JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

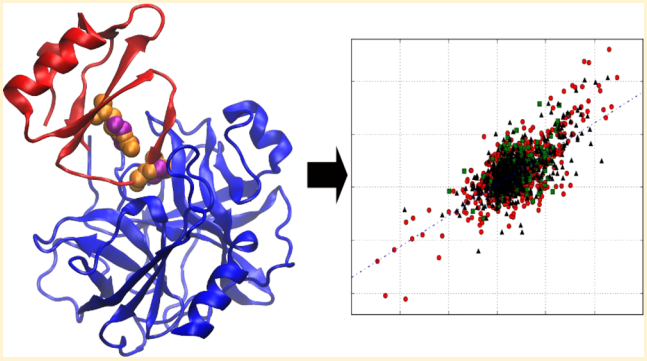
Intermolecular Contact Potentials for Protein−Protein Interactions Extracted from Binding Free Energy Changes upon Mutation

Iain H. Moal and Juan Fernandez-Recio*

Joint BSC-IRB Research Program in Computational Biology, Life Science Department, Barcelona Supercomputing Center, C/Jordi Girona 29, 08034 Barcelona, Spain

Supporting Information

ABSTRACT: Understanding and predicting the energetics of protein−protein interactions is fundamental to the structural modeling of protein complexes. Binding free energy can be approximated as a sum of pairwise atomic or residue contact energies, which are commonly inferred from contact frequencies observed in experimental protein structures. However, such statistically inferred potentials require certain assumptions and approximation. Here, we explore the possibility of deriving atomic and residue contact potentials directly from experimental binding free energy changes following mutation and present a number of such potentials. The first set of potentials is obtained by unweighted least-squares fitting and bootsrap aggregating. The second set is calculated using a weighting scheme optimized against absolute binding affinity data, so as to account for the over-representation of certain complexes, residues, and families of interactions. The congruence of the potentials with known physical chemistry is investigated. The potentials are further validated by ranking and clustering protein−protein docking poses.



1. INTRODUCTION

The physical interactions between proteins is one of the indispensable elements underlying biological functions, from signal transduction and enzyme regulation to the assembly of macromolecular machines. Thus, understanding the structural and thermodynamic principles that govern them has been a highly active area of research. Indeed, binding energy estimation is central to many fields, including complex structure prediction,[1] interface engineering,[2] peptide therapeutic design,[3−5] and locating binding hot-spots and surface motifs for rational drug design.[6−8] Determining exactly the energetics of complex systems, for instance from first principles, is extremely costly. However, empirical approaches can be used to drastically reduce computational expense. One highly popular approach is through statistical potentials. Since their inception, these have found many uses in the field of computational structural biology, from protein modeling,[9−13] design,[14,15] and docking[1,16,17] to the prediction of binding affinities,[18−23] protein stabilities, and how these can change upon mutation[24−31] and identifying hotspots residues,[32−34] the estimation of binding rate constants,[35] and interface design.[36] Although many different statistical potentials have been developed, most are based upon the same fundamental principles.[37,38] Specifically, the potential takes the form

$$P_{i,j} = -\log(N_{\mathrm{obs}i,j}/N_{\mathrm{exp}\ i,j}) \quad (1)$$

where $P_{i,j}$ is the potential for interacting pairs i and j, $N_{\mathrm{obs}i,j}$ is the number of such interactions observed in a database of solved structures, and $N_{\mathrm{exp}i,j}$ is the expected number of interactions assuming that i,j interactions are neither favored or disfavored. By applying the Boltzmann principle to eq 1, the log ratios can be expressed as potential energies. However, when interpreting these potentials, a number of approximations and assumptions must be invoked. For instance, in the scheme given, pairwise additivity is assumed. Additionally, when calculating N_{exp}, the way in which the reference state is defined can profoundly influence the derived potential and many different reference states have been proposed. Further, it has been shown that these derived potentials do not correspond to real energies,[39] even if, in practice, they can be used for structural modeling. For instance, Thomas and Dill showed that statistical potentials extracted from lattice models failed to reproduce the energies used to construct those models,[40] even though the strengths of pair interactions could be correctly ordered. In the study, they showed that the derived energies can be influenced by factors such as protein size. Statistical potentials may also be susceptible to systematic biases in the crystal structures of the proteins used to train them; while the atomic arrangements are global energy minima, they have also been shaped by evolutionary forces. Further, the proteins used are only those amenable to crystallization, and thus, they may be biased toward configurations that lend to binding or folding kinetics suitable for crystal formation. For instance, contacts that have higher prevalence in the database due to their enhancement of folding kinetics can

Received: January 29, 2013
Published: June 18, 2013

   dx.doi.org/10.1021/ct400295z | J. Chem. Theory Comput. 2013, 9, 3715−3727

have lower apparent energies. These biases can be avoided by fitting directly to experimental affinities. Therefore, new general empirical methodologies for binding energetics that are not optimized for a particular type of problem need to be developed.

We recently published SKEMPI, a large literature-derived database of 3047 experimental binding affinity changes upon mutation.[41] In the current paper, we present atomic and residue level contact potentials extracted from this database. While previous work has related energetic terms to mutation data,[42,43] this is, to the authors' knowledge, the first time such an approach has been undertaken to directly obtain energy functions. This is done by first modeling the mutant structures and calculating the change in number of *i,j* interacting pairs for each mutant. In the second step, we fit the potential, $P_{i,j}$, to the experimental $\Delta\Delta G$ values using linear least-squares multiple regression, under the assumption that $\Delta\Delta G$ can be approximated as a summation of $P_{i,j}$ terms, that is, using $P_{i,j}$ as independent variables and the change in number of contacts as basis functions. Bootstrap aggregating was employed to smooth out anomalies which may arise from sensitivity to the training data. Although pairwise additivity is still assumed, by fitting directly to energies, our approach circumnavigates some of the problem associated with potentials derived from the statistics of structural databases. However, it is susceptible to other sources of bias. For instance, mutations can alter binding affinity via structural rearrangements rather than direct changes in contacts. For instance, forced changes in backbone dihedral angles following mutation to proline may alter the overall complementarity of the interface. However, in the fitted potential, this lower affinity would be attributed to the presence of new proline contacts, artificially disfavoring such contacts. Similar biases could result from mutations in the interior of the protein, or from mutants with many mutations. To alleviate this, mutations to and from proline, and mutations in the interior of the protein, are removed from the training set, as are mutants with over three mutations in the binding site. Further, the fitting may be susceptible to biases toward over-represented classes of mutant in the training data, such as over-represented residues, binding sites, proteins, or protein families. To account for these sources of bias, we calculated a second set of potentials using a weighting scheme in which $\Delta\Delta G$ values associated with these over-represented classes are given less weight in the fitting, and the under-weighting is commensurate to the extent of over-representation. This weighting scheme is optimized against a diverse and nonredundant set of ΔG values, as well as $\Delta\Delta G$ error estimates derived from the bootstrap aggregating procedure. The final set of potentials was derived from the models that give an optimal balance between ΔG and $\Delta\Delta G$ performance, by averaging over the models of the Pareto frontier. The potentials are validated by comparison with the known physical chemistry of amino acids, by their ability to reproduce ΔG values and by their ability to rank docked poses.

2. METHODOLOGY

2.1. Selecting Mutants for Training. The $\Delta\Delta G$ data used to train the potentials are taken from the SKEMPI database,[41] against which a number of filters were used to remove mutants. First, the mutations for which the affinities were measured with unusual methods were removed (the mutations in 2JEL and the apparent affinity values of 1DAN). Further, mutants were filtered according to their location as determined using the scheme proposed by Levy;[44] mutants with a mutation in the interior (completely buried in the monomer) or support region (mostly buried in the monomer and completely buried upon binding) were removed, as were mutants with more than three mutations in the core (mostly exposed in the monomer and fully buried upon binding) and rim regions (mostly exposed in the monomer and partially buried upon binding). Mutants with mutations to or from proline were also removed. To alleviate redundancy, affinities that had been measured by multiple groups or experimental methods were averaged. After this filtering was applied, 1949 mutations remained, in 82 complexes. These mutations are shown in Supporting Information Table 1.

2.2. Calculating the Potentials. Here, we define two potentials. The first is a residue-level potential, where residues are defined as in contact if the geometrical centers of their side-chains (or C_α for glycine) are within 6.5 Å of one another. This value is chosen so as to encompass the first peak in the radial distribution around the residues.[37] The second potential is atomic-level. The most commonly used atom type definitions used in the literature consists of 18 atom types.[18] These are not adopted here, as they resulted in singular matrices when fitting; so instead, we use a reduced set of 12 atom types.[16] These atom types are defined such that they optimally distinguish between different protein environments.[45] We use the original atom type nomenclature with one correction; in the original type list,[16] the type containing atoms in the glutamic acid and aspartic acid side chains was erroneously labeled DE_mc, while the type containing the main chain atoms of these amino acids was labeled DE-. Thus, we have swapped these labels so that the suffixes reflect their constituent atoms. Two atoms are deemed to be in contact if they are within 6.0 Å of one another. This cutoff value was chosen because it was used in the construction of the atom type definitions.[16,45] Although quite large, it is commonly used in the literature and justified because it (1) includes all first-order contacts, (2) is the distance at which the two atoms exclude water, that is, 2.8 Å water diameter plus two heavy atom VDW radii (around 1.5−1.7 Å), and (3) satisfies the smooth density assumption, that is, is the radius at which the average number of contacts is comparable with a theoretical value estimated from a model of tightly packed atoms with average atomic volume.[18] For all the selected mutants, side-chains were remodelled using FoldX 3b5.1.[46] All intermolecular contacts were calculated for the native and the mutant structures, and the change in number of contacts for each *i,j* pair, $N_{i,j}$, was determined and collected into matrix **N**. The binding free energy changes were collected in the vector **E**. It was assumed that these changes could be expressed as a sum of energetic contributions from *m* contact types, given in vector **P**, plus residuals **r**.

$$E = \sum_{k=1}^{m} P_k N_k + r \tag{2}$$

We use weighted linear least-squares regression to determine the optimal *P* values, $\hat{P}$, which minimize the sum square residuals, as shown in eq 3. A full derivation of this equation is given in the Supplementary Methods section of the Supporting Information.

$$\hat{P} = (N^T W N)^{-1} N^T W E \tag{3}$$

Here, *W* is the diagonal weight matrix. In the first two potentials that we present (the unweighted residue potential and the unweighted atomic potential), the identity matrix is used; that is, the diagonal elements of this matrix are set to unity to give equal weight to all mutants. In the second set of

potentials (the weighted residue potential and the weighted atomic potential), the weights are set to compensate for the biases in the training data, as described below. As small quantities of data for some of the contacts can result in anomalies due to regression instability, we used the bootstrap aggregating (bagging) algorithm;[47] instead of fitting to all the $\Delta\Delta G$ data, we resample the training data 100 times using bootstrap sampling, derive the potential for each sample, and report the average potential. This popular resampling method is known to improve the accuracy and stability of regression models by "smoothing out" anomalous values caused by sensitivity to small variations in the training set. This also allows us to derive an error estimate. In each bootstrap sample, about 37% of the data is left out—the out-of-bag mutants. The potential calculated using that sample can be used to make blind predictions for these omitted mutants. By averaging the out-of-bag predictions for each sample, it is possible to use these predictions to estimate the error of the model, referred to as the out-of-bag root-mean-square error (RMSE).

2.3. Weighting the Training Data. The filtered data set used to train the potentials contains a number of biases, and a weighting scheme was devised in which over-represented classes of mutation are down-weighted. The weighting scheme was chosen to account for three sources of bias associated with the greatest sources of over-representation in the mutation database:

- Biases toward specific families of interaction. For instance, mutations in the serine protease/inhibitor family of interactions are highly over-represented, with 909 mutations, as are mutations in antibody−antigen complexes, with 320 mutations.
- Biases toward specific complexes. Some complexes contain many more mutations than others, such as the BLIP/TEM1 β-lactamase interaction, with 67 mutations.
- Biases toward specific residues. Within some complexes, certain residues are over-represented, such as the leucine residue in the P1 position of the interaction of OMTKY3 with *Streptomyces griseus* proteinase B. For this residue, $\Delta\Delta G$ values are available for mutations to all 19 other amino acids. Further, as crystal structures are available for 16 of these P1 mutants and the reverse mutations are also included in the database, a total of 35 $\Delta\Delta G$ values relate to this specific residue alone.

In doing so, a number of considerations were made, the first of which was simplicity. Sophisticated schemes could be devised where, for instance, two mutants with a mutation in the residue are more downweighted if they are both mutations from small amino acids to large amino acids or mutations in corresponding residues of two homologous interactions are more downweighted than noncorresponding residues. However, to avoid overly complicated schemes with many free parameters, we downweight mutants only based on the number of other mutants with mutations in the same residue, the same complex, or homologous complexes. Second, it was desired that the down-weighting of over-represented classes of mutation be commensurate with the extent of over-representation. Third, the mutual information contained in two mutations in the same residue was deemed to be different from that in two mutations in a different residue of the same complex, in turn different from the mutual information in two mutations in two homologous complexes. Thus, it was desired that each source of bias have a different magnitude of down-weighting. Lastly, we wanted to have a baseline weight, from which the above three sources of bias were detracted. In light of these considerations, the following scheme was devised.

$$w_i = \frac{1}{1 + xr_i + yc_i + zh_i} \tag{4}$$

The weight of mutant i is determined by r_i, the number of other mutants with mutations in the same residues of the same complex, c_i, the numbers of other mutants with mutations in other residues of the same complex, and h_i, the number of mutations in homologous interactions. The factors x, y, and z are parameters to be optimized. When calculating h_i, two interactions were deemed to be homologous, as described previously.[41] Briefly, two interactions are homologous if they share a mutual binding partner or homologous binding partner and also share 70% of interface residues. When calculating r_i, aligned residues in different PDB files of the same complex count as the same residue, as determined by sequence alignment using GAP4.[48] For mutants with multiple mutations, r_i and c_i are calculated as the mean r and c for each individual residue that is being mutated. Similarly, contributions to r_i and c_i of other mutants are attributed so that each residue contributes equally and the contributions for each mutant sum to one. For instance, if a complex has two mutants, the first a single mutant, and the second a double mutant of the same residue and another residue, then $r_1 = c_1 = r_2 = c_2 = 0.5$, or if complex has two double mutations which have one residue in common, then $r_1 = r_2 = 0.25$, and $c_1 = c_2 = 0.75$. In a slightly more complicated example, if a complex has three mutants, the first a single mutant, the second a double mutant with a residue in common with the first mutant and the other residue in common with the third mutant, and the third mutant is a double mutation, then $r_1 = 0.5$, $c_1 = 1.5$, $r_2 = 0.75$, $c_2 = 1.25$, $r_3 = 0.25$, and $c_3 = 1.75$. This example illustrates how this method of calculating c_i and r_i gives values that intuitively reflect the amount of mutual information in each set of $\Delta\Delta G$ experiments. All homology groupings and r_i, c_i, and h_i values are given in Supporting Information Table 1.

2.4. Deriving the Weighted Potentials. In the scheme given in eq 4, three parameters require adjustment. These values were sampled uniformly from 0 to 50 in jumps of 5, yielding 1331 combinations of x, y, and z. For each combination, the potential was calculated by bagging eq 3 and used to predict the 144 absolute binding affinities described in the binding affinity benchmark.[49] We also report predictions for the subset of 57 complexes for which the affinities are known with high confidence. These complexes are those for which similar affinities had been determined by multiple groups or multiple experimental methods.[23] The affinity benchmark is also subdivided into 71 rigid (IRMSD < 1.0 Å) and 73 flexible (IRMSD > 1.0 Å) complexes. For each of the x,y,z combinations, we plot the Pearson correlation between the predicted and the experimental ΔG values versus the $\Delta\Delta G$ out-of-bag RMS error. This plot is used to identify the models on the Pareto frontier, that is, those for which no other model can achieve a better ΔG performance without also incurring a lower $\Delta\Delta G$ performance and *vice versa*. The final weighted potentials were found by averaging the contact energies for all the models on the Pareto front.

2.5. Ranking Docked Poses. The ability to ability to dock protein complexes is a long-standing problem in structural bioinformatics and is frequently assessed in CAPRI, a series of community wide events which provide a blind test for docking algorithms. The four contact potentials were further evaluated for their ability to identify the 3D structure of protein complexes

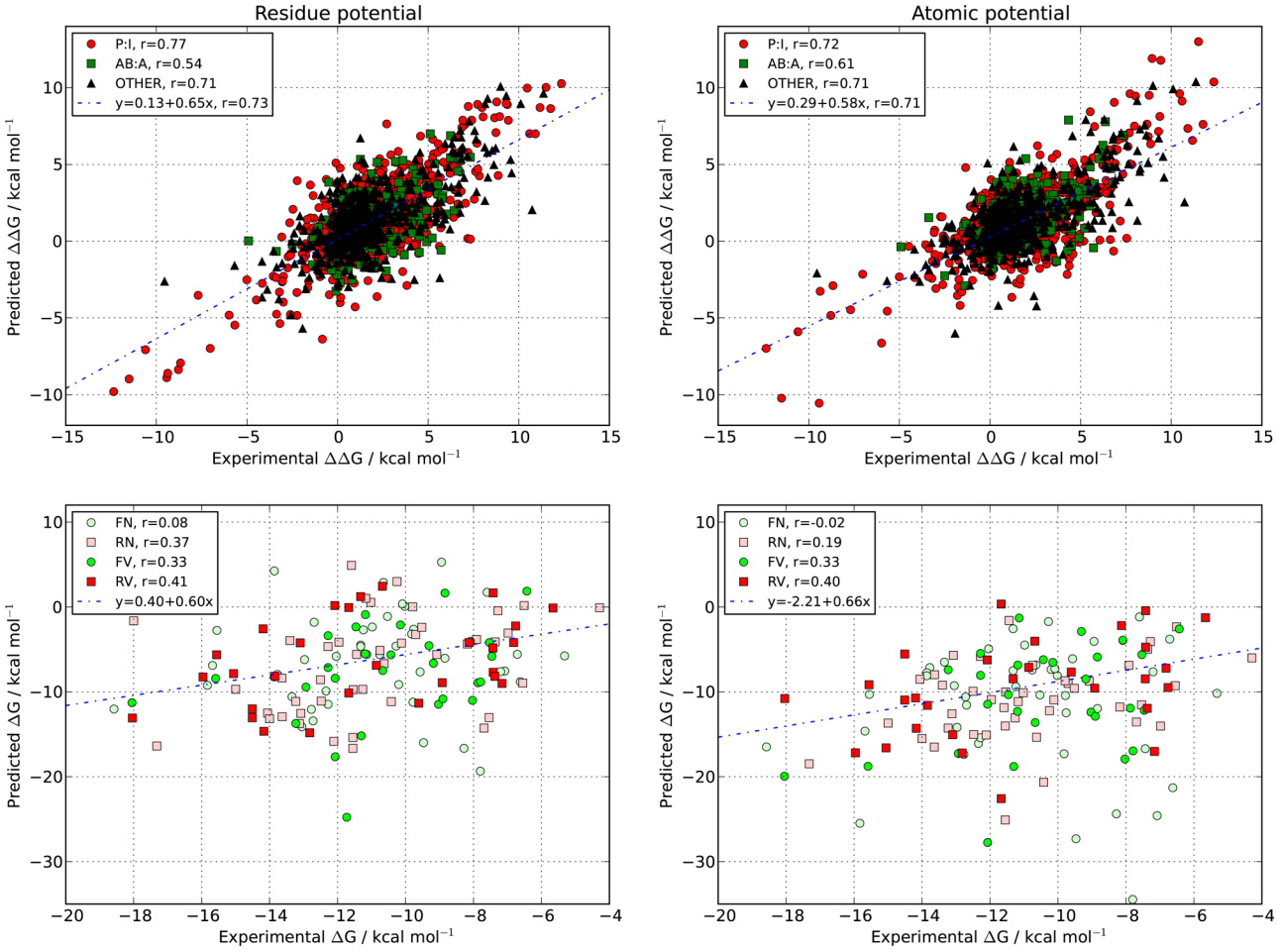


Figure 1. Modeled versus experimental $\Delta\Delta G$ and ΔG values for the unweighted residue and unweighted atomic potential. For the $\Delta\Delta G$ data, out-of-bag predictions are shown and the data are subdivided into mutations in serine protease inhibitor complexes (P:I), antibody−antigen complexes (AB:A) and other complexes (OTHER), with Pearson correlations reported for each subset. The line of best fit is shown, with the slope, intercept and overall correlation. The ΔG data are subdivided into rigid (R) and flexible (F) complexes, as well as validated (V) and nonvalidated (N) affinities, with correlations shown. The line of best fit, slope and intercept are given for the affinities in the validated rigid set. An outlier of very low predicted energy is omitted from atomic potential ΔG plot.

from the coordinates of their unbound constituents using a computational docking algorithm. Specifically, their ability to identify correctly docked structures generated using the SwarmDock algorithm from amidst incorrect docking decoys generated using the same method.[50−52] The decoy poses used are those that had been generated previously,[52] in which the algorithm was applied to all complexes in the protein−protein docking Benchmark 4.0.[53] This recent benchmark consists of 176 nonredundant complexes of diverse affinity, size, and conformational flexibility. For each complex, between 406 and 536 poses were generated. After scoring, the decoys were clustered at 3 Å resolution as per the standard SwarmDock protocol.[50,51] For each cluster, all but the lowest energy member were removed, leaving a ranked list of docked poses. For each of these poses, the standard CAPRI metrics were calculated:[54] interface RMSD (the RMSD of interface residues), ligand RMSD (the RMSD of the ligand after superimposing the receptors, chosen as the larger of the two subunits), and the fraction of native contacts. These were then used to classify the solutions as incorrect, acceptable, medium, or high quality. The standard CAPRI thresholds were used,[55] with structures assigned in descending order of quality. Thus, poses are deemed incorrect if fewer than 30% of the contacts are also present in the crystal structures of the complex, if the interface RMSD is greater than 4 Å, and if the ligand RMSD is greater than 10 Å.

3. RESULTS AND DISCUSSION

3.1. Unweighted Potentials Capture the Physical Chemistry of Most Amino Acid Interactions. *3.1.1. Unweighted Residue Potential.* The performance of the unweighted residue potential, when applied to the $\Delta\Delta G$ data, is shown in the top left corner of Figure 1 and shows a strong correlation (r = 0.73). These results are out-of-bag predictions, in which the model generated from each bootstrap sample was used to predict the $\Delta\Delta G$ of mutants not in that sample, and these out-of-bag predictions were averaged. Although these are "blind" predictions in this sense, it is still the case that often mutations in the same residue, interface, or homologoues interface are present in both the bootstrap sample and the out-of-bag mutants. As such, it cannot be used directly to estimate the generalization error. Ideally, the performance would be evaluated using a homology-corrected leave-complex-out cross-validation protocol, in which mutations in the same or homologous interfaces are simultaneously held out for validation.[41] However, doing so proved insoluble due to degenerate matrices, and thus, we instead resort to validating the potentials by their plausibility

Table 1. Unweighted Residue-Level Contact Potential[a]

	acidic		basic			hydrophobic									polar			special		
	ASP	GLU	HIS	LYS	ARG	ALA	VAL	LEU	ILE	PHE	TRP	MET	TYR	SER	THR	ASN	GLN	CYS	GLY	PRO
ASP	−0.049																			
GLU	−0.627	−0.098																		
HIS	−1.782	−1.535	−0.176																	
LYS	−1.568	−0.964	0.893	2.379																
ARG	−1.383	−1.426	0.231	0.494	0.568															
ALA	0.281	−0.021	0.490	0.666	0.382	0.856														
VAL	0.290	0.225	0.067	−0.433	0.191	0.499	−1.344													
LEU	0.065	−0.024	0.217	0.931	0.583	−0.293	−1.279	0.503												
ILE	1.546	2.919	−0.471	−0.416	0.953	−0.051	−1.309	−0.316	−1.016											
PHE	−0.051	−0.007	0.469	−0.143	−0.808	−0.617	−1.052	−0.422	−0.403	0.714										
TRP	0.279	−1.054	−0.052	−2.053	−1.674	−0.542	−0.041	−0.350	−0.301	−0.101	−0.935									
MET	0.231	1.478	−0.361	2.382	0.471	−0.477	−1.007	−1.489	0.247	0.068	−2.818	−1.241								
TYR	−0.722	0.099	−1.852	−1.073	−0.356	0.572	0.439	−0.016	−2.214	−1.057	−0.470	0.036	−0.144							
SER	0.220	−0.257	−0.202	0.188	0.363	−0.075	0.107	−0.140	0.002	−0.202	−0.905	−0.098	−0.173	0.299						
THR	0.751	0.263	−1.248	−2.316	−0.525	−0.773	−0.678	−0.381	−0.851	0.857	−0.564	0.400	−1.559	−0.861	0.171					
ASN	−0.192	−0.171	−0.486	0.740	0.108	0.825	0.123	0.574	0.166	−0.123	−1.374	−0.108	−0.585	0.030	−0.078	−0.059				
GLN	−0.600	0.359	0.131	−2.087	−0.158	0.378	0.859	−0.564	−0.057	0.670	−0.322	0.021	0.590	−0.763	0.026	−0.325	1.029			
CYS	0.487	−0.290	−0.092	−2.206	−0.054	0.248	0.150	−0.249	−0.406	0.001	0.325	−0.604	−0.181	0.015	0.649	−0.116	0.043	−0.534		
GLY	0.408	0.661	0.211	−0.458	0.019	−0.340	0.026	−0.794	0.380	−0.796	0.194	−0.150	−0.920	−0.108	0.679	−0.370	0.568	−0.275	1.165	
PRO	−0.076	1.579	−0.561	1.572	0.430	1.514	−0.244	−0.082	0.509	−0.108	−0.209	−1.260	−0.528	0.470	0.248	0.059	0.973	−0.229	−1.933	N/A

[a]Note that the proline−proline contact energy is omitted due to filtering of the training data. Values are shown in kcal/mol.

dx.doi.org/10.1021/ct400295z | J. Chem. Theory Comput. 2013, 9, 3715−3727

Table 2. Unweighted Atomic Contact Potential[a]

	-ve		+ve			hydrophobic		aromatic		polar		
	DE_mc	DE-	R+	K+	RHK_mc	ILV_sc	AILMV_mc	MFW_sc	WY_sc	polar	mc	CG
DE_mc-	0.120											
DE-	−0.069	−0.003										
R+	−0.011	−0.077	0.040									
K+	0.041	−0.124	−0.023	0.390								
RHK_mc	0.002	−0.077	0.014	0.053	0.065							
ILV_sc	−0.041	0.014	−0.004	0.009	0.054	−0.034						
AILMV_mc	−0.020	0.022	0.014	0.030	−0.015	−0.014	0.048					
MFW_sc	−0.008	0.016	−0.052	0.014	0.034	−0.003	−0.030	0.014				
WY_sc	−0.038	−0.011	−0.014	−0.081	−0.044	−0.023	−0.015	−0.048	0.006			
polar	0.004	−0.022	−0.006	−0.021	−0.031	0.003	−0.004	−0.008	−0.044	−0.005		
mc	0.043	0.032	0.024	0.054	−0.030	−0.008	0.040	−0.007	0.016	0.001	−0.016	
CG	−0.078	0.105	−0.011	−0.100	0.044	−0.061	−0.032	−0.012	0.038	0.006	−0.046	0.028

[a]Values are shown in kcal/mol.

and their ability to predict absolute binding affinities and to rank docked poses. The ability of this potential to predict ΔG, the free energy of binding, for a range of protein−protein complexes, is shown in the bottom left corner of Figure 1. This gives a low but significant overall correlation of 0.28 ($p < 0.001$). However, the correlation for the cases in the validated set, the complexes for which the affinity has been determined by multiple groups or experimental methods, is higher than that for the nonvalidated set, as expected (0.35, versus 0.23). Similarly, the rigid complexes have a higher correlation than for the flexible complexes (0.38, versus 0.17), and the highest correlation of all is observed for the complexes in the intersections of the validated and the rigid set ($r = 0.41$). This latter correlation should be seen as the most indicative of accuracy, as it corresponds to the most experimentally accurate data and is focused on the complexes for which unaccounted conformational energy change is least likely to add noise.

The contact energies for the potential are shown in Table 1. Many of the values are intuitively in accordance with known physical chemistry of the amino acids. For instance, with the exception of the histidine−histidine interaction, all of the contact energies between basic amino acids are positive. The lysine−lysine interaction is highly disfavored (2.4 kcal/mol), corresponding to the coupling of the localized charges on the ε-amino groups. In the arginine−arginine interaction, the uniting of the more polarizable and diffusely charged guanidinium moieties is, correspondingly, less disfavored (0.6 kcal/mol). Further, the interactions between the polar and basic amino acids are mostly favorable (up to −2.3 kcal/mol), resulting from their electrostatics and ability to form hydrogen bonds. Seven of the eight most disfavored interactions (>1.4 kcal/mol) correspond to the pairing of charged amino acids with either hydrophobic amino acids or proline (glu−met, ile−asp, pro−glu, pro−lys, met−lys, ile−glu), or the pairing of like charges (lys−lys), which are known to be destabilizing. Similarly, with the exception of one special case (pro−gly), all of the most favored interactions (<−1.4 kcal/mol) correspond to known stabilizing chemical interactions; four polar-basic interactions (thr−lys, gln−lys, and the special cases of lys−cys and tyr−his, due to the polar character of tyr and cys), two hydrophobic−hydrophobic interactions (met−leu, trp−met), four complementary charge interactions (lys−asp, his−asp, his−glu, arg−glu), two cation−π interactions (trp−lys, trp−arg), a polar−polar interaction (tyr−thr, due to the polar character of tyr), and a hydrophobic−aromatic interaction (tyr−ile). All of the interactions between hydrophobic and acidic amino acids have energies that are positive or close to zero, except for the tyr−asp interaction (−0.72 kcal/mol), which can be explained by the presence of the tyrosine hydroxyl group, and the tryptophan−glutamic acid interaction (−1.1 kcal/mol). Due to the electron-rich nature of the indole functional group, this contact energy cannot be rationalized as an anion−π interaction. However, it may arise via interactions with the δ^+ σ-framework, particularly the electron deficient region of the N_ε atom. The contact energies between hydrophobic and basic amino acids are, with a few notable exceptions, also positive as expected. The exceptions include interactions of histidine with isoleucine, methionine, tyrosine and tryptophan, which can be explained in part by the ability of histidine to become charge neutral around physiological pH, and in part by its ability to participate in aromatic/hydrophobic interactions. All but two of the remaining exceptions are cation−π interactions (lys and arg interacting with phe, trp, and tyr). The remaining two interactions are for lys with ile and val, consistent with the possibility for hydrophobic interactions via the lysine aliphatic chain. Further, 27 of the 36 contact energies between two hydrophobic amino acids are negative. The interactions of polar side-chain with themselves and with acidic, basic, hydrophobic, and special amino acids are more varied, also as expected.

Although the majority of the energies conform to expectations, there are a number of discrepancies. For instance, one would expect the interactions between two acidic amino acids to be highly disfavored due to electrostatic repulsion. However, the contact energies are either close to zero (glu−glu, asp−asp) or, for the glutamic acid−aspartic acid interaction, stabilizing (−0.6 kcal/mol). One possible explanation for this result is that the interactions are cation-mediated. Destruction of a Zn^{2+} mediated bridge is observed in three mutations of the idiotope−anti-idiotope complex between antibodies D1.3 and E5.2 (PDB 1DVF). Further, the crystal structures of two OMTKY3 mutants in complex with *Streptomyces griseus* protease B (PDB 1SGE and 1SGD), for which both the forward and backward mutations appear in the data set, shows that charge neutralization is achieved by a buried K^+ bridge. Although no additional evidence for cation bridges is seen for mutations in which acid−acid contacts are removed, such ions may not be resolved or reported. Further, as only the wild-type structures are available for the other complexes in which acid−acid contacts are created, it is difficult to assess the role of ion recruitment. Additionally, the leucine−leucine energy (0.5 kcal/mol) seems out of place with the commonly

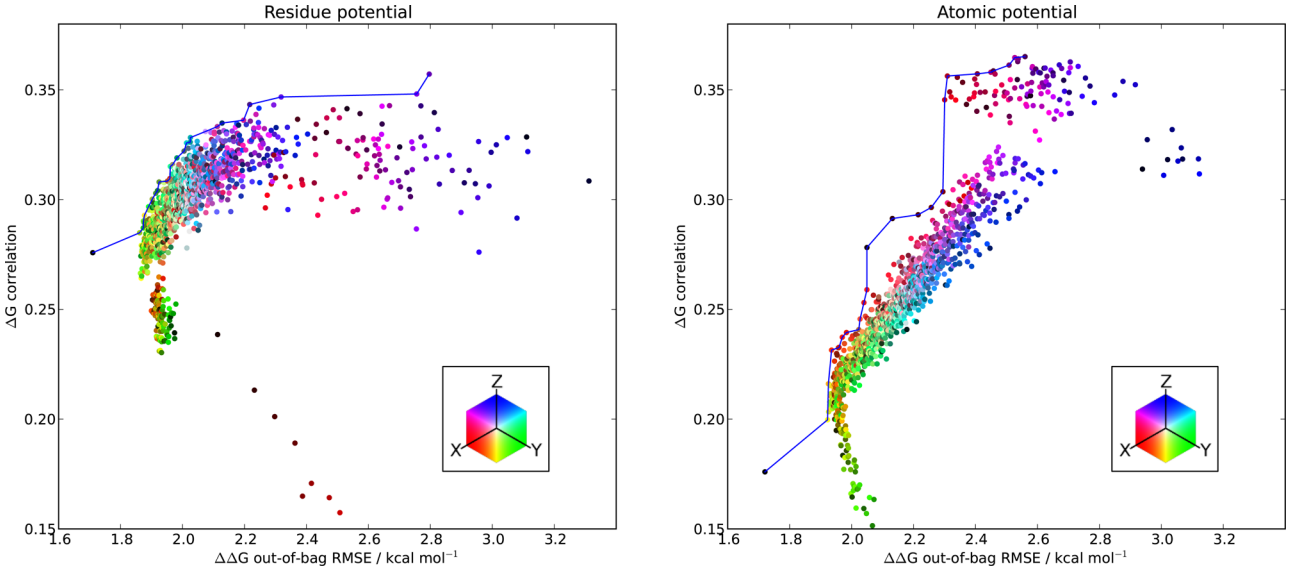


Figure 2. Pareto frontiers for the residue and atomic potentials. The $\Delta\Delta G$ out-of-bag RMSE is plotted against the ΔG Pearson correlation for each sampled combination of x, y, and z parameter. The (x, y, z) values for each model are shown by their respective (r, g, b) color, as indicated by the white corner facing color cube projection. The Pareto frontier is indicated by the blue line, and corresponds to the points which have an optimal trade off between $\Delta\Delta G$ and ΔG performance. The contact energies for these models are averaged to derive the weighted potential.

held notion that aliphatic residues are the most hydrophobic and compared to the other interaliphatic contact energies. The contact energy between two phenylalanine groups is also anomalous (0.7 kcal/mol), as one would expect a favorable $\pi-\pi$ interaction between the aromatic benzyl groups. The total number of changes in residue contacts, $\sum |N_{i,j}|$ for each i,j pair, are shown in Supporting Information Table 2. As a significant proportion of the training data comes from alanine scanning experiments, changes in alanine contacts are particularly frequent. On the other hand, changes in some i,j residue contacts are much less frequent. For instance, in the training data, only three lysine−lysine contacts are created or destroyed following mutation. This uneven distribution may offer an explanation for some of the counterintuitive values in Table 1. For instance, although the pairs for which there is a relative paucity of data ($\sum |N_{i,j}| < 15$) are in a minority, they include a number of anomalous results. These include the unfavorable interaction between two phenylalanines, as well as a number of the interactions that are unexpectedly favorable (asp−asp, glu−glu, his−his), including two of the three interactions between acidic pairs. Further, the surprisingly unfavorable leu−leu interactions has only 17 contacts in the training data. Nonetheless, the contact energies for most of the underrepresented pairs seem reasonable, and there are also anomalies for pairs that are well represented in the training data.

To see the extent to which the derived potential agrees with known statistical potentials, we calculated to correlation with 29 published potentials and propensities. For 24 of these, we use the same naming scheme used previously.[56,57] The remaining five are the two SIPPER amino-acid propensities,[58] SIPi and SIPs, the two intermolecular pairing preferences described by Glaser et al.,[59] GSVB and GSVBn, and the intermolecular contact potential described by Lu et al.,[19] SKOip. The correlations are given in Supporting Information Table 4. For 28 of the 29 statistical potentials, the correlation with the unweighted residue potential was of statistical significance (significance of correlation test, $n = 209$, $r > 0.114$, $p < 0.05$), with correlations up to 0.39. This indicates that while there are some commonalities with the statistical potentials as expected, the potential presented here has considerable differences with all of them. Previous studies have shown that although many of the statistical potentials correlate highly among themselves, others have weak or even negative correlations.[57] Indeed, on the basis of this, it is possible to group potentials by their agreement with one another. As the correlations with the other potentials are low, the presented potential would constitute a group of its own, indicating that the derived contact energies are considerably distinct from those obtained from the inverse Boltzmann equation. Despite this, however, there are a higher number of significant correlations with the residue potential and the other potentials than there are between some of the other potentials themselves, such as GKS, MSBM, RO, and GSVB. For instance, the correlation between GKS and MJ2H is −0.23, yet the correlations between the residue potential and GKS and MJ2H are 0.22 and 0.25 respectively, showing some degree of agreement even with different statistical potentials that are anticorrelated with one another.

3.1.2. Unweighted Atomic Potential. The unweighted atomic potential, when applied to the $\Delta\Delta G$ data, is shown in the top right corner of Figure 1. The overall correlation is similar to that observed for the residue potential (0.71 versus 0.73). The potential was used to predict the ΔG values, the free energy of binding, as shown in the bottom right corner of Figure 1. Again, we see a small but significant correlation ($r = 0.18$, $p = 0.015$). As with the residue potential, the correlation for the validated set is higher than for the nonvalidated set (0.32 versus 0.06), the correlation with the rigid complexes is higher than with the flexible complexes (0.28 versus 0.12) and the highest correlation is with the intersection between the rigid and validated sets ($r = 0.40$). The total number of changes in atom contacts, $\sum |N_{i,j}|$ for each i,j pair, are shown in Supporting Information Table 3. As there is a much greater number of atoms than residues, and fewer atom types than amino acids, the number of changes in atomic contacts is significantly higher. Subsequently, the risk of problems arising from the uneven distribution of data is lower. The values of the potential are

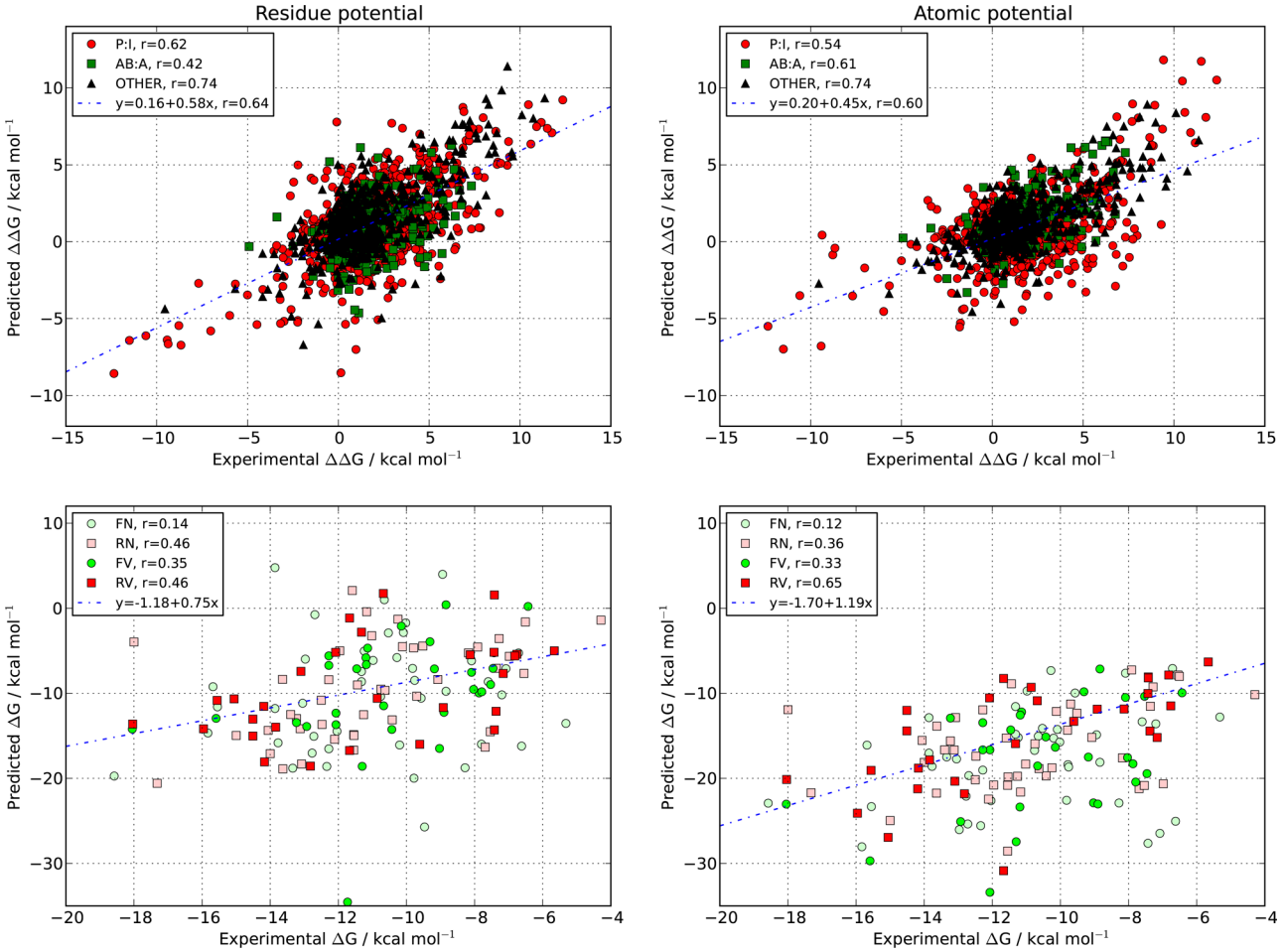


Figure 3. Modeled versus experimental $\Delta\Delta G$ and ΔG values for the weighted residue and weighted atomic potential. For the $\Delta\Delta G$ data, out-of-bag predictions are shown, and the data are subdivided into mutations in serine protease inhibitor complexes (P−I), antibody−antigen complexes (AB−A), and other complexes (OTHER), with Pearson correlations reported for each subset. The line of best fit is shown, with the slope, intercept, and overall correlation. The ΔG data are subdivided into rigid (R) and flexible (F) complexes, as well as validated (V) and nonvalidated (N) affinities, with correlations shown. The line of best fit, slope and intercept are given for the affinities in the validated rigid set. An outlier of very low predicted energy is omitted from atomic potential ΔG plot.

shown in Table 2. Although atoms are grouped into atom types based on the similarity of their environments and the types cannot be cleanly delineated by physiochemical properties, they can be roughly classified as positive, negative, hydrophobic, aromatic, and polar.[16] Going by this, the values in Table 2 also mostly conform to chemical intuition. Five of the seven most destabilizing atom contacts (>50 cal/mol) correspond to the repulsion of like charges (K+ RHK_mc, RHK_mc RHK_mc, DE_mc DE_mc, K+ K+) and charge-hydrophobic interactions (ILV_sc RHK_mc). Similarly, seven of the nine most stabilizing atom contacts (<−50 cal/mol) correspond to cation−π interactions (MFW_sc R+, K+ WY_sc), opposing charge interactions (DE- R+, DE- RHK_mc, K+ DE-) and charge-polar interactions (K+ CG, DE_mc CG). One exception is a favorable interaction between the side-chain and main-chain atoms of acidic residues (−69 cal/mol). Indeed, for aspartic and glutamic acid residues, it appears that the repulsions with themselves is falsely attributed to backbone interactions (120 cal/mol), rather than side-chain interactions (−3 cal/mol). A similar misattribution has occurred for the hydrophobic side-chain atom type ILV_sc, which has an unfavorable interaction (54 cal/mol) with the backbone atoms of basic residue, RHK_mc, and a negligible interaction with the arginine side-chain (−4 cal/mol). For interactions between basic residues, the lysine−lysine side-chain repulsions are correctly found (390 cal/mol), and indeed, all of the basic residue specific atom types repel one another, with the surprising exception of the arginine−lysine side-chain contact energy (−23 cal/mol). The interaction energies for oppositely charged residue specific atom types are also coherent, with particularly favorable side-chain interactions (−124 cal/mol), highlighting the importance of salt bridges. Further, the contact energies of hydrophobic side-chain specific atom types are favorable (−34 cal/mol), as are interactions between hydrophobic and aromatic atoms (as low as −30 cal/mol). Further, reasonable energies appear for cation−π and π−π contacts.

3.2. Weighting Coherently Augments the Potentials.

The weighted potentials were derived by performing uniform sampling of the x, y, and z parameters in eq 4 and calculating the potentials using the resultant weights with bootstrap aggregating. The contact energies were averaged over all the models that were found to be Pareto optimal with respect to the $\Delta\Delta G$ out-of-bag RMS error and the ΔG Pearson correlation, as shown in Figure 2. The correlations with the $\Delta\Delta G$ and ΔG data are shown in Figure 3. As the weighted potentials can be seen as optimized against both the $\Delta\Delta G$ data and the ΔG data, as opposed to just the $\Delta\Delta G$ data

for the unweighted potential, we see an improvement in correlation with the ΔG data compared to the unweighted potentials, as expected; 0.33 versus 0.25 for the residue potential and 0.30 versus 0.18 for the atomic potential. This increase in correlation is most visible when calculated over the complexes in the intersections of the set of validated affinities and the set of rigid complexes, with $r = 0.65$ for the atomic weighted potential. For comparison, the DComplex distance-dependent atomic potential has found wide use in the analysis of protein−protein interactions.[60] When applied to the same set complexes, it can reproduce the affinities with a correlation of 0.66.[23] Also as expected, we see that the increase in ΔG correlation is associated with a decrease in correlation for the $\Delta\Delta G$ data; 0.64 versus 0.73 for the residue potential and 0.60 versus 0.71 for the atomic. This decrease, however, is not evenly distributed over different types of complexes in the data set. The biggest decrease in performance is exhibited by the serine protease inhibitor complexes (0.62 versus 0.77 for the residue potential, 0.54 versus 0.72 for the atomic). These interactions form the most over-represented group in the training data, the group containing the most homologous interfaces, and the group within which there are many amino acids with multiple $\Delta\Delta G$ data. Less deleterious is the drop in $\Delta\Delta G$ performance for mutations in the antibody−antigen complexes (0.54 versus 0.42 for the residue potentials, and 0.61 versus 0.61 for the atomic), a group within which there are also many homologous interfaces, but with a less focused distribution of mutations. For the remaining complexes, however, we see a small increase in $\Delta\Delta G$ performance (0.74 versus 0.71 for both the residue and atomic potentials). This indicated that the weighting is behaving as expected, as the inclusion of absolute binding affinity data and the down-weighting of the over-represented groups is associated with an improvement in performance for modeling the energy changes associated with mutations in the more structurally and functionally diverse group of complexes.

For all cases there is one outlier in the ΔG plots, a flexible complex in the validated set (this outlier is omitted from the plots for the atomic potential, in the bottom left of Figures 1 and 3). When investigating the unbound structures of this complex for which the affinity is consistently overestimated, the p38 MAPK/MK2 interaction (2OZA), a large disorder-to-order transition in an interfacial loop is observed.

3.2.1. Weighted Residue Potential. Looking at the points on the Pareto front for the residue potential, on the left of Figure 2, we see a wide distribution of x,y,z parameters, but with a general gradient trend. For the models that perform best on the ΔG data, we see dark purple points, which correspond to large x and z parameters and a relatively small y parameter. This indicates that the greatest ΔG correlations arise when the downweighting is focused upon mutations in the same residue and mutations in homologous interfaces. As we move down the Pareto front and encounter models with increasing performance on the $\Delta\Delta G$ data, we see shades of amaranth, indicating an increasing y and decreasing z, then gray-cyan and gray-green points, indicating approximately equal x,y,z values, with slightly higher y and z parameters, and eventually chartreuse yellow, indicating a lower z value and higher x and y values. The model that performs best on the $\Delta\Delta G$ data, however, is the black point, which corresponds to the unweighted model ($x = y = z = 0$). Taken together, this shows that there is no particular parameter or parameter distribution associated with better performance on both the ΔG and $\Delta\Delta G$ data.

The contact energies for the weighted residue potential are shown in Table 3. As for the unweighted potential, the weighted residue potential was compared with the published statistical potential. Similarly, we see a significant, but low, correlation ($p < 0.05$) with 27 of the 29 other published potentials, as shown in Supporting Information Table 4. When looking at the contact energies, some of the anomalous entries associated with the unweighted potential have improved. For instance, the energy of the phenylalanine-phenylalanine contact is now negative (−0.6 kcal/mol), and the energey between charged histidine residues is now positive (0.5 kcal/mol). Further, the energies between hydrophobic residues are, in general, more stabilizing, with 30 of the 36 hydrophobic−hydrophobic interactions being negative, as opposed to 27 for the unweighted potential. Despite these improvements, and the fact that most of the values still conform to expectation, the discrepancy regarding negative energies between acidic residues persists. Further, a new discrepancy arises in the form of favorable arginine interactions with valine, and leucine, and the tryptophan−methionine energy is now exceptionally low at −5.3 kcal/mol. This value is much lower than that of previously estimated methonine−π interaction energies not in the context of a pre-existing binding site.[61] There are only 12 changes in the number of trp−met contacts in the data set (Supporting Information Table 2), and this extreme energy provides an interesting case study for how a deficit of training data can manifest itself. Both trp and met are common in binding hot-spots and are known to make large contributions to binding energy.[62] These hot-spots are often characterized by cooperative effects, such that mutation of the trp or met to alanine breaks the cooperativity and results in a large positive $\Delta\Delta G$. Without enough non-hotspot mutations involving changes in trp−met contacts, this large $\Delta\Delta G$ manifests itself as a very low trp−met contact energy.

3.2.2. Weighted Atomic Potential. The Pareto front for the weighted potential is shown on the right of Figure 2. Unlike the residue potential, a clear pattern can be seen. The Pareto front is dominated by red points, some of which are verging upon purple. This shows that the models with optimal performance have high x parameters, lower y parameters, and z parameters that are lower still. The interpretation of this is clear and intelligible. The x value represents the extent to which sets of mutations in the same residue are downweighted, y represents the extent to which sets of mutations in different residues of the same complex are downweighted, and z represents the extent to which sets of mutations in homologous interfaces are downweighted. As we find that $x > y > z$, this suggests that the mutual information in a pair of mutations in the same residue is greater than that in a different residue in the same interface, which is greater than that in a different homologous interface. This should come as no surprise, because pairs of mutations in a single residue can convey information to one another about that specific location in the binding interface and about the binding site in general, while only the latter can be conveyed between mutations in different residues of the same interface. Even less mutual information can be found in pairs of mutations in different homologous interfaces compared to mutations in the same interface. The atomic contact energies for the weighted atomic potential are shown in Table 4. Although we still get a misattribution of energies in the acidic residue specific atom types, there are a number of places in which the energies fit better with expectations compared to the unweighted potenial. For instance, now all the energies between basic amino acid specific atom types are positive. Further, all the

Table 3. Weighted Residue-Level Contact Potential[a]

	acidic		basic			hydrophobic								polar			special		
	ASP	GLU	HIS	LYS	ARG	ALA	VAL	LEU	ILE	PHE	TRP	MET	TYR	SER	THR	ASN	GLN	CYS	GLY	PRO
ASP	−0.037																			
GLU	−0.134	−0.405																		
HIS	−1.466	−2.001	0.488																	
LYS	−1.235	−0.730	1.885	1.978																
ARG	−1.286	−1.193	0.156	0.385	0.865															
ALA	0.308	−0.027	0.161	0.617	0.644	0.406														
VAL	−0.186	−0.888	0.044	−0.384	−0.243	0.048	−1.638													
LEU	−0.497	0.604	0.117	−0.364	−0.285	−0.493	−0.657	0.563												
ILE	1.126	1.707	−0.498	−0.898	0.502	−0.023	−0.812	−0.708	−1.920											
PHE	−0.151	−0.072	−0.410	−0.002	−0.576	−0.379	−0.958	−0.598	−0.705	−0.611										
TRP	−0.547	−1.239	0.084	−1.763	−1.499	−0.780	−0.934	−0.555	−0.765	−0.653	−0.616									
MET	1.022	1.285	−0.064	1.717	0.723	−0.358	−1.194	−1.535	−0.146	−0.268	−5.266	−1.827								
TYR	−0.927	−0.163	−2.124	−0.880	−0.426	0.248	0.562	−0.749	−1.792	−1.353	−0.158	−0.680	0.049							
SER	0.167	−0.521	−0.591	0.221	0.375	−0.121	0.234	−1.019	0.320	−0.464	−0.610	−0.822	−0.566	0.156						
THR	1.105	0.377	−1.155	−2.581	−0.684	−0.749	−1.044	−0.950	−0.972	0.939	−2.514	−0.484	−2.148	−0.740	0.294					
ASN	−0.032	−0.810	−0.065	0.437	0.183	0.390	−0.611	0.527	−0.178	0.350	−1.892	−0.045	−0.701	0.027	0.418	−1.420				
GLN	−0.936	−0.377	1.495	−0.970	−0.048	0.260	0.891	−0.978	−0.646	0.808	−0.644	−0.223	0.822	−1.382	−0.622	−0.854	1.951			
CYS	0.533	0.326	−1.384	−2.140	0.473	0.670	0.733	−0.259	−0.627	0.388	1.162	0.129	0.632	0.088	1.068	−0.355	−0.302	−0.443		
GLY	−0.128	0.293	0.113	−0.877	−1.013	−0.156	−0.822	−0.039	0.548	−1.077	0.252	0.192	−0.170	0.072	0.616	−1.136	1.031	−0.551	0.847	
PRO	−1.262	1.927	−1.333	2.530	0.395	2.593	0.586	−0.124	1.270	−0.582	0.975	−1.430	−0.041	0.537	0.060	0.343	1.113	0.400	−1.498	N/A

[a]Note that the proline−proline contact energy is omitted due to filtering of the training data. Values are shown in kcal/mol.

dx.doi.org/10.1021/ct400295z | J. Chem. Theory Comput. 2013, 9, 3715−3727

Table 4. Weighted Atomic Contact Potential[a]

	-ve		+ve			hydrophobic		aromatic		polar		
	DE_mc	DE-	R+	K+	RHK_mc	ILV_sc	AILMV_mc	MFW_sc	WY_sc	polar	mc	CG
DE_mc	0.066											
DE-	0.004	−0.103										
R+	−0.050	−0.040	0.031									
K+	0.072	−0.089	0.043	0.294								
RHK_mc	0.014	−0.085	0.116	0.017	0.018							
ILV_sc	−0.017	−0.034	−0.047	−0.099	0.017	−0.057						
AILMV_mc	0.000	−0.029	−0.026	0.024	−0.020	−0.030	−0.001					
MFW_sc	0.001	−0.024	0.001	0.074	−0.014	−0.014	−0.029	−0.031				
WY_sc	−0.067	−0.017	−0.012	−0.036	−0.033	−0.055	−0.047	−0.075	−0.047			
polar	0.053	−0.024	0.008	−0.059	−0.030	−0.011	−0.020	−0.008	−0.056	−0.020		
mc	0.007	0.021	0.005	0.003	−0.034	0.017	0.044	−0.033	0.041	−0.018	0.011	
CG	−0.079	0.048	−0.038	−0.028	−0.053	−0.045	−0.043	−0.009	0.054	0.002	−0.062	0.027

[a]Values are shown in kcal/mol.

Table 5. Performance of the Potentials for the Ranking of Docked Poses Generated Using SwarmDock, for All 176 Structures in the Docking Benchmark 4.0[a]

	unweighted						weighted					
	residue			atomic			residue			atomic		
	high	med.	acc.	high	med.	acc.	high	med.	acc.	high	med.	acc.
top 1%	0	**4**	7	0	2	6	0	3	5	**1**	**9**	**10**
	0.0	**2.4**	**4.0**	0.0	1.1	3.4	0.0	1.7	2.8	**0.6**	**5.1**	**5.7**
top 5%	0	11	**22**	0	9	18	0	**13**	20	**2**	**19**	**27**
	0.0	6.2	**12.5**	0.0	5.1	10.2	0.0	**7.4**	11.4	**1.1**	**10.8**	**15.3**
top 10 %	0	16	31	1	17	35	**1**	**22**	**33**	**3**	**26**	**39**
	0.0	9.1	17.6	0.6	9.7	19.9	**0.6**	**12.5**	**18.8**	**1.7**	**14.8**	**22.2**
top 50%	2	37	71	5	45	77	2	**40**	**73**	**6**	**47**	77
	1.1	21.0	40.3	2.8	25.6	43.8	1.1	**22.7**	**41.5**	**3.4**	**26.7**	43.8
top 100%	**7**	**55**	**98**	5	61	**102**	4	54	97	7	**63**	98
	4.0	**31.2**	**55.7**	2.8	34.7	**58.0**	2.3	30.7	55.1	**4.0**	**35.8**	55.7
all %	7	71	126	7	72	124	6	72	125	8	72	124
	4.0	40.3	71.6	4.0	40.9	70.5	3.4	40.9	71.0	4.5	40.9	70.5

[a]The potentials were used to rank and cluster the solutions. The number and percentage of complexes for which a high/medium/acceptable solution was found as the lowest energy member of the top 1/5/10/50/100/all clusters is shown. For comparison of the unweighted potentials with the weighted potentials, the values highlighted in bold on in the unweighted columns are higher than their corresponding values in the weighted columns, and *vice versa*.

hydrophobic−hydrophobic, hydrophobic−aromatic, and aromatic−aromatic atomic contact energies are negative.

3.3. Potentials Can Be Used to Score Docked Poses. As a further validation of the potentials, they were used to rank and cluster flexible unbound docking poses generated using SwarmDock.[50−52] These results appear in Table 5. The most common metric for comparing docking performance is the top 10 acceptable or better success rate, as this is the measure used in the CAPRI experiment, that is, the number of cases in which there is an acceptable or better quality docking solution within the top scoring 10 docking poses. Going by this, the top performing potential is the weighted atomic potential (22%), followed by the unweighted atomic potential (20%), the weighted residue potential (19%), and the unweighted residue potential (18%). Taking a more general view of Table 5, this trend seems to hold, with the atomic potentials performing better than the residue potentials and the weighted potentials performing better then their unweighted counterparts. This is particularly true when considering the medium or better quality solutions. As the docking reranking is a blind test of the potentials, this indicates that accounting for biases in the training data using the weighting scheme does improve the quality of the resultant potentials. When considering only the complexes in the "rigid-body" category, the number of complexes with an acceptable or better structure in the top 10 increases for all cases, up to between 22% for the unweighted residue potential and 27% for the weighted atomic potential. For comparison, the PatchDock-FireDock docking protocol[63,64] was recently evaluated using the same test set and evaluation criteria as reported here.[65] It was shown that the method could find an acceptable or better solution in the top 1/10/100 with a success rate of 10%, 24%, and 49% respectively. Although the top 1 and top 10 success rates reported here are lower, the top 100 success rates are higher. It should be noted that, unlike the FireDock scoring function, the potentials presented here have not been optimized for protein−protein docking.

4. CONCLUSION

We have presented here initial investigations into a novel method of deriving potentials for protein−protein interactions. Our approach was to model proteins with mutations at the binding interface, and fit the potentials to the binding free energies changes using the changes in intermolecular contacts

as basis functions. As the energies are fitted directly, this approach can bypass some of the approximations and assumptions inherent in methods derived from the statistics of structural databases. However, this method is susceptible to other limitations, including structurally disruptive mutations and biases in the training data. To address these, we implemented a protocol for selecting mutations that are unlikely to disrupt the monomers and devised a weighting scheme to counteract biases toward specific residues, complexes, and families of interaction. We qualitatively validate the derived potentials by comparing them to known physical chemistry. Although some peculiarities are observed, these can be explained in part by the biases and uneven distribution of contact changes in the training data. We also quantitatively evaluated the potentials by predicting $\Delta\Delta G$ values, ΔG values, and by scoring docked poses. For all of the presented potentials, we see significant correlations with the ΔG data and docking performance comparable to state of the art methods. Further, for the residue potentials, we see some degree of agreement with 27 to 28 of the 29 statistical contact potentials previously published in the literature, although significant differences exist with all of them. Taken together, these results provide strong evidence that the approach does capture many of the energetic factors working at the interface without having to make the assumptions required to extract potentials from structural databases and can be used for the construction of general-purpose energy estimators. Given that statistical potentials have received almost three decades of development and have benefited from massive increases in the availability of structural data, it is no surprise that modern atomic-resolution distance-dependent variants currently have greater predictive value for calculating $\Delta\Delta G$ than those presented here,[23] and this will likely also be true for the scoring of docked poses. Nevertheless, these preliminary investigations may serve as a starting point for several future avenues of exploration, toward improved weighting schemes and contact definitions, improving the methods by using more training data, investigating whether the performance can be boosted by enhanced mutant side-chain modeling, and possibly extending the method to include orientation and distance dependencies.

■ ASSOCIATED CONTENT

Supporting Information

Table 1: mutation data used to train the potential. Table 2: total number of changes in residue−residue contacts. Table 3: total number of changes in atom−atom contacts. Table 4: correlations of the residue potentials with 29 published potentials. Supplementary Methods: derivation of the equation used to calculate the potentials. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: juanf@bsc.es.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement PIEF-GA-2012-327899. This work was also funded by the Spanish Ministry of Science (BIO2010-22324).

■ REFERENCES

(1) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins* **2002**, *47*, 409−443.
(2) Kortemme, T.; Baker, D. *Curr. Opin. Chem. Biol.* **2004**, *8*, 91−97.
(3) Zhou, P.; Wang, C.; Ren, Y.; Yang, C.; Tian, F. *Curr. Med. Chem.* **2013**, *20*, 1985.
(4) Szymkowski, D. E. *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 590−600.
(5) Hwang, I.; Park, S. *Drug Discovery Today Technol.* **2008**, *5*, e43−e48.
(6) Grosdidier, S.; Fernandez-Recio, J. *Curr. Pharm. Des.* **2012**, *18*, 4607−4618.
(7) Morrow, J. K.; Zhang, S. *Curr. Pharm. Des.* **2012**, *18*, 1255−1265.
(8) Fernandez-Recio, J. *WIREs Comput. Mol. Sci.* **2011**, *1*, 680−698.
(9) Sippl, M. J. *Curr. Opin. Struct. Biol.* **1995**, *5*, 229−235.
(10) Godzik, A. *Structure* **1996**, *4*, 363−366.
(11) Bordner, A. J. *Methods Mol. Biol.* **2012**, *857*, 83−106.
(12) Sippl, M. J. *J. Comput. Aided Mol. Des.* **1993**, *7*, 473−501.
(13) Skolnick, J. *Curr. Opin. Struct. Biol.* **2006**, *16*, 166−171.
(14) Poole, A. M.; Ranganathan, R. *Curr. Opin. Struct. Biol.* **2006**, *16*, 508−513.
(15) Russ, W. P.; Ranganathan, R. *Curr. Opin. Struct. Biol.* **2002**, *12*, 447−452.
(16) Mintseris, J.; Pierce, B.; Wiehe, K.; Anderson, R.; Chen, R.; Weng, Z. *Proteins* **2007**, *69*, 511−520.
(17) Feliu, E.; Aloy, P.; Oliva, B. *Protein Sci.* **2011**, *20*, 529−541.
(18) Zhang, C.; Vasmatzis, G.; Cornette, J. L.; DeLisi, C. *J. Mol. Biol.* **1997**, *267*, 707−726.
(19) Lu, H.; Lu, L.; Skolnick, J. *Biophys. J.* **2003**, *84*, 1895−1901.
(20) Jiang, L.; Gao, Y.; Mao, F.; Liu, Z.; Lai, L. *Proteins* **2002**, *46*, 190−196.
(21) Liu, S.; Zhang, C.; Zhou, H.; Zhou, Y. *Proteins* **2004**, *56*, 93−101.
(22) Su, Y.; Zhou, A.; Xia, X.; Li, W.; Sun, Z. *Protein Sci.* **2009**, *18*, 2550−2558.
(23) Moal, I. H.; Agius, R.; Bates, P. A. *Bioinformatics* **2011**, *27*, 3002−3009.
(24) Jernigan, R. L.; Bahar, I. *Curr. Opin. Struct. Biol.* **1996**, *6*, 195−209.
(25) Rooman, M. J.; Wodak, S. J. *Protein Eng.* **1995**, *8*, 849−858.
(26) Gilis, D.; Rooman, M. *J. Mol. Biol.* **1997**, *272*, 276−290.
(27) Gilis, D.; Rooman, M. *J. Mol. Biol.* **1996**, *257*, 1112−1126.
(28) Zhou, H.; Zhou, Y. *Protein Sci.* **2002**, *11*, 2714−2726.
(29) Hoppe, C.; Schomburg, D. *Protein Sci.* **2005**, *14*, 2682−2692.
(30) Dehouck, Y.; Grosfils, A.; Folch, B.; Gilis, D.; Bogaerts, P.; Rooman, M. *Bioinformatics* **2009**, *25*, 2537−2543.
(31) Ota, M.; Isogai, Y.; Nishikawa, K. *Protein Eng.* **2001**, *14*, 557−564.
(32) Tuncbag, N.; Gursoy, A.; Keskin, O. *Bioinformatics* **2009**, *25*, 1513−1520.
(33) Xia, J. F.; Zhao, X. M.; Song, J.; Huang, D. S. *BMC Bioinformatics* **2010**, *11*, 174.
(34) Tuncbag, N.; Keskin, O.; Gursoy, A. *Nucleic Acids Res.* **2010**, *38*, W402−406.
(35) Moal, I. H.; Bates, P. A. *PLoS Comput. Biol.* **2012**, *8*, e1002351.
(36) Fleishman, S. J.; et al. *J. Mol. Biol.* **2011**, *414*, 289−302.
(37) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534−552.
(38) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859−883.
(39) Ben-Naim, A. *J. Chem. Phys.* **1997**, *107*, 3698−3706.
(40) Thomas, P. D.; Dill, K. A. *J. Mol. Biol.* **1996**, *257*, 457−469.
(41) Moal, I. H.; Fernandez-Recio, J. *Bioinformatics* **2012**, *28*, 2600−2607.
(42) Khatun, J.; Khare, S. D.; Dokholyan, N. V. *J. Mol. Biol.* **2004**, *336*, 1223−1238.
(43) Zhou, H.; Zhou, Y. *Proteins* **2002**, *49*, 483−492.
(44) Levy, E. D. *J. Mol. Biol.* **2010**, *403*, 660−670.

(45) Mintseris, J.; Weng, Z. *Genome Inf.* **2004**, *15*, 160−169.
(46) Guerois, R.; Nielsen, J. E.; Serrano, L. *J. Mol. Biol.* **2002**, *320*, 369−387.
(47) Breiman, L. *Mach. Learn.* **1996**, *24*, 123−140.
(48) Huang, X.; Brutlag, D. L. *Nucleic Acids Res.* **2007**, *35*, 678−686.
(49) Kastritis, P. L.; Moal, I. H.; Hwang, H.; Weng, Z.; Bates, P. A.; Bonvin, A. M.; Janin, J. *Protein Sci.* **2011**, *20*, 482−491.
(50) Moal, I. H.; Bates, P. A. *Int. J. Mol. Sci.* **2010**, *11*, 3623−3648.
(51) Li, X.; Moal, I. H.; Bates, P. A. *Proteins* **2010**, *78*, 3189−3196.
(52) Torchala, M.; Moal, I. H.; Chaleil, R. A.; Fernandez-Recio, J.; Bates, P. A. *Bioinformatics* **2013**, *29*, 807−809.
(53) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. *Proteins* **2010**, *78*, 3111−3114.
(54) Lensink, M. F.; Mendez, R.; Wodak, S. J. *Proteins* **2007**, *69*, 704−718.
(55) Mendez, R.; Leplae, R.; Lensink, M. F.; Wodak, S. J. *Proteins* **2005**, *60*, 150−169.
(56) Feng, Y.; Kloczkowski, A.; Jernigan, R. L. *BMC Bioinformatics* **2010**, *11*, 92.
(57) Pokarowski, P.; Kloczkowski, A.; Jernigan, R. L.; Kothari, N. S.; Pokarowska, M.; Kolinski, A. *Proteins* **2005**, *59*, 49−57.
(58) Pons, C.; Talavera, D.; de la Cruz, X.; Orozco, M.; Fernandez-Recio, J. *J. Chem. Inf. Model* **2011**, *51*, 370−377.
(59) Glaser, F.; Steinberg, D. M.; Vakser, I. A.; Ben-Tal, N. *Proteins* **2001**, *43*, 89−102.
(60) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. *J. Med. Chem.* **2005**, *48*, 2325−2335.
(61) Tatko, C. D.; Waters, M. L. *Protein Sci.* **2004**, *13*, 2515−2522.
(62) Ma, B.; Nussinov, R. *Curr. Top. Med. Chem.* **2007**, *7*, 999−1005.
(63) Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. *Nucleic Acids Res.* **2005**, *33*, W363−367.
(64) Andrusier, N.; Nussinov, R.; Wolfson, H. J. *Proteins* **2007**, *69*, 139−159.
(65) Schneidman-Duhovny, D.; et al. *Bioinformatics* **2012**, *28*, 3282−3289.