x given λ_b, denoted $r(x|\lambda_b)$, is equal to the Boltzmann distribution

$$r(x|\lambda_b) = \Lambda \exp(\beta F(\lambda_b) - \beta E(x, \lambda_b)) \quad (5)$$

where $F(\lambda_b)$ is the free energy of the bound state, $E(x,\lambda_b)$ is the potential energy of the structural snapshot, Λ is the integral over all momenta, and β is equal to the reciprocal of the Boltzmann constant multiplied by the temperature. As discussed, our information theoretic approach to scoring function error means that x represents a particular snapshot of a protein ligand complex, without any loss of information.

To understand scoring function error, we consider the probabilistic relationship between structure x and affinity y. For a given protein ligand complex θ, we denote the PDF of observing x and y as $p(x,y|\theta)$. The question "what is the binding affinity of a particular protein-ligand complex θ?" is implicitly asking "what is the free energy difference between states defined by λ_b and λ_u?". In other words, θ is really a surrogate label for λ_b and λ_u. From the above discussion, x and y are dependent on λ_b and λ_u and not on each other. Thus, x and y are conditionally independent for a given complex, so that

$$\begin{aligned} p(x, y|\theta) &= p(x, y|\lambda_b, \lambda_u) \\ &= r(x|\lambda_b)s(y|\lambda_b, \lambda_u) \end{aligned} \quad (6)$$

Critically, by sampling the structure x from the equilibrium distribution as in eq 5, it is apparent that each $p(x,y|\theta)$ is *unique* for a particular θ. This follows by acknowledging that altering either the protein or the ligand will change the Boltzmann distribution. Therefore, as $r(x|\lambda_b)$ is not the same for different protein−ligand complexes, neither is $p(x,y|\theta)$.

The Transferability of Structure-Based Models. We have shown that the PDF of structures and affinities for a particular protein−ligand complex, $p(x,y|\theta)$, is different for each molecular pair. Yet structure-based scoring functions are fitted and applied to many different protein−ligand complexes. By using regression to train a model, one implicitly assumes that the complexes in a data set set have been sampled in a probabilistic manner. We denote the probability for selecting a complex θ for a particular data set as $\alpha(\theta)$. A scoring function is not trained on the complexes themselves, but on their structures and affinities sampled from the weighted sum

$$p_\alpha(x, y) = \sum_\theta p(x, y|\theta)\alpha(\theta) \tag{7}$$

where the sum is over all protein−ligand complexes, and we have made the dependency of $p_\alpha(x,y)$ on $\alpha(\theta)$ explicit with a subscript. The optimal scoring function for complexes sampled from $\alpha(\theta)$ is encoded in the conditional PDF

$$p_\alpha(y|x) = \frac{p_\alpha(x, y)}{p_\alpha(x)} = \frac{\sum_\theta p(y, x|\theta)\alpha(\theta)}{\sum_\theta p(x|\theta)\alpha(\theta)} \tag{8}$$

By inserting the PDFs defined in eqs 6 and 5 into the above, it is apparent that $p_\alpha(y|x)$ contains contributions from all the individual Boltzmann distributions of each complex, so that different protein−ligand sampling probabilities will, in general, result in distinct optimal models.

Using the above definitions, we now investigate the error of a scoring function that occurs when it is applied to a set of complexes sampled from $\alpha(\theta)$ but is trained on complexes sampled from a *different* distribution, denoted as $\beta(\theta)$. Following the Theory section, the bias incurred by applying the scoring function that is optimal on data sampled from $\beta(\theta)$, which is encoded in $p_\beta(y|x)$, to complexes sampled from $\alpha(\theta)$ is given by $D(p_\alpha(y|x)||p_\beta(y|x))$. Our first main result (proven in Supporting Information Section 3.1) is that

$$D(p_\alpha(y|x)||p_\beta(y|x)) \le D(\alpha(\theta)||\beta(\theta)) \tag{9}$$

Thus, the relative entropy between the complex selection probabilities is an upper bound to the bias of a misapplied scoring function. Scoring function bias is minimized when $\alpha(\theta) = \beta(\theta)$, so that the protein−ligand complexes in the training and test sets of a scoring function have been sampled from the same probability mass function. However, if $\alpha(\theta)$ and $\beta(\theta)$ do not overlap, then $D(\alpha(\theta)||\beta(\theta))$ is unbounded, implying that scoring function error can be *arbitrarily* large. This can occur when the probability of finding any complex from the training set in the test set is zero, such as when a protein-specific QSAR model is applied to another protein.

A universally applicable structure-based model should, by definition, have the lowest possible error when applied to the widest conceivable range of protein−ligand complexes. If a training set is composed of a diverse range of proteins and ligands, we know from eq 9 that for a scoring function's bias to be zero, the test set should be similarly diverse. In a typical virtual screen, however, scoring functions that have been trained on a diverse range of protein−ligand complexes are applied only to ligands binding to a *single* protein, implying different complex selection probabilities between the training and test sets and a potentially large bias.

The Errors of Generalized Structure-Based Models. To investigate the error of a universal scoring function, we consider N different proteins that one can use to construct data sets of protein−ligand structures x and affinities y. Protein-specific complexes are sampled according to $\alpha_i(\theta)$, i = 1,2,3,...,N, with corresponding joint PDFs $p_i(x,y)$. Many samples from $p_i(x,y)$ results in a data set of bound structures and affinities of different ligands bound to protein i. To model the creation of a data set composed of a diverse range of protein−ligand complexes, we select which $p_i(x,y)$ to sample from with probability ω_i. Similar to eq 6, the appropriate joint PDF for this diverse scoring function is given by the weighted sum

$$p_\omega(x, y) = \sum_i \omega_i p_i(x, y) = \sum_i \omega_i \sum_\theta p(x, y|\theta)\alpha_i(\theta) \tag{10}$$

The corresponding conditional PDF, $p_\omega(y|x)$, encodes for the optimal diverse scoring function. If N is sufficiently large and ω suitably broad, then $p_\omega(y|x)$ represents a 'universally' applicable structure-based model. However, it is important to note that this generalized optimality is defined only for its particular sampling probabilities.

From the previous section, we know that a scoring function that has been designed for a diverse range of protein−ligand complexes will have a nonzero regret when applied to a protein-specific data set. A relevant question is, therefore, how large is the average bias incurred by applying a scoring function defined from $p_\omega(y|x)$ to complexes sampled from $\alpha_i(\theta)$. Our second main result (see Supporting Information Section 3.2 for the proof) is that for an arbitrary scoring function encoded by $q(y|x)$

$$\sum_i^N \omega_i D(p_i(y|x)||q(y|x)) \ge \sum_i^N \omega_i D(p_i(y|x)||p_\omega(y|x)) \tag{11}$$

meaning the generalized model has the lowest average bias over all the protein−ligand complexes sampled by ω. Although the universal model $p_\omega(y|x)$ is optimal over the all N proteins, we now show, perhaps counterintuitively, that it is not optimal for each specific protein. In our information theoretic perspective (see eq 2), the minimum error of the general model $p_\omega(y|x)$ over the N proteins is given by its conditional Shannon entropy, denoted $h_\omega(Y|X)$. Also, the minimum achievable error of an optimal protein-specific model $p_i(y|x)$ on protein i is denoted $h_i(Y|X)$. Our third main result (see Supporting Information Section 3.3 for the proof) is that

$$\sum_i^N \omega_i h_i(Y|X) < h_\omega(Y|X) \tag{12}$$

Hence, the minimum error of a generalized structure-based model is greater than the average minimum error for protein specific models. Thus, on average, a scoring function targeted for a specific protein will outperform a scoring that has been

dx.doi.org/10.1021/ct4004228 | J. Chem. Theory Comput. 2013, 9, 4266−4274

designed for a diverse range of protein−ligand complexes. This result explains previously reported studies.[22] Eq 12 also shows that a universal model has a broader error distribution than the average single-protein model.

For a given model, there is an unavoidable trade-off between accuracy over a broad spectrum of complexes and accuracy for certain individual cases. It is important to highlight that the generalized PDF, $p_\omega(y|x)$, and the PDF for a specific data set $p_i(y|x)$ are, in general, different. As these conditional distributions encode their own optimal functional relationships between structure and affinity, the model that best predicts the binding affinities for many proteins may be very different from the best model for specific proteins.

The Optimization of Scoring Functions. The previous section showed that regret is integral to a generalized empirical scoring function. There is a similar trade-off in accuracy when optimizing a scoring function for any set of complexes.

As described, the relative entropy quantifies the error cost of misapplying a scoring function to a data set. This cost is a bias, and as it ensures that the minimum error is not achieved, it contributes to the regret of a model. The bias can be reduced for a particular sampling regime of protein−ligand complexes by refitting a scoring function for those complexes. However, when the regret for those complexes decreases, the regret between the scoring function and another sampling regime may increase. This occurs by virtue of the fact that different complex sampling probabilities results in distinct structure-affinity PDFs and that the relative entropy is a convex function of two distributions. This means that a structure-based model that achieves the lowest possible error on a group of protein−ligand complexes will necessarily perform poorly on another group. This represents a 'no free lunch'[35] scenario for structure-based models: without any a priori information, no structure-based model can be said to outperform another on any particular data set.

In conceptualizing the practical implications of this, it is fruitful to consider two protein-specific conditional PDFs, $p_\alpha(y|x)$ and $p_\beta(y|x)$, as normal distributions with variances approximately equal to σ^2 and corresponding optimal scoring functions $f_\alpha(x)$ and $f_\beta(x)$ respectively. Following from Heskes,[36] we show in the Supporting Information section 2.2 that

$$D(p_\alpha(y|x)||p_\beta(y|x)) \approx \frac{1}{2\sigma^2}\int p_\alpha(x)(f_\alpha(x) - f_\beta(x))^2 \, dx \tag{13}$$

This equation implies that we can represent the theoretically optimal scoring functions $f_\alpha(x)$ and $f_\beta(x)$ as being embedded in a space that preserves the mean squared distances between them. Optimizing a scoring function to a particular protein can then be considered as moving through this space toward the optimal model, increasing the distance from another model which is optimal on a different protein. A schematic diagram of this hypothetical scoring function space is shown in Figure 1.

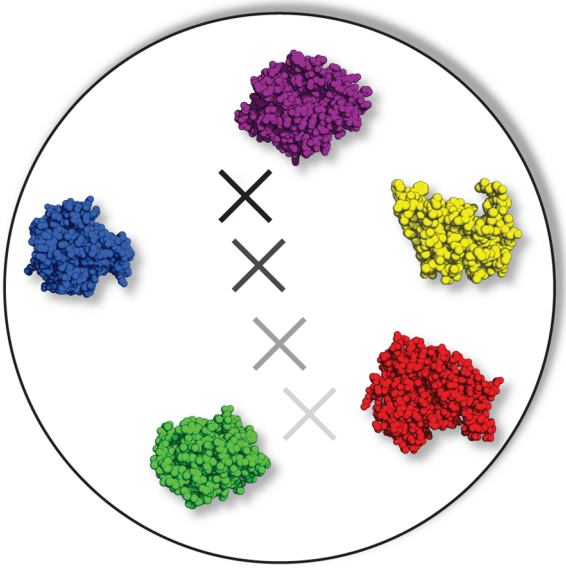


Figure 1. Schematic representation of a hypothetical scoring function space (circle). Every location in the space corresponds to a scoring function, and each colored protein represents the optimal model for a particular set of protein−ligand complexes. Beginning with a fitted model, shown as the lightest gray cross, that performs close to optimally on the green data set, its large distance from the purple model means that the bias on this data set is also large. Optimizing this model for the latter, shown by darkening gray crosses, increases the distance from the green and hence the error of the model on this data set. Having a low error on one set of complexes necessarily means a scoring function performs poorly on others.

VERIFICATION OF ANALYTICAL RESULTS

Four of our test sets were chosen to match the single-protein tests sets used by a study by Cheng et al.,[7] so that we could compare our scoring function to the thirty-three popular scoring functions they tested. Our scoring function achieved a higher Spearman rank coefficient than 79%, 61%, 39%, and 100% of the thirty-three models on the HIV1 protease (B), trypsin (C), carbonic anhydrase II (E), and thrombin (G) data sets respectively. A full comparison between our model and the best and worst models evaluated in the study by Cheng et al. is shown in Table S2 of the Supporting Information. These results show our model is a relatively good scoring function, and we take the following results to be representative of the state of the art.

In 2010, a study by Kramer and Gedeck showed that while the scoring function, RF-score,[16] performed well on a diverse range of protein−ligand complexes, its accuracy was highly variable on single-protein data sets.[8] Our information theoretic analysis indicates that this behavior is independent of the functional form of the model and the protein−ligand complex descriptors. To verify this assertion, we assessed the performance of our own diverse scoring function on the data sets shown in Table 1. Our own scoring function is linear and uses interaction terms such as hydrogen bond strengths. As Figure 2 demonstrates, our diverse scoring function correlates well with experimental affinities on the diverse protein−ligand data set and worse on average on the single-protein data sets, in agreement with what has been observed in previous scoring functions.[7,8]

Figure 2 shows that, also like many popular scoring functions,[6,7] the performance of our diverse model is heavily dependent on the data set it is applied to. This ultimately follows from the 'no free lunch' theorem for supervised learning;[35] a scoring function that performs well on one sampling regime of protein−ligand complexes necessarily means it will perform poorly on others. Also, we know from eq 9 that the regret of a misapplied scoring function can be arbitrarily large if the complexes in the training and test sets have been sampled in a significantly different manner. We further illustrate this difficulty in scoring function optimization

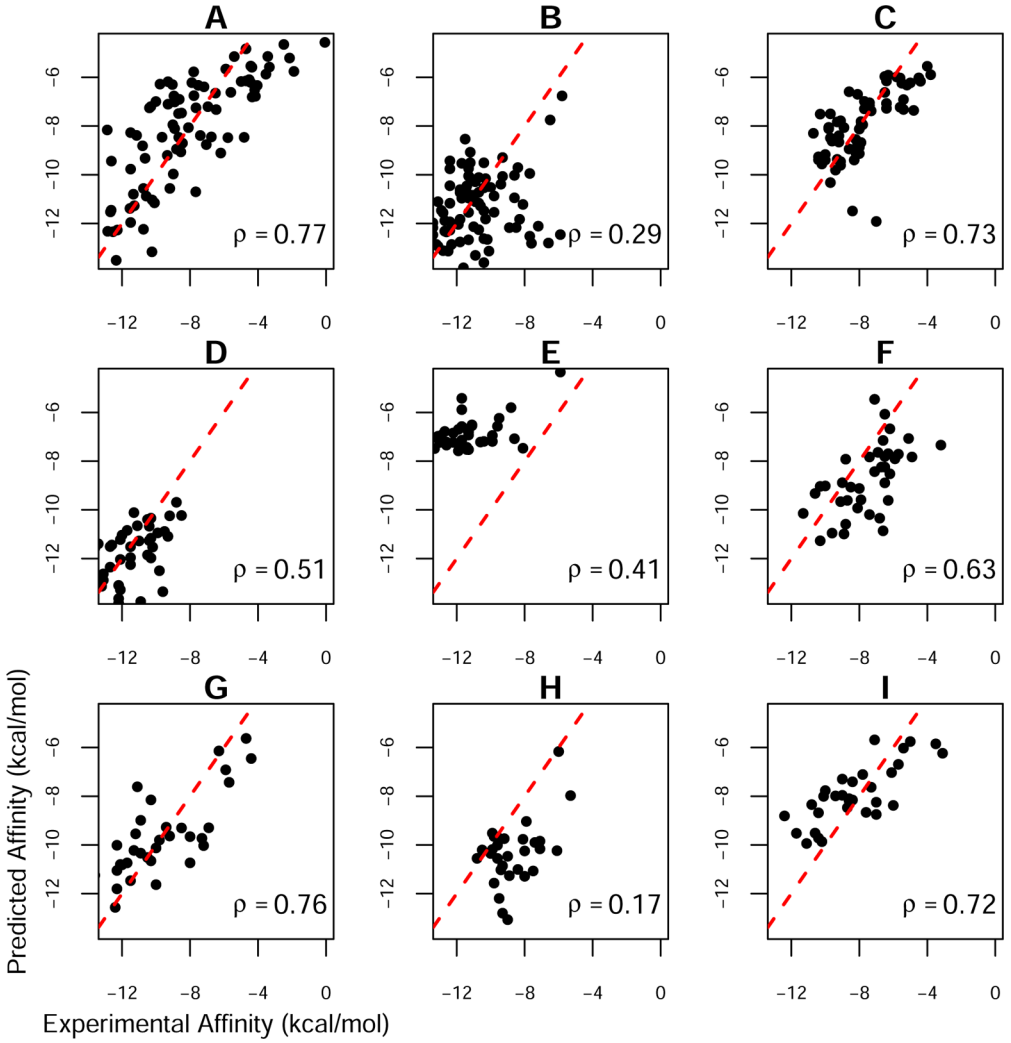


Figure 2. The predictions of the diverse scoring function on each of our test sets (see Table 1). The dashed red line indicates the line of perfect prediction and the Spearman rank coefficient, ρ, for each data set is shown. In Table S2 of the Supporting Information, these rank coefficients along with the Pearson correlation coefficients and standard deviations from a linear fit are shown for data sets B, C, E, and G and compared against the performance of the scoring functions tested by Cheng et al.[7] While the rank order of the ligand affinities is well predicted for the diverse range of complexes (A), the accuracy varies dramatically for the single-protein data sets (B–I). This behavior is similar to what has been observed previously with other scoring functions[6−8] and is explained by our analysis.

by calibrating the free parameters (see Methods) of the diverse scoring function for each protein-specific data set. Changing these parameters can be thought of as moving through a hypothetical scoring function space as depicted in Figure 1 toward the location of an optimal protein-specific model. The relative mean absolute error of the scoring function for each free parameter pair on each data set is shown in Figure 3. No single choice of parameter pair yields the minimum error on all of the sets, and Figure 3 shows that optimizing the scoring function for a data set can increase the error on others.

Equation 11 shows that, for a particular sampling regime, the conditional PDF that encodes for the scoring function with the lowest error over a diverse range of complexes also has the lowest average bias than any other. Yet eq 12 shows that this comparatively low bias is compensated by an intrinsic error that is larger than the average minimum error for specific protein models. This implies that protein-specific scoring functions will have a lower error when applied to their respective protein than the general model. Similarly, misapplying a protein-specific model to the wrong protein will have a higher average error than a generalized scoring function. These analytical results explain the results observed with our own single-protein and diverse scoring functions (see Methods). The average error from cross-validation of each single-protein model is 1.2 kcal/mol compared to an average error of 1.8 kcal/mol of the diverse scoring function. By applying each single-protein scoring functions to the other data sets, we found the average misapplied error to be 2.7 kcal/mol. The relative sizes of each of these errors are in complete agreement with our analytical results. The mean absolute errors of these models on each data set are shown in Figure 4.

By assuming that the cross-validation error of the single-protein models is the minimum error achievable for our scoring function descriptors, we can approximate the regret of the diverse scoring function on a single-protein data set by the difference between the error it achieves and the cross-validation error. On the carbonic anhydrase II data set (E), the diverse scoring function has an exceptional large regret of around 4 kcal/mol. Removing this data set from the analysis, the diverse-scoring function has an apparently encouraging average regret

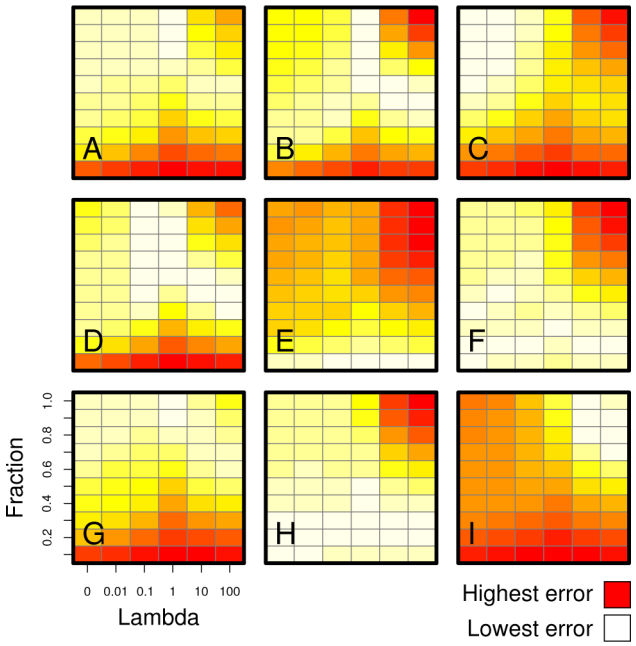


Figure 3. Grid of heatmaps showing the mean absolute error for our own scoring function when applied to data sets A–I (see Table 1). Data set A is composed of many different protein–ligand complexes, while data sets B–I are single-protein data sets. Each heatmap shows the relative error of the model on the data set as the two free parameters of the scoring function - trained using elastic net regularization - are varied (*x*- and *y*-axis). The color gradation indicates parameter-pair choices that give rise to the lowest (white) through to the highest (red) error for that data set. Coloring by relative error highlights that no single parameter-pair choice achieves the lowest error on all nine data sets, so that a model that has the lowest average error over all data sets will not be the best on each individual data set. To compare the absolute values of the errors, Figure S3 of the Supporting Information utilizes an absolute coloring scale, and Table S3 shows the absolute values of the highest and lowest errors on each data set.

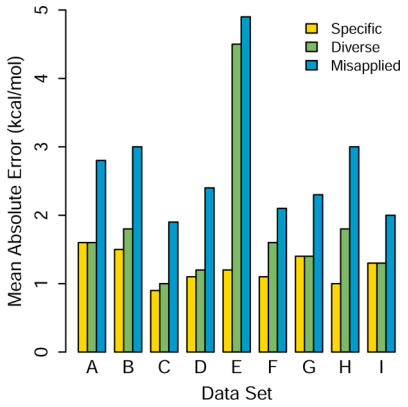


Figure 4. The mean absolute errors of our scoring model on our test sets (see Table 1) when fitted in three different ways. Yellow bars show the cross-validation error of the model when it is fitted to each specific data set; green bars show the error of the model when it is fitted to a diverse range of protein–ligand complexes; blue bars show average errors of the protein-specific scoring functions when misapplied to that data set. In agreement with our mathematical analysis, the error of a protein-specific scoring function is on average less than the error of a general scoring function, which itself is more accurate than the average error of a misapplied protein-specific scoring function.

of 0.3 kcal/mol. In actuality, a truly protein-specific scoring function would be have to be designed from the bottom up, and may include descriptors designed especially for the protein, have a particular functional form, and have the method of regression chosen after experimentation. Our own protein-specific scoring functions are simply recalibrations of our own diverse scoring function. Thus, the regret of our diverse model is likely to be larger when truly protein-specific models are considered.

We attribute the large error on carbonic anhydrase II to the relative infrequency of complexes with metal–ligand interactions appearing in the diverse training set. Carbonic anhydrase II has a catalytic zinc ion in its binding site that interacts with its inhibitors. In the diverse training set, only 37 out of the 207 ligands are within a hydrogen bonding distance to a metal ion. Thus, the data set contains relatively little information about metallic interactions and a correspondingly large regret for these type of complexes is to be expected. Metalloproteins are notoriously difficult to account for in scoring functions, and further optimization or extra protocols are typically required for specific systems.[37−41] The substantially lower error our scoring function attains when recalibrated for carbonic anhydrase II is indicative of this general trend. As our analytical results show, the accuracy of an empirical structure-based scoring function depends on the degree of informational overlap between the training and test sets, such that specific scoring functions have a lower error on average than generalized models.

CONCLUSIONS

By formally analyzing the structure-based modeling process, we have conclusively proven that protein-specific scoring functions will on average achieve a lower error than the very best universal model. We have shown how training and applying a scoring function on different sets of protein–ligand complexes can result in an arbitrarily large error and that a model which performs optimally on one set of complexes necessarily performs poorly on another set by virtue of the 'no free lunch theorem' for supervised learning.

Our results follow from the fact that data sets of protein–ligand structures and affinities are, in general, governed by distinct probability distributions, so that there may be a cost in transferring empirical scoring functions between data sets. This cost is a bias that contributes to the regret of a model. Regret would typically occur in a virtual screening context, where a scoring function that is trained on a diverse range of complexes is used to predict affinities of ligands binding to a single protein. We employed an information theoretic analysis to demonstrate that this error is independent of the way protein–ligand interactions are modeled and is a property of the data itself. Thus, error via bias is fundamental to the nature of protein–ligand scoring. While previous research into the sources of scoring function error has focused on errors from energetic calculations,[20] bias-derived error has remained unreported and explains the variability of scoring function performance on different protein data sets.

Our work demonstrates that there is no 'one size fits all' empirical scoring function and nor will there be. In contrast, techniques based on statistical mechanics that utilize well sampled molecular simulations - in principal - represent universally applicable methodologies. However, even these methods are not free from error, as their accuracy hinges on the type of forcefield used and its parametrization; understanding

forcefield errors is an active area of research.[44−46] As atomic resolution simulations represent a 'bottom-up' approach to free energy calculations and empirical scoring functions a 'top-down' approach, some of the sources of error in both methods may be fundamentally different. As such, the extent to which both approaches are complementary merits investigation.

The speed of empirical models mean they remain ideal for virtual screens. Although generalized empirical models are often used, prior knowledge often exists that may help optimize the scoring function to the protein. One may know some of the active ligands and may be fortunate to have a three-dimensional structure of the target available. A fruitful way forward may be to develop scoring functions that are able to robustly incorporate any prior information one has of the protein into the model. Bayesian methods are ideal for such an approach, and while they have previously been applied to predict ligand activity and protein selectivity,[42,43] the formal exploitation of all prior information, to our knowledge, has not yet been explored. Indeed, rigorous physics-based approaches and empirical models could be combined within a Bayesian framework. If one has extremely limited data of the protein, the development of protein class or family specific models may be a sufficient compromise between generality and accuracy. When fitting to small data sets, these generalized models can also provide the prior distributions on the regression coefficients for Bayessian regression, as it is well established that Bayesian regression greatly stabilizes fitted models in such cases.[23]

Given that model regret is intrinsic to protein−ligand scoring, it remains of paramount importance to be able to estimate its magnitude in a predictive setting and is the subject of our ongoing research. Any further understanding regarding the limits of scoring function accuracy is vital for a more informed and efficient synergy between theoretical affinity predictions and experimentally driven drug development.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional background on the relationship between cross entropy and mean squared error, a detailed description of the scoring function we used to test our theoretical predictions, and Tables S1, S2 and S3, which contain the PDB codes of the complexes in the single-protein data sets, the comparison of our scoring function to other popular models, and the absolute error values of the maximum and minimum errors of Figure 3, respectively. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: philip.biggin@bioch.ox.ac.uk.

Present Address

[§]Crysalin Ltd., Cherwell Innovation Centre, 77 Heyford Park, Upper Heyford OX2 55HD, United Kingdom.

Notes

The authors declare the following competing financial interest(s): G.M.M. states that for the duration of the project he was an employee of InhibOx Ltd.

■ ACKNOWLEDGMENTS

We thank Dr. Rafael Perera, Prof. W. Graham Richards, and Ms. Jessica B. McGillen for critical reading of the manuscript. G.A.R. is funded through the EPSRC Systems Approaches to Biomedical Sciences Doctoral Training Centre and InhibOx Ltd.

■ REFERENCES

(1) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.

(2) Shirts, M.; Mobley, D.; Chodera, J. Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time? *Annu. Rep. Comput. Chem.* **2007**, *3*, 41−59.

(3) Michel, J.; Essex, J. W. Prediction of Protein-Ligand Binding Affinity by Free Energy Simulations: Assumptions, Pitfalls and Expectations. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 639−658.

(4) Steinbrecher, T.; Case, D. A.; Labahn, A. A Multistep Approach to Structure-Based Drug Design: Studying Ligand Binding at the Human Neutrophil Elastase. *J. Med. Chem.* **2006**, *49*, 1837−1844.

(5) Jorgensen, W. L.; Bollini, M.; Thakur, V. V.; Domaoal, R. A.; Spasov, K. A.; Anderson, K. S. Efficient Discovery of Potent Anti-HIV Agents Targeting the Tyr181Cys Variant of HIV Reverse Transcriptase. *J. Am. Chem. Soc.* **2011**, *133*, 15686−15696.

(6) Warren, G. L.; Andrews, W. W.; Capelli, A.-M. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.

(7) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* **2009**, *49*, 1079−1093.

(8) Kramer, C.; Gedeck, P. Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 1961−1969.

(9) Smith, R. D.; Dunbar, J. B.; Ung, P. M.; Esposito, E. X.; Yang, C.-Y.; Wang, S.; Carlson, H. A. CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. *J. Chem. Inf. Model.* **2011**, *51*, 2115−2131.

(10) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.

(11) Schneider, G. Virtual Screening: An Endless Staircase? *Nat. Rev. Drug. Discovery* **2010**, *9*, 273−276.

(12) Huang, S.-Y.; Grinter, S. Z.; Zou, X. Scoring Functions and Their Evaluation Methods for Protein-Ligand Docking: Recent Advances and Future Directions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12899−12908.

(13) Artemenko, N. Distance Dependent Scoring Function for Describing Protein-Ligand Intermolecular Interactions. *J. Chem. Inf. Model.* **2008**, *48*, 569−574.

(14) Sato, T.; Honma, T.; Yokoyama, S. Combining Machine Learning and Pharmacophore-Based Interaction Fingerprint for in Silico Screening. *J. Chem. Inf. Model.* **2010**, *50*, 170−185.

(15) Durrant, J. D.; McCammon, J. A. NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1865−1871.

(16) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein-Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26*, 1169−1175.

(17) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(18) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.

(19) Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 4. Are Popular Scoring Functions Accurate for This Class of Proteins? *J. Chem. Inf. Model.* **2009**, *49*, 1568−1580.

(20) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz, K. M. Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein-Ligand Complexes. *J. Chem. Theory Comput.* **2011**, *7*, 790−797.
(21) Merz, K. M. Limits of Free Energy Computation for Protein-Ligand Interactions. *J. Chem. Theory Comput.* **2010**, *6*, 1769−1776.
(22) Seifert, M. H. J. Targeted Scoring Functions for Virtual Screening. *Drug Discovery Today* **2009**, *14*, 562−569.
(23) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, USA, 2009.
(24) Burnham, K.; Anderson, D. *Model Selection and Multimodel Inference*; Springer: New York, USA, 2002.
(25) Principe, J. C. *Information Theoretic Learning*; Springer: New York, USA, 2010.
(26) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley-Interscience: New York, USA, 2006.
(27) Loomes, G.; Sugden, R. Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal* **1982**, *92*, 805−824.
(28) Verdu, S. Mismatched Estimation and Relative Entropy. *IEEE Trans. Inf. Theory.* **2010**, *56*, 3712−3720.
(29) Kellogg, G. E.; Abraham, D. J. Hydrophobicity: Is LogPo/w More than the Sum of Its Parts? *Eur. J. Med. Chem.* **2000**, *35*, 651−661.
(30) 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.
(31) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *J. Comput. Chem.* **2007**, *28*, 1145−1152.
(32) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.
(33) Zou, H.; Hastie, T. *Regularization and variable selection via the Elastic Net* **2005**, 301−320.
(34) R Development Core Team, *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2011.
(35) Wolpert, D. H. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Comput.* **1996**, *8*, 1341−1390.
(36) Heskes, T. Selecting Weighting Factors in Logarithmic Opinion Pools. *Adv. Neural Inf. Process. Syst.* **1998**, 266−272.
(37) Hu, X.; Shelver, W. H. Docking studies of matrix metalloproteinase inhibitors: zinc parameter optimization to improve the binding free energy prediction. *J. Mol. Graphics Modell.* **2003**, *22*, 115−126.
(38) Hu, X.; Balaz, S.; Shelver, W. H. A practical approach to docking of zinc metalloproteinase inhibitors. *J. Mol. Graphics Modell.* **2004**, *22*, 293−307.
(39) Schiffmann, R.; Neugebauer, A.; Klein, C. D. Metal-mediated inhibition of Escherichia coli methionine aminopeptidase: structure-activity relationships and development of a novel scoring function for metal-ligand interactions. *J. Med. Chem.* **2006**, *49*, 511−522.
(40) Jain, T.; Jayaram, B. Computational protocol for predicting the binding affinities of zinc containing metalloprotein-ligand complexes. *Proteins: Struct., Funct., Bioinf.* **2007**, *67*, 1167−1178.
(41) Röhrig, U. F.; Grosdidier, A.; Zoete, V.; Michielin, O. Docking to heme proteins. *J. Comput. Chem.* **2009**, *30*, 2305−2315.
(42) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47*, 4463−4470.
(43) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguiz, R. M.; Huang, X. P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; Simeons, F. R. C.; Stojanovski, L.; Prat, A.; Seidah, N. G.; Constam, D. B.; Bickerton, G. R.; Read, K. D.; Wetsel, W. C.; Gilbert, I. H.; Roth, B. L.; Hopkins, A. L. Automated Design of Ligands to Polypharmacological Profiles. *Nature* **2012**, *492*, 215−220.
(44) Faver, J. C.; Yang, W.; Merz, K. M. The Effects of Computational Modeling Errors on the Estimation of Statistical Mechanical Variables. *J. Chem. Theory Comput.* **2012**, *8*, 3769−3776.
(45) Rocklin, G. J.; Mobley, D. L.; Dill, K. A. Calculating the Sensitivity and Robustness of Binding Free Energy Calculations to Force Field Parameters. *J. Chem. Theory Comput.* **2013**, *9*, 3072−3083.
(46) Di Pierro, M.; Elber, R. Automated Optimization of Potential Parameters. *J. Chem. Theory Comput.* **2013**, *9*, 3311−3320.