J Comput Aided Mol Des (2012) 26:647–659
DOI 10.1007/s10822-012-9567-9

Prediction of trypsin/molecular fragment binding affinities by free energy decomposition and empirical scores

Mark L. Benson · John C. Faver · Melek N. Ucisik · Danial S. Dashti · Zheng Zheng · Kenneth M. Merz Jr.

Received: 24 October 2011 / Accepted: 21 March 2012 / Published online: 4 April 2012

Abstract Two families of binding affinity estimation methodologies are described which were utilized in the SAMPL3 trypsin/fragment binding affinity challenge. The first is a free energy decomposition scheme based on a thermodynamic cycle, which included separate contributions from enthalpy and entropy of binding as well as a solvent contribution. Enthalpic contributions were estimated with PM6-DH2 semiempirical quantum mechanical interaction energies, which were modified with a statistical error correction procedure. Entropic contributions were estimated with the rigid-rotor harmonic approximation, and solvent contributions to the free energy were estimated with several different methods. The second general methodology is the empirical score LISA, which contains several physics-based terms trained with the large PDBBind database of protein/ligand complexes. Here we also introduce LISA+, an updated version of LISA which, prior to scoring, classifies systems into one of four classes based on a ligand's hydrophobicity and molecular weight. Each version of the two methodologies (a total of 11 methods) was trained against a compiled set of known trypsin binders available in the Protein Data Bank to yield scaling parameters for linear regression models. Both raw and scaled scores were submitted to SAMPL3. Variants of LISA showed relatively low absolute errors but also low correlation with experiment, while the free energy decomposition methods had modest success when scaling factors were included. Nonetheless, re-scaled LISA yielded the best predictions in the challenge in terms of RMS error, and six of these models placed in the top ten best predictions by RMS error. This work highlights some of the difficulties of predicting binding affinities of small molecular fragments to protein receptors as well as the benefit of using training data.

Keywords SAMPL · Docking and scoring · Error analysis · Protein–ligand interactions

Introduction

The ability to accurately predict the binding affinity of a small molecule ligand to a protein receptor is one of the most challenging problems in computational chemistry. The problem is of great interest in pharmaceutical research, where large libraries of compounds are often screened in silico for their ability to bind to protein targets. These tools can help predict which sets of molecules are more likely to be active and thereby direct the focus of a drug design effort toward more productive regions of chemical space. However, accurately predicting such binding affinities has proven to be a very difficult task. While significant progress has been made in the in silico docking of ligands (i.e. predicting bound structures of ligands in a receptor's active site), there is still room for improvement in scoring functions (predicting free energies of binding) [1–4]. Thus there is a need for prospective studies such as the SAMPL challenge in the docking and scoring community, as they introduce less bias and offer fair evaluation of scoring protocols [5].

Electronic supplementary material The online version of this article (doi:10.1007/s10822-012-9567-9) contains supplementary material, which is available to authorized users.

M. L. Benson · J. C. Faver · M. N. Ucisik · D. S. Dashti · Z. Zheng · K. M. Merz Jr. (✉)
The Quantum Theory Project, The University of Florida, 2328 New Physics Building, P.O. Box 118435, Gainesville, FL 32611-8435, USA
e-mail: merz@qtp.ufl.edu

The SAMPL challenge reflects the real life scenario of attempting to predict the unknown binding affinities of small molecules [6]. One major emphasis of the exercise is to help provide a benchmark for different approaches, and to inspire the community to improve and innovate. In this spirit, we introduce a novel procedure for the estimation and correction of energy function errors into our results. Our lab has been investigating the effects of energy function errors and means of correcting for these errors in different applications [7–9]. In this work we applied our error estimation methods to the rescoring of docked ligands, to test whether it would improve our ability to rank ligands by affinity. In addition, our lab has developed a knowledge-based score function LISA which has demonstrated excellent performance in earlier benchmark studies [10]. We were eager to test the performance of LISA and its variants on a blind test such as the SAMPL challenge as well. We also introduce here an alternate form of LISA, named LISA+, which classifies protein/ligand complexes into one of four classes based on ligand molecular weight and hydrophobicity. The classification scheme offers reduced errors in binding affinity predictions in several test sets.

For the purposes of this exercise, we also compiled a set of trypsin inhibitors with known binding affinities, which was used as a training set for our scoring methods. Predictions from each of our models were fit to linear regression models with the training set. This set was compiled from the Binding MOAD database [11, 12] and was used to improve our energy models by tailoring them specifically to trypsin inhibitors. In total, 24 submissions were generated for the SAMPL3 exercise including predictions from two families of our own methodologies as well as Schrodinger's Glide XP [13–17] score. In this manuscript we report each of the methods and protocols used, and then compare their predictions with each other as well as the experimental measurements.

Methods

Scoring method family 1: free energy decomposition

The first family of scoring functions comprises end-point methods utilizing the thermodynamic cycle of Fig. 1, which models the free energy of binding in solution ΔG_b^s as

$$\Delta G_b^s = \Delta G_b^g + \Delta G_{solv}^{PS} - \Delta G_{solv}^{S} - \Delta G_{solv}^{P} \quad (1)$$

where P and S represent the protein and substrate systems, ΔG_{solv}^x represent solvation free energies, and ΔG_b^x represent binding free energies in either the gas(g) or solution(s) phase. Our free energy models separate solvation free energies from the gas-phase binding free energy, and further decompose the gas-phase binding free energy into enthalpic and entropic contributions as shown in Eq. 2 The individual terms in Eq. 2 can be estimated separately with various methods. We generated four such different models for ΔG_b^s, hereafter referred to as M1, M2, M3, and M4. The various components of each model will be described in the following sections.

$$\Delta G_b^g = \Delta H_b^g - T\Delta S_b^g \quad (2)$$

Enthalpy and statistical corrections

The gas-phase enthalpy contribution for all free energy decomposition models was approximated with differences in heats of formation (Eq. 3) calculated with the PM6-DH2 semiempirical quantum Hamiltonian using the MOPAC2009 program [18, 19]. The input structures came directly from the docking procedure described in later sections. This is a significant approximation since it only considers enthalpy contributions from single microstates, the presumed energy minima, for each ligand, receptor, and complex system. The MOZYME keyword was used to carry out the receptor and complex calculations, to utilize MOPAC's linear-scaling localized molecular orbital algorithm.

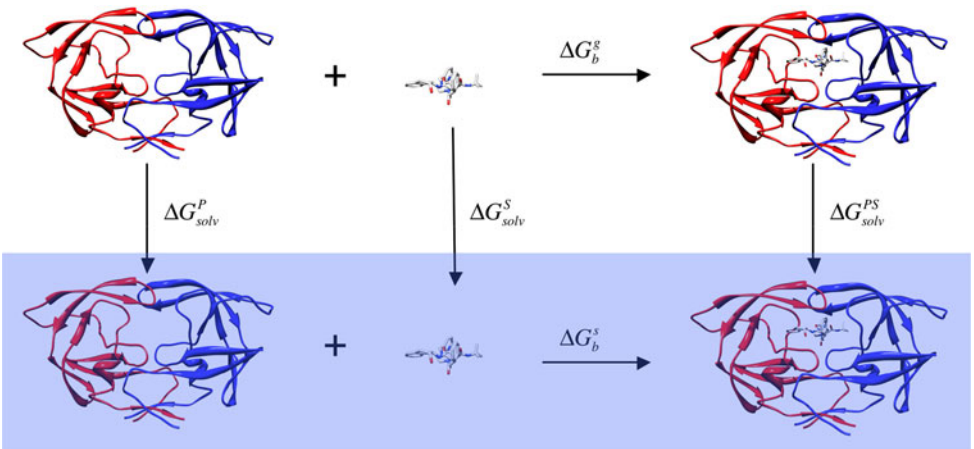

Fig. 1 The thermodynamic cycle used to formulate the free energy of protein–ligand binding in solution. The free energy decomposition schemes separate solvent contribution from the gas-phase binding free energy, and further decompose the binding free energy into enthalpic and entropic components

$$\Delta H_b^g = \Delta H_{complex}^f - \left(\Delta H_{\text{Receptor}}^f + \Delta H_{Ligand}^f\right) \quad (3)$$

The PM6-DH2 interaction energy model was selected for both its computational speed and its accuracy with respect to reproducing interaction energies calculated by high-level quantum methods performed on biologically relevant molecular systems. In addition, similar free energy decomposition schemes using PM6-DH2 enthalpies have been used and they have shown significant correlation with binding affinities of HIV-1 protease inhibitors [20]. In our own previous work concerning energy model error estimation procedures, we generated a database of biomolecular fragments and calculated interaction energies from several different computational methods [7, 8]. With respect to our reference method (coupled cluster with singles, doubles, and perturbative triples with complete basis set extrapolation, i.e. CCSD(T)/CBS), PM6-DH2 showed distributions of interaction energy errors with means and standard deviations of -0.09 ± 0.32 kcal/mol for van der Waals type interactions ($N = 42$), and 0.61 ± 1.39 kcal/mol for polar/hydrogen bonding type interactions ($N = 50$). Figure 2 contains a histogram of the individual error magnitudes in the fragment database along with a plot of the fitted Gaussian probability density functions describing the interaction energy errors of PM6-DH2 for both classes of interactions. For more information on this analysis, see [7] and [8].

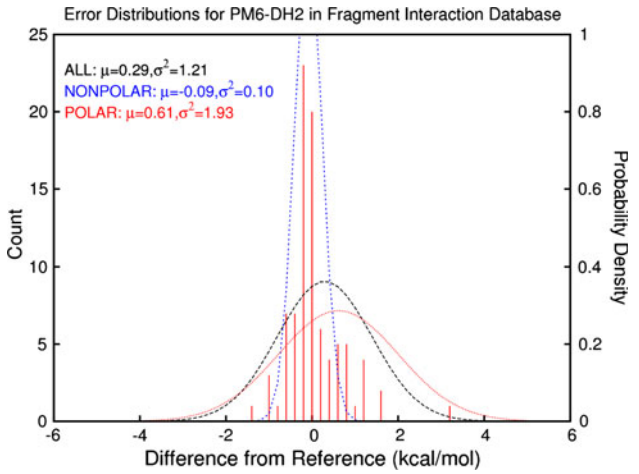

Fig. 2 Histogram and Gaussian probability density functions describing errors in the PM6-DH2 energy model for different classes of molecular interactions with respect to a CCSD(T)/CBS reference. Based on our previous analysis on a set of small interacting biomolecular fragments, PM6-DH2 tends to model nonpolar/van der Waals contacts with an error distribution of -0.09 ± 0.32 kcal/mol (mean and standard deviation), and polar/hydrogen bonding contacts with an error distribution of 0.61 ± 1.39 kcal/mol. These distributions were used to estimate new fragment-based errors which were propagated to yield overall systematic errors and uncertainties in enthalpy calculations with PM6-DH2

The probability density functions in Fig. 2 describing the modeling errors in PM6-DH2 were not only used to justify using the energy model, but were also used to correct systematic errors and produce minimum uncertainty estimates in our enthalpy calculations. Each docked pose was analyzed for the number and class of each molecular interaction present at the protein–ligand interface, and each interaction's contribution to overall error was estimated with the appropriate probability density function and propagated with the formulas in Eq. 4. Propagated systematic errors were always subtracted from the enthalpy calculations, and random errors were used to estimate minimum uncertainties in our free energy estimates. This error propagation procedure assumes independence of fragment contributions to enthalpy, an approximation which is supported by computational and experimental evidence [21, 22]. It also neglects errors from the entropy and solvation terms of the overall free energy score. Systematic errors in the enthalpy and remaining terms may work in the same or opposite directions, either magnifying or reducing total systematic errors. Random errors, on the other hand, will only increase as random errors from the entropy and solvation terms are propagated with the enthalpy random error. Thus our estimated random error should be considered a lower bound of the uncertainty in the free energy score.

$$Error_{Systematic} = N_{vdW}(-0.09) + N_{Polar}(0.61) \text{ and}$$
$$Error_{Random} = \sqrt{N_{vdW}(0.10) + N_{Polar}(1.93)} \quad (4)$$

Entropy

Entropy contributions to gas-phase free energy of binding for all energy models were calculated using Schrödinger's MacroModel [23]. Contributions to entropy from the translational, rotational, and vibrational degrees of freedom were calculated using the rigid rotor harmonic approximation (RRHO). The sum of these components was used as the total entropy for the system.

Solvation

Four different models for the solvation free energy component were used. M1 used the MM/GBSA solvation free energy calculated by Prime from Schrödinger [24–26]. M2 used COSMO implicit solvation energies [27]. The COSMO solvation energy contribution was estimated by performing each enthalpy calculation in MOPAC [18] with and without using the COSMO continuum solvent model, and using the formula in Eq. 5. The dielectric constant was set to 78.4 for each of these calculations. It should be noted that the COSMO solvation term is a severe approximation since it neglects entropy contributions to solvation free energy.

$$\Delta G_{Cosmo} = \left[\Delta H^{PS}_{COSMO} - \Delta H^{PS}_{GAS}\right] - \left[\Delta H^{P}_{COSMO} - \Delta H^{P}_{GAS}\right] - \left[\Delta H^{S}_{COSMO} - \Delta H^{S}_{GAS}\right] \quad (5)$$

M3 and M4 used solvation free energies from MM/PBSA and MM/GBSA calculations in AMBER. The AMBER MM/PBSA [28, 29] and MM/GBSA [28, 30] solvation free energy estimations were extracted from 4 individual 500 ps-long MD simulations run with the ff99SB force field [31]. For each calculation we utilized 40 frames. The calcium ion was stripped from the structure for the analysis along with the added counter chloride ions and water molecules. The salt concentration was set to 0.100 M for the GBSA calculations while the ionic strength was adjusted to 0.100 mM for PBSA.

Scoring method family 2: LISA and LISA+

The second family of scoring functions was the empirical scoring function LISA (Ligand Identification Scoring Algorithm) and its variants developed in our group, where the binding free energy (expressed in pK_d units) is represented by a linear model of van der Waals (VDW) contacts, hydrogen bonds, desolvation effects and metal chelation terms [10]. A training set of 492 protein–ligand complexes from the PDBbind v2010 database [32] was selected for its parameterization. LISA contains 20 different atom types based on common interacting atoms found in the PDBbind training set. The score can be written as [10]:

$$\begin{aligned} pK_d = {} & c_1 M_{VDW\,C3-C3} + c_2 M_{VDW\,C3-C2/Car} + c_3 M_{VDW\,C3-N3/Npl3} + c_4 M_{VDW\,C3-N4} \\ & + c_5 M_{VDW\,C3/C2/Car-S} + c_6 M_{VDW\,C2-C2} + c_7 M_{VDW\,C2-O3} + c_8 M_{VDW\,C2-O2} \\ & + c_9 M_{VDW\,C2-Npl3} + c_{10} M_{VDW\,Car-Car} + c_{11} M_{VDW\,Car-O2} + c_{12} M_{VDW\,Car-N3} \\ & + c_{13} M_{VDW\,Car-N2} + c_{14} M_{VDW\,O-N} + c_{15} M_{HB\,O-O} + c_{16} M_{HB\,O-N} \\ & + c_{17} M_{SASA} + c_{18} M_{chelation} \end{aligned} \quad (6)$$

LISA/LISA+ Van der Waals interactions

Van der Waals interactions are some of the most important interactions present in protein–ligand complexes. Their potential energy contributions depend on the distances between pairs of atoms. The Lennard-Jones 6–12 term is employed in LISA to reflect van der Waals interactions when two atoms approach one another during the binding process between a protein and a ligand.

$$M_{VDW\,AB} = \varepsilon_{AB} \sum_{i \in L} \sum_{j \in P} f_{ij}(x, y, z) \quad (7)$$

$$f_{ij}(x, y, z) = \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \quad (8)$$

In Eq. 7 and 8 r_{ij} is the distance between atom i in the ligand and atom j in the protein. σ_{ij} is the interatomic separation at which repulsive and attractive forces balance (the sum of the van der Waals radii of atom i and atom j). ε_{AB} is the potential well depth, subscripts A and B refer to atom type A and B.

LISA/LISA+ hydrogen bonding

Hydrogen bonding is another important interaction type found in most protein–ligand complexes. The principal variables associated with hydrogen bonding are the distance between the hydrogen bond donor and hydrogen bond acceptor, d_{HA}, the bond angle between the hydrogen bond donor and acceptor, $\theta_{D\text{-}H\text{-}A}$, and the H—A-AA angle defined by the hydrogen bond acceptor, $\sigma_{H\text{-}A\text{-}AA}$. In LISA, we model hydrogen bonding with Eq. 9. In this description of hydrogen bonding, d_{HA}, $\theta_{D\text{-}H\text{-}A}$ and $\sigma_{H\text{-}A\text{-}AA}$ have defined optimal values. Deviation in d_{HA}, $\theta_{D\text{-}H\text{-}A}$ and $\sigma_{H\text{-}A\text{-}AA}$ from these optimal values destabilizes the hydrogen bond interaction. See [10] for details.

$$\begin{aligned} M_{h-bond} &= f_1(d_{HA}) f_2(\theta_{D-H\cdots A}) f_3(\sigma_{H\cdots A-AA}) \\ f_1(d_{HA}) &= \varepsilon \left[\left(\frac{r_0}{r_{ij}}\right)^{12} - 2\left(\frac{r_0}{r_{ij}}\right)^{6} \right] \\ f_2(\theta_{D-H\cdots A}) &= \cos^2(\theta_{D-H\cdots A} - \theta_0) \\ f_3(\sigma_{H\cdots A-AA}) &= \cos^2(\sigma_{H\cdots A-AA} - \sigma_0) \end{aligned} \quad (9)$$

LISA/LISA+ desolvation

Desolvation causes changes in the entropy as well as in enthalpy of the binding event. In LISA, we associate the free energy change caused by the desolvation effect with the binding surface area. A grid-based algorithm was developed to generate the binding surface area for LISA. The algorithm and parameterization is outlined in [10].

LISA/LISA+ metal chelation

Metal chelates are observed in numerous metalloprotein–ligand complexes as metal binding "warheads". LISA

contains a term specifically for Zn chelation, but it was not utilized in this work on trypsin inhibitors [10].

LISA+ development and validation

While testing LISA, we found that it had relatively poor predictive ability when it came to ranking ligands within the low affinity (pK_d, $pK_i < 5$) region and also ranking those within the high affinity (pK_d, $pK_i >= 8$) region. Table 1 displays the RMSE (root mean square error) values of LISA predictions in different binding regions in the first test set for LISA validation [10]. There was a trend of the ligands' carbon atom number fraction and ligand molecular weight (MW) increasing from the low binding region to the high binding region. This suggested that ligand polarity and size were potential factors in our scoring function prediction.

In order to improve the prediction ability of LISA, we classify ligands into groups based on their size and polarity and use different parameter sets to evaluate the binding affinity. So, in LISA+, before the program does any scoring, it first categorizes ligands into different groups based on the ratio of carbon atoms in the whole ligand and the molecular weight (MW). The first group has a low carbon ratio and low molecular weight (carbon ratio <=0.65 and MW <=350), the second has a low carbon ratio, high molecular weight (carbon ratio <=0.65 and MW >350), the third has a high carbon ratio, low molecular weight (carbon ratio >0.65 and MW <=350), and the last has a high carbon ratio, high molecular weight (carbon ratio >0.65 and MW >350). Four different sets of scoring parameters are applied to these four groups. The previous data set with 1,399 complexes were used for LISA+ parameterization. The linear fitting results are shown in Table 2. For some of the four classes, certain terms were found to be insignificant in determining the binding affinity and these terms are neglected in evaluating the LISA+ score.

For validation of LISA+, 486 protein–ligand complexes were analyzed with both LISA and LISA+, and the two scores were compared (Fig. 3). LISA gave an R^2 of 0.4221, RMSE of 1.577 and a standard error of 0.0715, while LISA+ yielded an R^2 of 0.4717, RMSE of 1.419 and a Standard Error of 0.0644. From this analysis, we could see that LISA+ offers minor improvement over LISA. As knowledge-based scoring functions, LISA and LISA+ both have acceptable ability to predict protein–ligand binding affinities. However, the detailed atom type assignment used in the LISA+ scoring function is important to improve overall prediction ability.

Receptor selection

The protein model used for docking the 34 fragments of the SAMPL3 binding affinity test set was selected after careful evaluation of a large pool of trypsin/inhibitor complex crystal structures. A set of 43 trypsin complex crystal structures annotated with published binding data was constructed with entries from Binding MOAD. All 43 structures had resolutions of 2.5 Å or better, and had binding data ranging from 5 nM up to 36 mM. The ligands from the 43 structures were then filtered to remove duplicates and very large ligands, resulting in a set of 35 unique complexes. To identify which of the structures would serve as the most suitable protein receptor, a cross-docking procedure was performed. A library of the 35 different ligands was generated and prepared for docking using Schrodinger's LigPrep [33]. The library of ligands was then docked into the pool of protein crystal structures using Glide XP [15].

After docking, the resulting poses were compared against each other in terms of correct ranking of the ligands by binding affinity, RMSD from the crystal structures, and Glide XP scores. However, choosing the ideal receptor based on the previous criteria failed to identify a single ideal receptor—partly because of the high structural similarity of the protein structures, and their overall abilities to bind the ligand set. Nonetheless, the protein model from PDB ID: 1O3G [34] was chosen as the best candidate for docking based on its overall ligand binding properties. The 1O3G protein model (1) bound its ligand extremely tightly, (only 1TPS bound tighter with an $IC_{50} = 10$ nM, while 1O3G was inhibited at $K_i = 0.11$ nM), (2) demonstrated the highest ligand efficiency (defined as the ratio of Gibbs free energy to the number of non-hydrogen atoms in the ligand), (3) did not have a large peptide-like ligand like 1TPS (the expected targets were expected to not be peptide-like), and (4) had a very high-quality crystal structure (1.55 Å resolution). The choice of 1O3G as a protein model was validated by the docking of the training set. For the

Table 1 Relationship between LISA's prediction ability and ligand molecular properties (with standard errors, SE) in the LISA test set with 1,399 complexes

pK_d or pK_i	Number of ligands	RMS error	Standard error	Average C atom ratio	C atom ratio SE	Average molecular weight (g/mol)	Molecular weight SE
<5	301	1.795	0.044	0.5783	0.0076	365.78	6.71
5–8	743	1.589	0.031	0.6678	0.0052	417.40	4.66
>8	355	2.419	0.057	0.7257	0.0056	474.68	7.21

Table 2 Parameters derived from linear fitting for four different sets of scoring functions

Interaction type	Weight	95 % confidence interval		Interaction type	Weight	95 % confidence interval	
Low carbon ratio and low molecular weight				*Low carbon ratio, high molecular weight*			
sp3 C–sp2 C	0.2365	0.0207	0.4524	sp3 C–sp2 C	0.2460	0.0374	0.4546
sp3 C–sp2 O	0.2056	0.0706	0.3405	sp3 C–sp3 O	0.0989	0.0022	0.1955
sp3 C–sp3 N	0.4360	0.0189	0.8531	sp3 C–sp2 O	0.1223	0.0012	0.2433
sp3 C–sp2 N	−0.1343	−0.2497	−0.0189	sp3 C–sp2 N	−0.2792	−0.5439	−0.0144
sp3 C–N cation	3.4010	1.1635	5.6385	sp3 C–N cation	3.7510	1.1302	6.3719
sp3 C–S	1.2208	0.3358	2.1057	sp3 C–S	0.5418	0.0459	1.0376
sp2 C–sp2 C	0.1228	0.0132	0.2325	sp2 C–sp2 O	0.3239	0.0010	0.6468
sp2 C–sp3 O	0.0941	0.0025	0.1857	sp2 C–sp3 N	0.1276	0.0420	0.2131
sp2 C–sp2 O	0.3247	0.0322	0.6172	sp2 C–sp2 N	0.4712	0.0206	0.9218
sp2 C–N cation	−1.5279	−2.5542	−0.5016	sp2 C–N cation	−0.9372	−1.8463	−0.0280
HB O–H…N	1.9492	0.0662	3.8322	HB O–H…O	0.9482	0.1412	1.7552
HB N–H…O	1.1315	0.3392	1.9239	HB O–H…N	0.8587	0.0217	1.6957
Surface area	−0.0449	−0.0628	−0.0271	HB N–H…O	2.6710	1.9054	3.4365
				Surface area	−0.0229	−0.0355	−0.0103
High carbon ratio, low molecular weight				*High carbon ratio, high molecular weight*			
sp3 C–sp3 C	0.3559	0.1568	0.5551	sp3 C–sp3 C	0.1701	0.0334	0.3067
sp3 C–sp2 C	0.2168	0.1056	0.3280	sp3 C–sp2 C	0.0803	0.0012	0.1593
sp3 C–sp3 O	−0.1627	−0.0129	−0.3125	sp3 C–sp3 O	−0.1218	−0.2328	−0.0108
sp3 C–sp2 N	0.2194	0.0233	0.4155	sp3 C–sp3 N	0.3575	0.0988	0.6162
sp3 C–N cation	1.6176	0.1159	3.1193	sp3 C–S	1.1676	0.7129	1.6223
sp3 C–S	1.7532	0.9358	2.5707	sp2 C–sp2 C	0.1480	0.0061	0.2899
sp2 C–sp2 C	0.0859	0.0054	0.1664	sp2 C–sp3 O	0.6735	0.3280	1.0190
sp2 C–sp3 O	0.2253	0.0137	0.4370	sp2 C–sp2 O	0.0842	0.0011	0.1673
sp2 C–sp3 N	0.2278	0.0097	0.4459	sp2 C–sp3 N	0.1699	0.0061	0.3337
sp2 C–N cation	−1.6420	−2.7613	−0.5228	sp2 C–N cation	−0.8892	−1.7745	−0.0038
sp2 C–S	1.2274	0.4191	2.0357	sp2 C–S	0.6735	0.0036	1.3433
HB O–H…O	1.3705	0.3657	2.3753	HB O–H…O	0.6379	0.0365	1.2393
HB N–H…O	1.1544	0.0241	2.2847	HB O–H…N	3.9650	0.7266	7.2033
Surface area	−0.0469	−0.0561	−0.0377	HB N–H…O	0.6306	0.0688	1.1925
				Surface area	−0.0371	−0.0440	−0.0302

The interaction weights are given along with the lower and upper bounds of the 95 % confidence interval

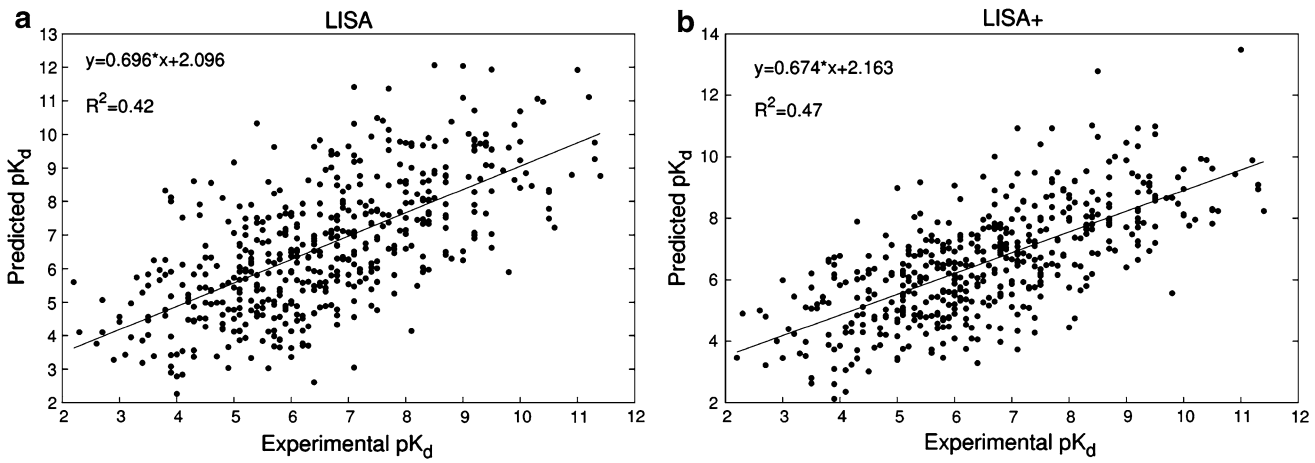

Fig. 3 LISA and LISA+ estimated pK_d versus the experimental pK_i or pK_d for the PDBbind v2010 test set of 486 protein–ligand complexes

docking exercises, the 1O3G receptor was stripped of its chlorine ion and most waters, but the calcium ion was retained with three nearby water molecules as well as three waters near the binding site as illustrated in Fig. 4.

Training set docking

The quality of the docked poses generated with the chosen model receptor is an essential component of accurate binding affinity prediction. To assess the accuracy of pose generation and the general protocol for calculating binding affinities, we constructed a small training set, which consisted of known binders of trypsin. The ligands were extracted from the following complexes deposited in the PDB: 1TX7 [35], 1QB6 [36], 1G3B [37], 1Y3V [38], 1Y3W [38], 1QBN [36], 1BJU [39], 1C5T [40], 1UTP [41], 1UTO [41], 1K1M [42], 1F0T [43], 1BJV [39], 1K1I [42], 1Y5U [44], 1Y3X [38], 1OYQ [45], 1UTL [41], 1G36 [45], 1F0U [43], and 1Y5B [44]. These ligands covered a range of molecular weights from 122 to 525 g/mol and they spanned a range of binding affinities from 36 mM to 40 nM. The variety in size and binding affinity was desired in order to test the compatibility of the model receptor of choice (1O3G) to bind a diverse set of ligands. The experimental binding affinities were then used to train our different scoring methods (described in the "Scaling procedure"). The structures of the training set ligands are shown in the supplementary information.

The training set ligands were all docked successfully into the 1O3G protein model using the protocol outlined in the supplementary information. RMSD values of the docked poses (as compared to their poses in the native complex structures) were calculated to qualitatively evaluate the docking results, but were not assigned significant weight as RMSD is known to be a poor metric for evaluating poses [46, 47]. As the Glide XP score has been shown to be a reasonable metric to evaluate poses [15], the correlation of the known binding affinities of the training set ligands to the GLIDE XP docking scores was calculated, and this yielded a square correlation coefficient of $R^2 = 0.747$. All of the Glide XP docking scores were lower than the experimental free energies of binding and there were four ligands that were predicted to have binding values between 3 and 4 kcal/mol tighter than what was observed. Overall, this exercise validated the protein model and the docking procedure, and provided a set of poses to be evaluated with our free energy models for training.

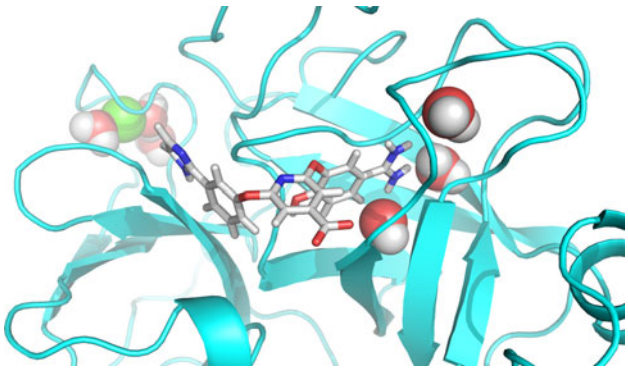

Fig. 4 Example docked pose showing the three retained crystal water molecules in the active site of the 1O3G receptor and the distant calcium ion bound to three additional waters

Scaling procedure

The four free energy decomposition methods were fit by least-squares minimization to the training set free energies assuming the linear form shown in Eq. 10. The LISA methods were fit to a linear model to reproduce the training set pK_d values (Eq. 11). Furthermore, LISA was trained with two pose sets: (1) the training set poses selected by PM6-DH2 enthalpies, and (2) the poses taken directly from the PDB entries of the training set. In order to distinguish between these parameter sets, we denote each training set with either s (for semiempirical), g (for Glide XP), or p(for PDB) in parentheses, e.g. M1(s).

$$\Delta G = a\Delta H_{\text{int}} + b\Delta G_{Solv} + c\Delta TS_{RRHO} + d \quad (10)$$

$$\Delta G = a(pK_d)_{LISA} + b \quad (11)$$

Test set docking

After the model receptor 1O3G was chosen, the SAMPL3 test set was finally docked to the model receptor. The 3D structures of the ligands were generated from SMILE strings [48, 49] using CORINA [50]. The test set ligands were processed into a screening library using Schrodinger's LigPrep, and were then docked to the 1O3G model using Schrodinger's Glide XP with the same settings that were used for the training set. Two sets of poses of the test set were used in post-processing which were selected by two criteria: (1) the most favorable PM6-DH2 interaction energy between the receptor and all the available poses of each particular ligand, and (2) the best Glide XP docking score obtained from the docking run.

Discussion

Training set

As seen in Fig. 5, the training set ligands fill the subpockets S2, S3, and S4 in addition to the S1 specificity pocket. The Asp189 residue of the S1 pocket is responsible for selecting the basic residues (e.g. Lys and Arg) for cleavage by forming a bidentate salt bridge between its carboxylate

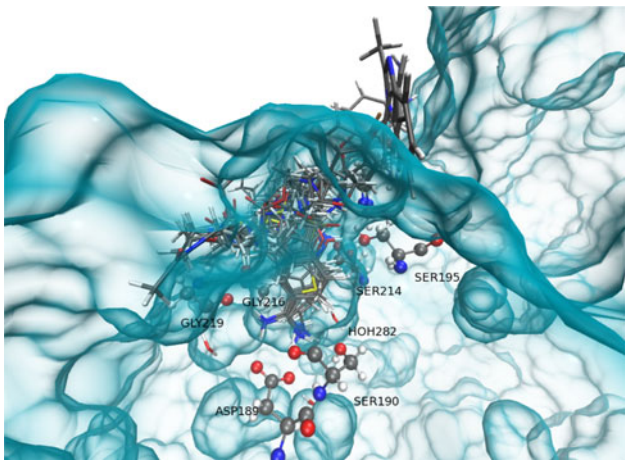

Fig. 5 Training set ligands superimposed in the active site of the model receptor 1O3G. The benzamidine group forms a bidentate salt bridge with Asp189, which normally determines the orientation of lysine and arginine residues of the substrate for the peptide bond cleavage. The benzamidine moiety commonly fills the S1 pocket while the additional subpockets are filled with varying chemical groups of the ligands. The poses in all the figures are those selected by the lowest PM6-DH2 protein–ligand interaction energy

moiety and the substrate's positively charged side chains [42, 51]. The majority of the training set ligands possessed benzamidine moieties, which interact with Asp189 and form a nearly symmetric salt bridge in the S1 pocket. This step is essential because it leads to the correct alignment of the peptide substrate for bond breaking. However, lacking those scissile peptide bonds, the training set ligands inhibit the enzyme from endopeptidic cleavage [52, 53].

The free energy decomposition methodologies (M1–4) and the LISA variants were refit to reproduce the experimentally determined free energies of the trypsin inhibitor training set as described in the methods section. The fitting procedure resulted in the scaling parameters shown in Table 3. Each of the free energy decomposition methods predicted very high magnitudes of free energies, which were scaled to the range of the training set free energies by the small scaling parameters. Thus the raw values of M1–4 scores were of little use for predicting absolute binding affinities, and they needed to be scaled to be able to match experimental values. The combination of the scaling procedure and enthalpic error corrections greatly improved correlation with the training set as shown in Fig. 6.

The SAMPL3 test set

Visual inspection of the poses generated for the test set suggests that the carboxylate side chain of Asp189 at the base of the S1 pocket acts as the main H-bond acceptor forming H-bonds not only to benzamidine or primary amine groups as seen in the training set (see Figs. 5 and 7a),

Table 3 Fitting the energy models to the training set

Method	a	b	c	d	R^2
M1(s)[a]	0.0229	0.0266	−0.145	−3.636	0.675
M2(s)	0.0711	0.0712	−0.170	−2.112	0.735
M3(s)	0.0121	0.0134	−0.144	−5.135	0.459
M4(s)	0.0195	0.0222	−0.159	−4.695	0.493
LISA(s)	0.767	0.607	–	–	0.814
LISA(p)	0.901	0.429	–	–	0.751
LISA+ (s)	0.985	−0.873	–	–	0.810
LISA+ (p)	1.12	−0.769	–	–	0.880

Each scoring function was fit to a linear model to reproduce the experimentally determined free energies of binding of our trypsin binding training set, leading to these parameter sets

[a] The labels in parentheses refer to the set of ligand poses used to train the scoring function: *s* semiempirical (PM6-DH2 selected poses), *p* PDB structure

but also to hydroxyl groups and secondary amines as shown in Fig. 7b (SAMPL3 ligand numbers: 1, 14, 15, 16, 18, 19, 21, 23, 24).

Another significant H-bond acceptor in the S1 pocket is Ser190 which participates in the H-bonding network through different means depending on the ligand type: for the ligands having benzamidine or primary amino functionalities (ligand numbers 2, 4, 6, 8, 10–13, 17, 20, 22, 26–34), it establishes a H-bond through its side chain hydroxyl whereby this hydroxyl group is further stabilized by H-bonding to the deeply buried water molecules. The amidine groups are stabilized by the carbonyl oxygen of Gly219, as shown in Fig. 7a. On the other hand, for the rest of the test set ligands, which replace these amidine/primary amine groups with hydroxyl or secondary amine functionalities (Fig. 7b), the side chain of Ser190 loses its importance in the H-bonding network and its backbone carbonyl serves as the H-bond acceptor. The binding affinities calculated for those ligands did not show a clear trend. They covered the similar range of magnitudes as the ones associated with amidine/primary amino compounds, although experimentally the vast majority of these (except 15 and 16) were found not to bind.

All methods we explored predicted the SAMPL test set to be weaker binders than the ligands of the training set. This is consistent with the observation that the larger training set ligands fill multiple subpockets and should bind tighter due to the increased number of contacts with the receptor. The test set ligands, on the other hand, with molecular weights ranging from 158 to 242 g/mol (average molecular weight: 182 g/mol) filled only the S1 pocket without expanding to the rest of the subpockets. Thus, having fewer contacts, their binding affinities were predicted to be lower in magnitude.

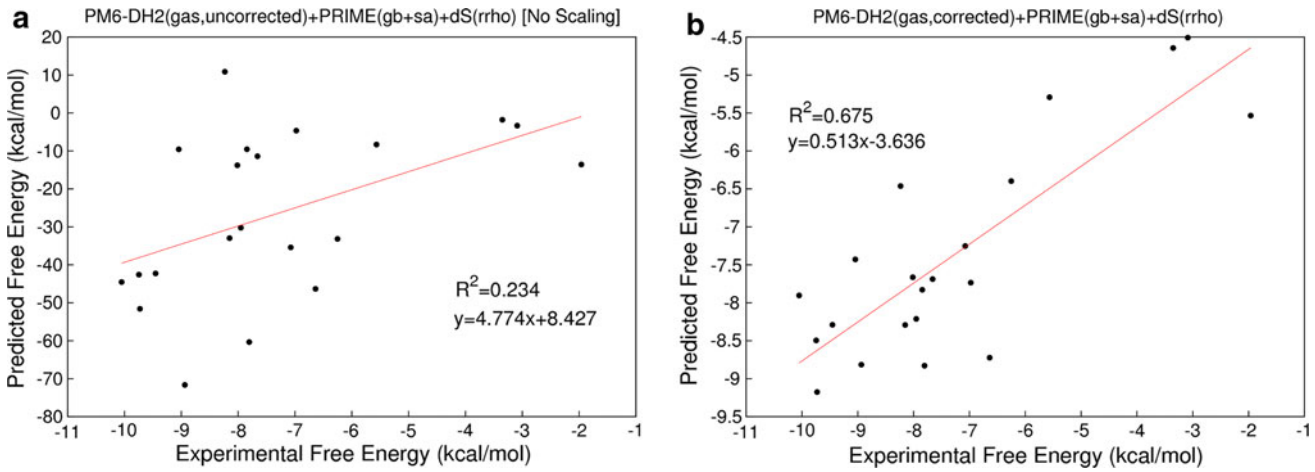

Fig. 6 **a** Raw M1 scores versus experimental free energies of the training set of trypsin inhibitors. **b** Scaled and enthalpy-corrected M1 scores versus experimental free energies of the training set

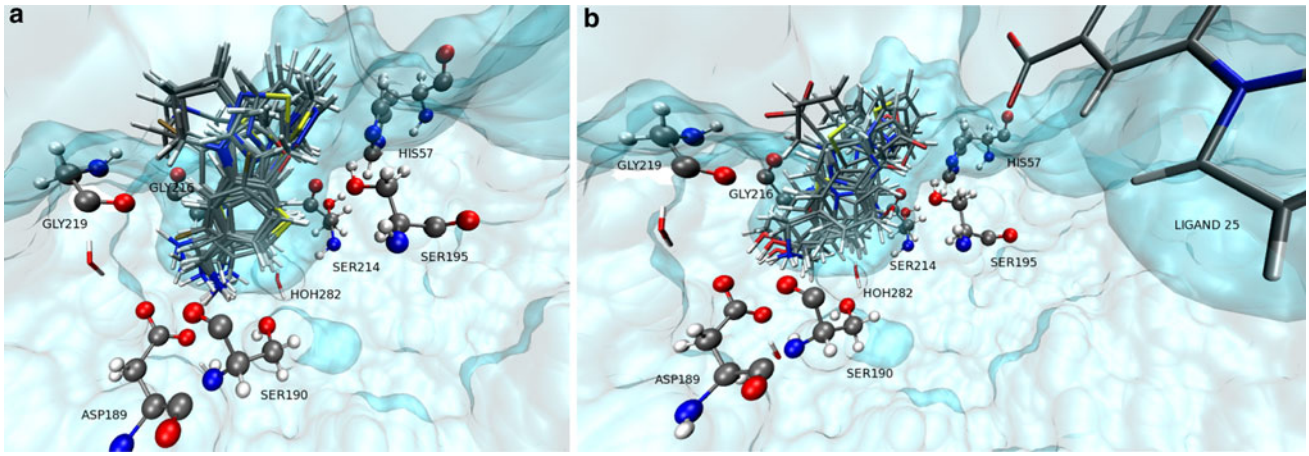

Fig. 7 **a** Subset of the SAMPL3 test set poses displaying interactions between an amidine or primary amine interacting with the Asp189 site. **b** The remainder of the test set poses lacking an amidine or primary amine group, but displaying other interactions with Asp189

The correlation coefficients between different methods and experiment are shown in Table 4, and mean absolute errors (MAE) are shown in Table 5. We expected that the free energy decomposition methods would correlate very closely since the only deviation in their construction was the solvation free energy term, but this was not the case. All methods yielded poor correlation with the experimental values for active binders in the test set, but yielded more acceptable MAE values. This is likely due to the small range of experimental binding affinities, and the errors inherent in each of the scoring methods. LISA and LISA+ were the only methods to reach the <1 kcal/mol MAE threshold, while the scaled M1, M3, and M4 methods had MAE values ranging 1–5. M2 performed the worst among the scaled methods, but this is understandable since its COSMO solvation free energy term neglected entropy contributions. M3 generally scored the best among the free energy decomposition methods, with its MM/PBSA solvent contribution. M3 performed slightly better when using poses selected by GlideXP score. In each of our methods, scaling our scores based on the training set of trypsin binders improved predictions on the SAMPL test set. The magnitude of improvement due to scaling is demonstrated specifically for M3 in Fig. 8. Scaling had the opposite effect on the Glide XP results, however, where the MAE grew dramatically upon scaling.

Part of the motivation behind our participation in this exercise was to attempt to improve predictions and place confidence intervals on them using the error estimation procedure discussed previously. Each of the free energy decomposition methods used error corrected enthalpies and was assigned error bars representing minimum

Table 4 Pearson correlation coefficients and standard errors between different methods used to score the SAMPL test set

	M1(s)	M2(s)	M3(s)	M4(s)	LISA(s)	LISA+(s)	GlideXP(s)	Experiment[a]
M1(s)	1	−0.35	0.53	0.49	0.13	−0.041	−0.12	0.21
M2(s)		1	−0.35	−0.35	−0.026	0.13 ± 0.01	0.58	0.12
M3(s)			1	0.98	−0.23	−0.25	−0.51	0.23
M4(s)				1	−0.19	−0.16	−0.54	0.25
LISA(s)					1	0.86	0.32	0.26
LISA+(s)						1	0.12 ± 0.01	0.34
GlideXP (s)							1	0.043

Standard errors are omitted when below 0.01. All free energy decomposition scores and LISA variants employ the scaling procedure

[a] Comparisons with experimental values only included ligands with known affinities (i.e. the 17 active binders)

Table 5 Bootstrapped mean absolute errors including standard error of different scoring methods with respect to the 17 SAMPL test set active binders

Method	MAE	Method	MAE
M1(s)-unscaled	31.0 ± 0.767	M1(s)-scaled	2.69 ± 0.00438
M2(s)-unscaled	20.2 ± 0.0381	M2(s)-scaled	4.46 ± 0.00578
M3(s)-unscaled	25.2 ± 0.211	M3(s)-scaled	1.90 ± 0.00556
M4(s)-unscaled	28.5 ± 0.155	M4(s)-scaled	2.45 ± 0.00686
M1(g)-unscaled	29.7 ± 0.0582	M1(g)-scaled	2.72 ± 0.00419
M2(g)-unscaled	157 ± 0.492	M2(g)-scaled	6.25 ± 0.0254
M3(g)-unscaled	9.65 ± 0.0542	M3(g)-scaled	1.77 ± 0.00489
M4(g)-unscaled	13.1 ± 0.0514	M4(g)-scaled	2.33 ± 0.00470
LISA(s)-unscaled	0.873 ± 0.00576	LISA(s)-scaled	0.775 ± 0.00344
LISA(s)+ -unscaled	1.72 ± 0.00694	LISA+ (s)-scaled	0.726 ± 0.00519
LISA(p)	0.810 ± 0.00551	LISA+ (p)	1.52 ± 0.00672
Glide-unscaled	1.18 ± 0.00514	Glide-scaled	33.8 ± 0.0516

Units are kcal/mol

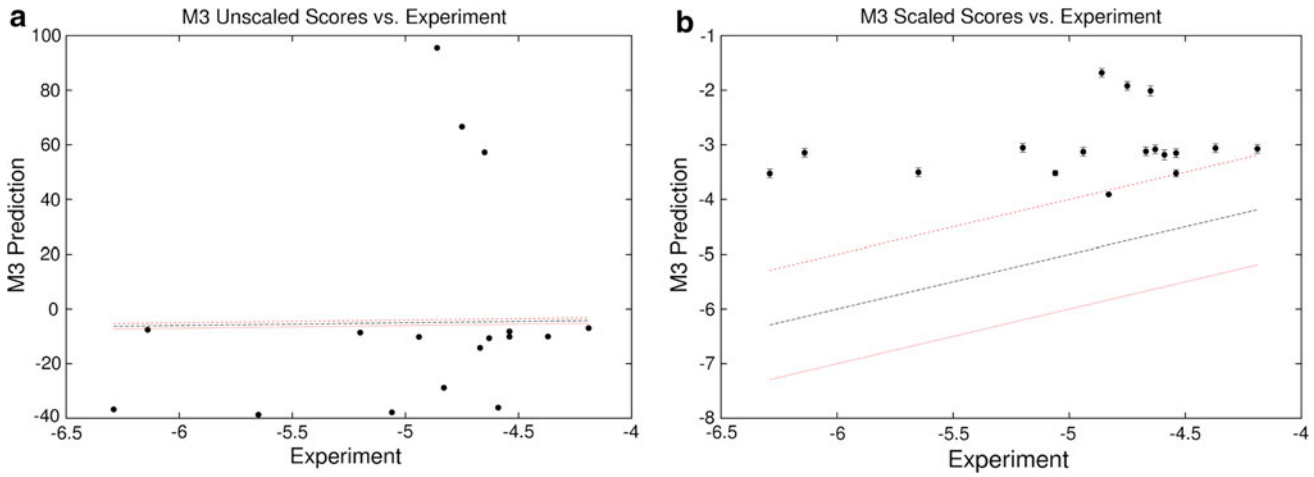

Fig. 8 **a** Unscaled M3 scores of the SAMPL3 test set versus experimental values. **b** Scaled M3 scores of the SAMPL3 test set. In both plots, the y = x ± 1 lines are drawn for reference. Unscaled scores spanned a large range of energies, which were scaled to a very narrow range around −3 kcal/mol in (**b**). Because of the scaling procedure, M3 scaled scores predicted most ligands to bind more weakly than the detection limit of the experiment. Furthermore, predicted enthalpy *error bars* were very small compared to overall predicted free energy magnitudes, and they become very small after the scaling procedure

uncertainties in free energy. These error estimations were of little benefit unfortunately, because the errors in the remaining terms drove predictions to very high magnitudes, making the relatively small error corrections have little effect on relative energies. The distributions of estimated systematic and random errors in enthalpy over

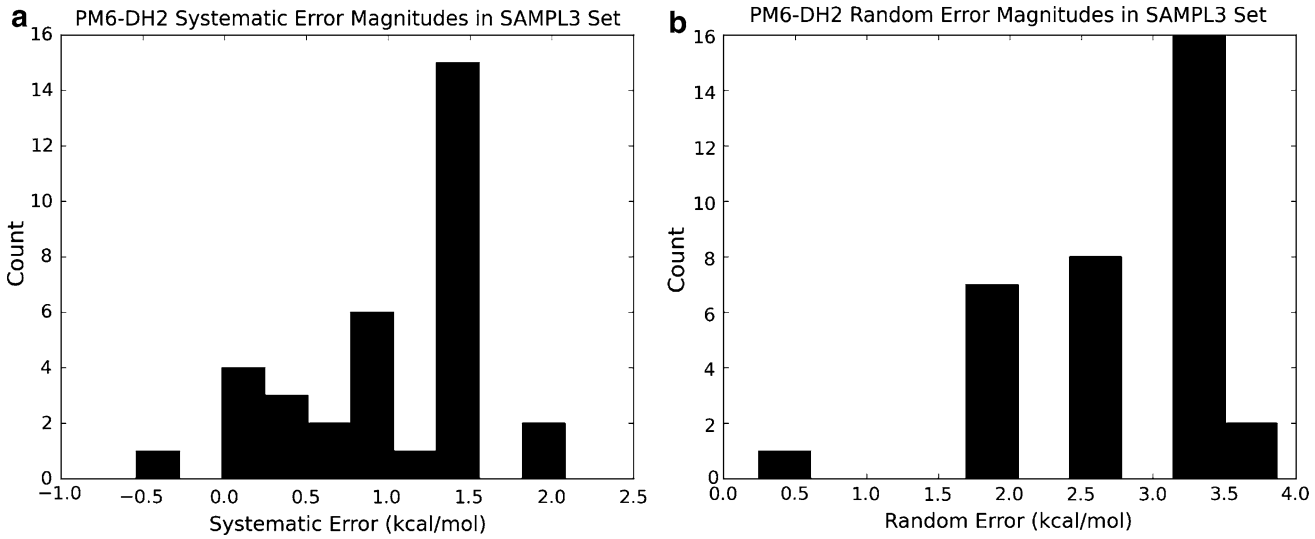

Fig. 9 **a** Distribution of estimated systematic errors in PM6-DH2 enthalpies over the SAMPL set poses. **b** Distribution of estimated random errors in PM6-DH2 enthalpies over the SAMPL set poses. Most systematic error corrections were below 2 kcal/mol, and these corrections became even smaller (in the 0.01–0.1 kcal/mol range) after applying the scaling procedure for M1–4

the SAMPL set ligands (PM6-DH2 selected poses) are shown as an example in Fig. 9. The systematic error correction magnitudes were <2.0 kcal/mol, and when used along with the scaling procedure, these became very small indeed (<0.1 kcal/mol). Thus the error corrections and uncertainty estimations were of little use when combining them with the scaling procedure. In order to improve error estimation in these models, additional error estimation of the entropy and solvation free energy components may be required, which would allow us to possibly forego the scaling procedure and more accurately correct systematic errors in free energy.

The scoring function with the best performance in this test was LISA refit with the training set. Figure 10 shows scatter plots of four of the LISA variants against the experimental binding affinities. LISA was able to predict the majority of binders within 1 kcal/mol range even before the rescaling procedure. LISA+, on the other hand, benefited greatly from the rescaling procedure as shown in Fig. 10c and d. LISA+ tended to predict binding affinities to be too strong, which was corrected by scaling its predictions based on the training set. Overall, the scaled LISA+ approach outperformed the rest of the methods employed in the SAMPL3 challenge by RMS error.

Conclusion

Of all our submissions the empirical score LISA+, refit to our training dataset, showed the best performance among all methods presented on the SAMPL3 test set as reflected by its mean absolute error of <1 kcal/mol. Using the training set to scale the results clearly improved all of the predictions made in this study. Having some knowledge about the binding attributes of some particular set of compounds provides a benchmark allowing for easier prediction of binding affinities.

A number of issues are highlighted from the analysis of our results. Predicting affinities using the methodologies of this study are clearly dependent on the poses used. The generation of reasonable poses for scoring is critical to making successful binding affinity predictions. Furthermore, using a single pose to represent the binding process may not be sufficient. Using an ensemble of poses may improve our results but this hypothesis will need to be tested in future work.

The results from the free energy decomposition methods indicate that much work is still needed to ensure the accurate prediction of binding affinities. The different solvation models resulted in significantly different predictions. The intricate interplay between enthalpy and entropy, and the error associated with each is not easy to fully understand—especially in light of entropy/enthalpy compensation that has been seen for ligands in a congeneric series [22, 51]. A better understanding of the errors associated with each calculation is essential considering the fortuitous cancellation of errors that is sometimes observed. Caution is recommended against placing too much confidence in calculations until one can accurately understand the accurate range of errors associated with each step of the calculation and how the error is propagated.

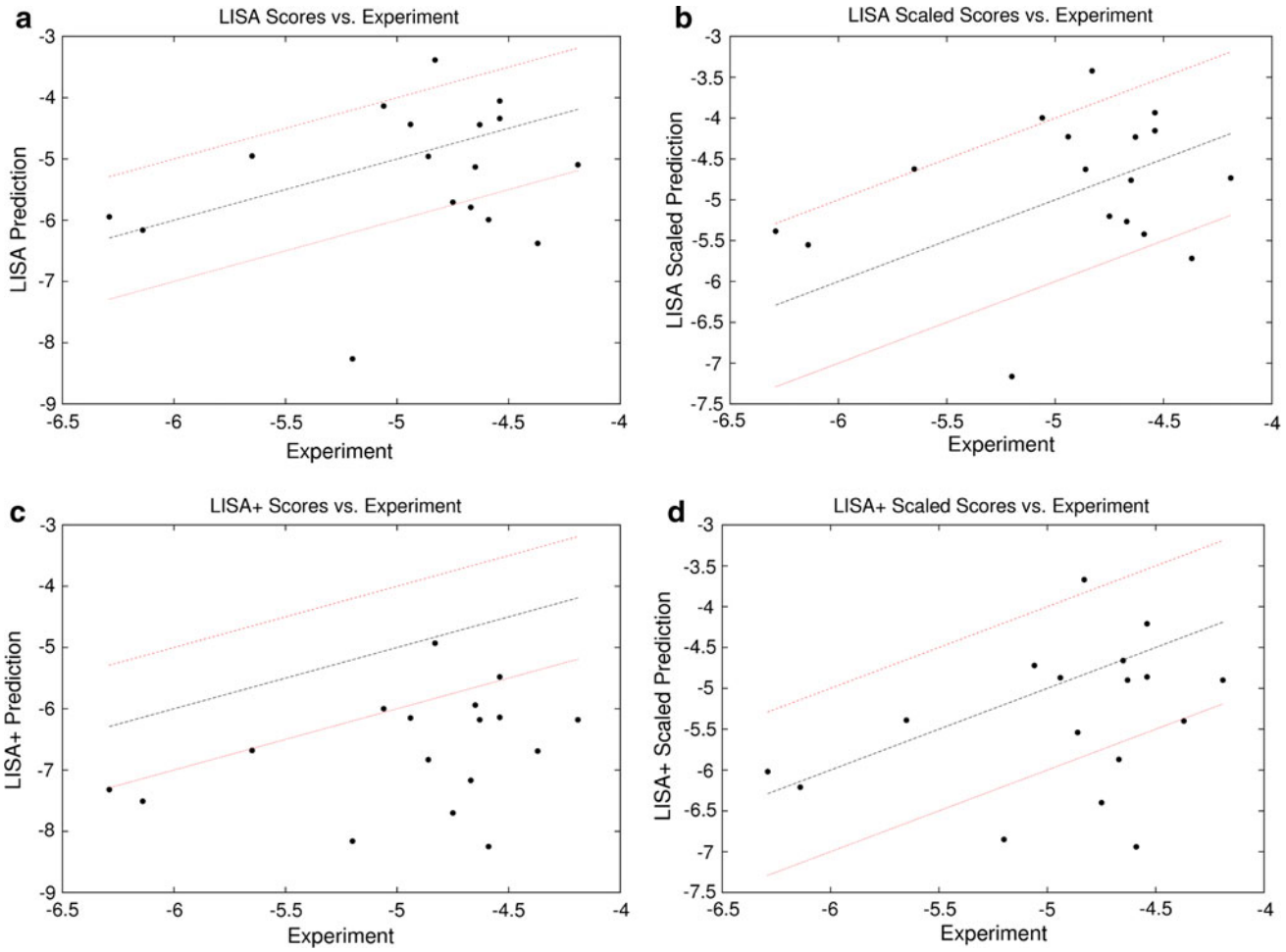

Fig. 10 Scatter plots of 4 LISA variants versus the SAMPL test set active binders. **a** LISA raw scores **b** LISA scaled scores **c** LISA+ raw scores. **d** LISA+ scaled scores

References

1. Andrusier N, Mashiach E, Nussinov R, Wolfson HJ (2008) Proteins Struct Func Bioinfo 73(2):271
2. Halperin I, Ma BY, Wolfson H, Nussinov R (2002) Proteins Struct Func Genet 47(4):409
3. Leach AR, Shoichet BK, Peishoff CE (2006) J Med Chem 49(20):5851
4. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) J Med Chem 49(20):5912
5. Kolb P, Irwin JJ (2009) Curr Top Med Chem 9(9):755
6. Deng YQ, Roux B (2009) J Phys Chem B 113(8):2234
7. Faver JC, Benson ML, He X, Roberts BP, Wang B, Marshall MS, Sherrill CD, Merz KM (2011) PLoS ONE 6(4):e18868
8. Faver JC, Benson ML, He X, Roberts BP, Wang B, Marshall MS, Kennedy MR, Sherrill DC, Merz KM (2011) J Chem Theor Comput 7(3):790
9. Merz KM (2010) J Chem Theor Comput 6(5):1769
10. Zheng Z, Merz KM (2011) J Chem Inf Model 51(6):1296
11. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, Carlson HA (2008) Nucleic Acids Res 36:D674
12. Hu LG, Benson ML, Smith RD, Lerner MG, Carlson HA (2005) Proteins Struct Func Bioinf 60(3):333
13. Glide. Version 5.7. New York, NY: Schrödinger, LLC; 2011
14. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) J Med Chem 47(7):1739
15. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT (2006) J Med Chem 49(21):6177
16. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) J Med Chem 47(7):1750
17. Park MS, Gao C, Stern HA (2011) Proteins Struct Func Bioinf 79(1):304
18. Stewart JJP (2008) MOPAC2009. Colorado. Stewart Computational Chemistry, Springs, CO, USA
19. Korth M, Pitonak M, Rezac J, Hobza P (2010) J Chem Theor Comput 6(1):344
20. Fanfrlik J, Bronowska AK, Rezac J, Prenosil O, Konvalinka J, Hobza P (2010) J Phys Chem B 114(39):12666
21. Ucisik MN, Dashti DS, Faver JC, Merz KM (2011) J Chem Phys 135:085101
22. Baum B, Muley L, Smolinski M, Heine A, Hangauer D, Klebe G (2010) J Mol Biol 397(4):1042
23. MacroModel. Version 9.9. New York, NY: Schrödinger, LLC; 2011

24. Prime. Version 3.0. New York, NY: Schrödinger, LLC; 2011
25. Jacobson MP, Friesner RA, Xiang ZX, Honig B (2002) J Mol Biol 320(3):597
26. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA (2004) Proteins Struct Func Bioinf 55(2):351
27. Klamt A, Schuurmann G (1993) J Chem Soc-Perkin Trans 2(5):799
28. Srinivasan J, Cheatham TE, Cieplak P, Kollman PA, Case DA (1998) J Am Chem Soc 120(37):9401
29. Massova I, Kollman PA (1999) J Am Chem Soc 121(36):8133
30. Kollman PA, Massova I, Reyes C, Kuhn B, Huo SH, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE (2000) Account Chem Res 33(12):889
31. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Proteins Struct Func Bioinfo 65(3):712
32. Wang RX, Fang XL, Lu YP, Wang SM (2004) J Med Chem 47(12):2977
33. LigPrep. Version 2.5. New York, NY: Schrödinger, LLC; 2011
34. Katz BA, Elrod K, Verner E, Mackman RL, Luong C, Shrader WD, Sendzik M, Spencer JR, Sprengeler PA, Kolesnikov A, Tai VWF, Hui HC, Breitenbucher G, Allen D, Janc JW (2003) J Mol Biol 329(1):93
35. Cui J, Marankan F, Fu WT, Crich D, Mesecar A, Johnson ME (2002) Bioorg Med Chem 10(1):41
36. Whitlow M, Arnaiz DO, Buckman BO, Davey DD, Griedel B, Guilford WJ, Koovakkat SK, Liang A, Mohan R, Phillips GB, Seto M, Shaw KJ, Xu W, Zhao ZC, Light DR, Morrissey MM (1999) Acta Crystallogr Sect D-Biol Crystallogr 55:1395
37. Toyota E, Ng KKS, Sekizaki H, Itoh K, Tanizawa K, James MNG (2001) J Mol Biol 305(3):471
38. Fokkens J, Klebe G (2006) Angewandte Chem Int Ed 45(6):985
39. Presnell SR, Patil GS, Mura C, Jude KM, Conley JM, Bertrand JA, Kam CM, Powers JC, Williams LD (1998) Biochemistry 37(48): 17068
40. Katz BA, Mackman R, Luong C, Radika K, Martelli A, Sprengeler PA, Wang J, Chan HD, Wong L (2000) Chem Biol 7(4):299
41. Leiros HKS, Brandsdal BO, Andersen OA, Os V, Leiros I, Helland R, Otlewski J, Willassen NP, Smalas AO (2004) Protein Sci 13(4):1056
42. Dullweber F, Stubbs MT, Musil D, Sturzebecher J, Klebe G (2001) J Mol Biol 313(3):593
43. Maignan S, Guilloteau JP, Pouzieux S, Choi-Sledeski YM, Becker MR, Klein SI, Ewing WR, Pauls HW, Spada AP, Mikol V (2000) J Med Chem 43(17):3226
44. Di Fenza A, Heine A, Koert U, Klebe G (2007) ChemMedChem 2(3):297
45. Nar H, Bauer M, Schmid A, Stassen JM, Wienen W, Priepke HWM, Kauffmann IK, Ries UJ, Hauel NH (2001) Structure 9(1):29
46. Yusuf D, Davis AM, Kleywegt GJ, Schmitt S (2008) J Chem Inf Model 48(7):1411
47. Baber JC, Thompson DC, Cross JB, Humblet C (2009) J Chem Inf Model 49(8):1889
48. Weininger D (1988) J Chem Inf Comput Sci 28(1):31
49. Weininger D, Weininger A, Weininger JL (1989) J Chem Inf Comput Sci 29(2):97
50. Sadowski J, Gasteiger J, Klebe G (1994) J Chem Inf Comput Sci 34(4):1000
51. Brandt T, Holzmann N, Muley L, Khayat M, Wegscheid-Gerlach C, Baum B, Heine A, Hangauer D, Klebe G (2011) J Mol Biol 405(5):1170
52. Creighton TE (1984) Proteins: structure and molecular properties. Freeman and Company, New York, NY
53. Hubbard RE (2006) Structure-based drug discovery: an overview. Royal Society of Chemistry, Cambridge