Biophysical Journal, Volume *115*

Supplemental Information

Contact Potential for Structure Prediction of Proteins and Protein Complexes from Potts Model

Ivan Anishchenko, Petras J. Kundrotas, and Ilya A. Vakser

SUPPOPRTING MATERIAL

Finding energy parameters by pseudo-likelihood maximization

Substituting Eq. 8 into Eq. 7 in the main text gives the pseudo-likelihood function L_p of parameters $\vec{\mathbf{h}}$ and $\mathbf{J}$:

$$L_p\left(\vec{\mathbf{h}},\mathbf{J}\right)=\prod_{i=1}^{L}\frac{\exp\left(-\beta U\left(x_i^{\text{nat}};N_i^{\text{nat}}\right)\right)}{\sum_{m=1}^{q}\exp\left(-\beta U\left(m;N_i^{\text{nat}}\right)\right)}. \tag{A1}$$

$U\left(k;N_i\right)$ is the energy of a single interaction center in state *k* surrounded by the set of neighbors N_i, which includes all atoms (residues) connected to site *i* by an edge in the graph (shown in blue in Fig. 1A , B of the main text), and "nat" indicates that all the neighbors are in their native states. Atom (residue) types x_i^{nat} along with the neighbors N_i^{nat} are from the native structures of the proteins in the training set and are fixed throughout the computations. For computational efficiency, we convert the pseudo-likelihood in Eq. A1 to the negative pseudo-log-likelihood function, which transforms the optimization problem (Eq. 7 in the main text) to

$$-\log\left(L_p\right)=\sum_{i=1}^{L}\left[\beta U\left(x_i^{\text{nat}};N_i^{\text{nat}}\right)+\log\sum_{k=1}^{q}\exp\left(-\beta U\left(k;N_i^{\text{nat}}\right)\right)\right]\xrightarrow[\vec{\mathbf{h}},\mathbf{J}]{}\min \tag{A2}$$

The gradient of the negative pseudo-log-likelihood function has components

$$\begin{cases}\dfrac{\partial\left(-\log L_p\right)}{\partial h_a}=\beta\sum_{i=1}^{L}\left[\delta_{a,x_i^{\text{nat}}}-P_i\left(a\right)\right],\\[2ex]\dfrac{\partial\left(-\log L_p\right)}{\partial J_{aa}}=\beta\sum_{i=1}^{L}\left(\sum_{j\in N_i^{\text{nat}}}\delta_{a,x_j^{\text{nat}}}\right)\left[\delta_{a,x_i^{\text{nat}}}-P_i\left(a\right)\right],\\[2ex]\dfrac{\partial\left(-\log L_p\right)}{\partial J_{ab}}=\beta\sum_{i=1}^{L}\left(\sum_{j\in N_i^{\text{nat}}}\delta_{a,x_j^{\text{nat}}}\right)\left[\delta_{a,x_i^{\text{nat}}}-P_i\left(a\right)\right]+\beta\sum_{i=1}^{L}\left(\sum_{j\in N_i^{\text{nat}}}\delta_{b,x_j^{\text{nat}}}\right)\left[\delta_{b,x_i^{\text{nat}}}-P_i\left(b\right)\right],\quad a<b\end{cases} \tag{A3}$$

where $a,b=1,...,q$, $\delta_{a,b}$ is the Kronecker delta and

$$P_i(a)=\frac{\exp\left(-\beta U\left(a;N_i^{\text{nat}}\right)\right)}{\sum_{k=1}^{q}\exp\left(-\beta U\left(k;N_i^{\text{nat}}\right)\right)} \tag{A4}$$

is the conditional probability of observing site i in state a, provided all neighboring sites N_i^{nat} are in their native states. We explicitly force the coupling matrix **J** to be symmetric by aggregating off-diagonal contributions from J_{ab} and J_{ba} into one derivative (3rd line in Eq. A3) and ommiting the lower triangular part of **J** (i.e. $a>b$) from computations. This reduces the total number of unknowns to $q+q\cdot(q+1)/2$. Given analytic derivatives in Eq. A3, the optimization problem Eq. A2 can be efficiently solved (e.g. by a Quasi-Newton method), until the requirement $\nabla\left(-\log L_p\right)\simeq 0$ (Eq. A3) is met.

Table S1. Docking accuracy according to CAPRI criteria

Quality category	Condition
High	f_{nat}[(1)] ≥ 0.5 *and* (L-RMSD[(2)] ≤ 1.0 Å *or* I-RMSD[(3)] ≤ 1.0 Å)
Medium	f_{nat} ≥ 0.3 *and* (1.0 < L-RMSD ≤ 5.0 Å *or* 1.0 < I-RMSD ≤ 2.0 Å)
Acceptable	f_{nat} ≥ 0.1 *and* (5.0 < L-RMSD ≤ 10.0 Å *or* 2.0 < I-RMSD ≤ 4.0 Å)
Incorrect	f_{nat} < 0.1 *and* (L-RMSD > 10.0 Å *and* I-RMSD > 4.0 Å)

(1) Fraction of predicted native residue–residue contacts

(2) C^{α} ligand RMSD when receptors are optimally aligned

(3) Interface C^{α} RMSD calculated over the set of native interface residues after a structural superposition of these residues

Table S2. *Details of various energy functions performance in the best model recognition from CASP decoys*. Best model's Z-score, its normalized rank 1 – *R*, and Pearson's correlation coefficient *r* of the energy score and GDT_TS score of models, all averaged over 224 CASP decoy sets, are shown in columns 6, 9 and 2 respectively. 95% confidence interval for the correlation coefficient averaged over 224 decoys is in column 3. 14 energy functions were ordered according to their *r* values, and one- and two-sided Wilcoxon signed-rank test was applied to compare samples of 224 correlation coefficients, Z-scores and normalized ranks between AACE18 and the other 13 energy functions. Corresponding *p*-values are in columns 4-5, 7-8 and 10-11. *P*-values < 0.05 are in blue.

potential	*r*	95% confidence interval	*p*-value for *r*		Z-score	*p*-value for *Z*-score		rank	*p*-value for rank	
			2-sided	1-sided		2-sided	1-sided		2-sided	1-sided
1	2	3	4	5	6	7	8	9	10	11
AACE18	0.606	(0.533;0.670)	-	-	1.09	-	-	0.809	-	-
GOAP	0.587	(0.511;0.654)	8.61E-02	4.31E-02	1.13	3.10E-01	8.45E-01	0.821	2.10E-01	8.95E-01
AACE167	0.585	(0.504;0.647)	7.99E-02	4.00E-02	1.08	9.82E-01	4.91E-01	0.808	9.26E-01	4.63E-01
DFIRE	0.562	(0.483;0.632)	3.62E-03	1.81E-03	0.90	1.28E-02	6.40E-03	0.796	7.46E-01	3.73E-01
dDFIRE	0.547	(0.468;0.617)	2.24E-03	1.12E-03	0.87	2.97E-03	1.49E-03	0.790	3.92E-01	1.96E-01
AACE20	0.540	(0.460;0.610)	5.01E-04	2.51E-04	0.88	1.43E-03	7.17E-04	0.755	2.44E-03	1.22E-03
RF-CB-SRS-OD	0.533	(0.451;0.606)	7.62E-06	3.81E-06	0.99	1.38E-01	6.88E-02	0.791	1.89E-01	9.44E-02
RRCE20	0.531	(0.450;0.603)	3.24E-05	1.62E-05	0.88	1.52E-03	7.58E-04	0.751	7.04E-04	3.52E-04
RW	0.524	(0.441;0.599)	1.05E-05	5.27E-06	0.83	1.62E-03	8.11E-04	0.769	1.64E-01	8.21E-02
RWplus	0.518	(0.434;0.594)	2.25E-06	1.12E-06	0.83	2.86E-03	1.43E-03	0.770	3.62E-01	1.81E-01
OPUS-PSP	0.515	(0.430;0.590)	3.68E-07	1.84E-07	1.04	6.79E-01	3.39E-01	0.796	6.99E-01	3.50E-01
DOPE	0.508	(0.422;0.584)	6.74E-09	3.37E-09	0.89	5.16E-02	2.58E-02	0.787	7.50E-01	3.75E-01
MJ3h	0.493	(0.407;0.571)	1.06E-10	5.31E-11	0.74	4.64E-07	2.32E-07	0.710	8.94E-08	4.47E-08
RF-HA-SRS	0.424	(0.331;0.510)	8.87E-21	4.44E-21	1.04	2.23E-02	1.12E-02	0.796	3.41E-02	1.71E-02

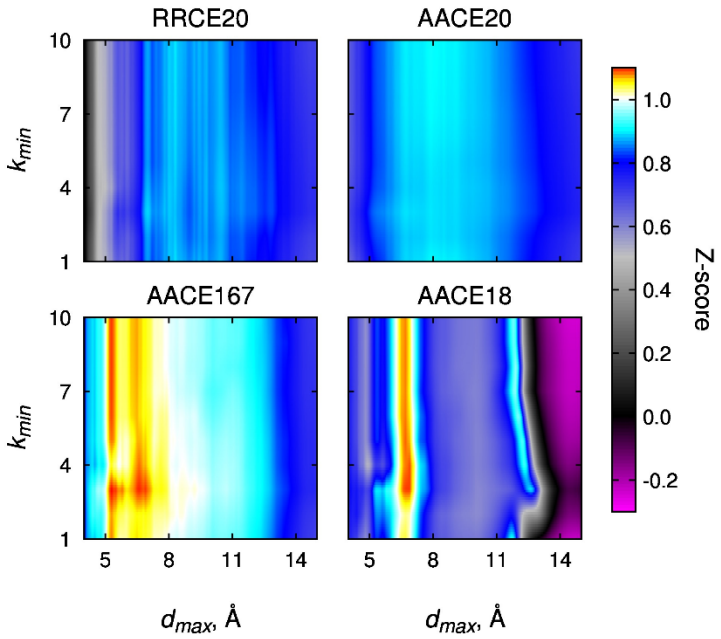

Figure S1: *Performance of the residue-residue and atom-atom contact potentials in best model recognition from CASP decoys*. The potentials derived at different values of sequence separation k_{min} and distance cut-off d_{max} were used to score models of 224 protein domains submitted to CASP rounds X and XI. The performance, measured as Z-score of the best model (the one with the highest GDT_TS score) averaged over all 224 evaluation units, is shown as heat map for RRCE20, AACE20, AACE167 and AACE18 potentials.

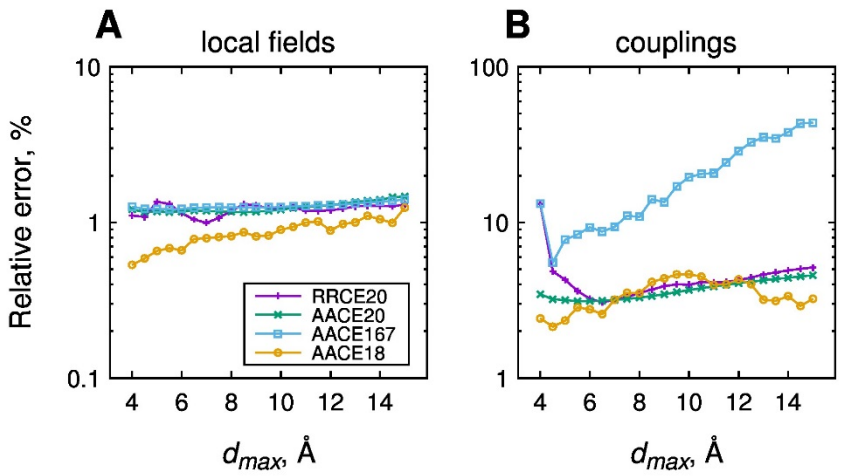

Figure S2: *Accuracy of the contact potentials parameters at different distance cut-offs.* The initial training set of 6,319 proteins was randomly split into halves. Each of the resulting subsets was used to train the statistical potentials at different distance cut-off d_{max} = 4 – 15Å with 0.5Å step, yielding in each case two sets of parameter estimates $\vec{\mathbf{h}}^{(1)}$, $\mathbf{J}^{(1)}$ and $\vec{\mathbf{h}}^{(2)}$, $\mathbf{J}^{(2)}$. Relative error was then calculated separately for (A) local fields $\vec{\mathbf{h}}$ and (B) couplings $\mathbf{J}$ using equation $\delta_{relative} = \left\|\vec{r}^{(1)} - \vec{r}^{(2)}\right\| \Big/ \left\|\vec{r}^{(1)} + \vec{r}^{(2)}\right\|$, where $\|\cdot\|$ is the l_2 vector norm. In the case of local fields, vector $\vec{r}$ is identical to vector $\vec{\mathbf{h}}$. For the coupling constants, $\vec{r}$ is composed of the upper triangle of matrix $\mathbf{J}$ plus the diagonal elements ($\mathbf{J}$ is symmetric, so the lower triangle was omitted). Relative errors $\delta_{relative}$ were calculated for five different random splits of the initial training set, and only the average values are shown on the plots.

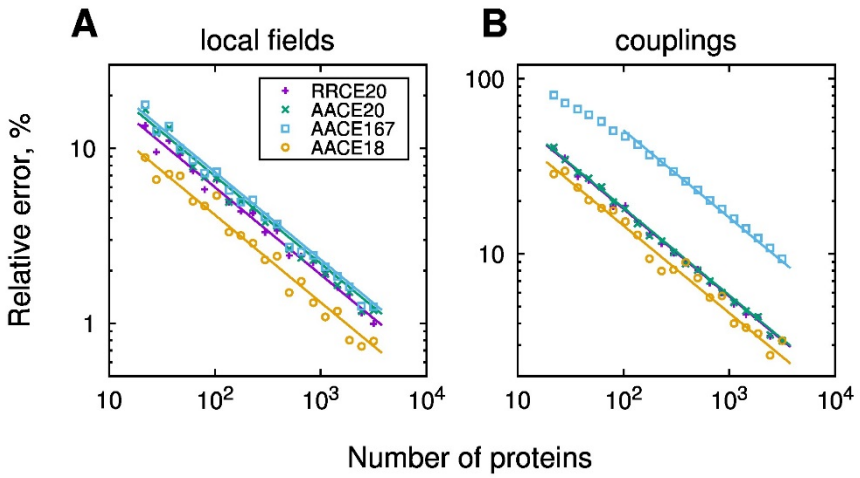

Figure S3: *Accuracy of the contact potentials parameters with varying sizes of the training set.* Using the procedure described in Figure S1, relative errors $\delta_{relative}$ for (A) local fields and (B) coupling constants were calculated for the randomly selected training subsets of different sizes ranging from 22 to 3159. The computed errors were fit by an empirically matched dependence $\delta_{relative} \sim 1/\sqrt{N}$, where N is the number of proteins used for training. Slight deviation of the AACE167 potential from this dependence (blue squares on the right-hand panel) is potentially caused by a very large number of parameters (~15,000), so that the system of equations (8) is underdetermined at small training set sizes N.

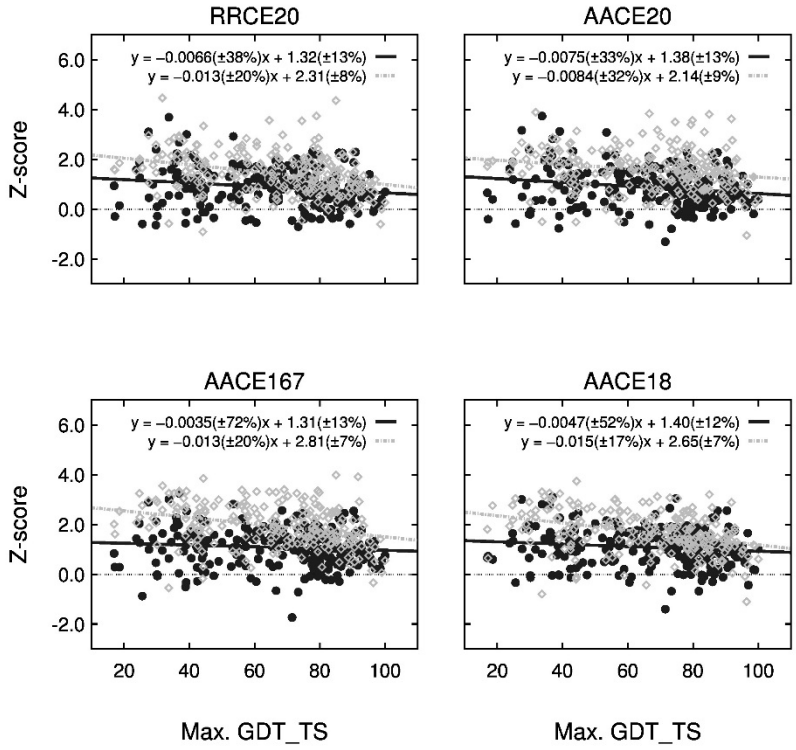

Figure S4: *Z-scores of the native structure (gray) and the highest accuracy model (black) in the CASP decoys depending on the decoys quality.* The GDT_TS score of the highest accuracy model (the best model according to CASP) was used as the measure of the decoys quality. For each of the 224 CASP decoy sets, the energy was calculated by the four contact potentials (see Methods in the main text),

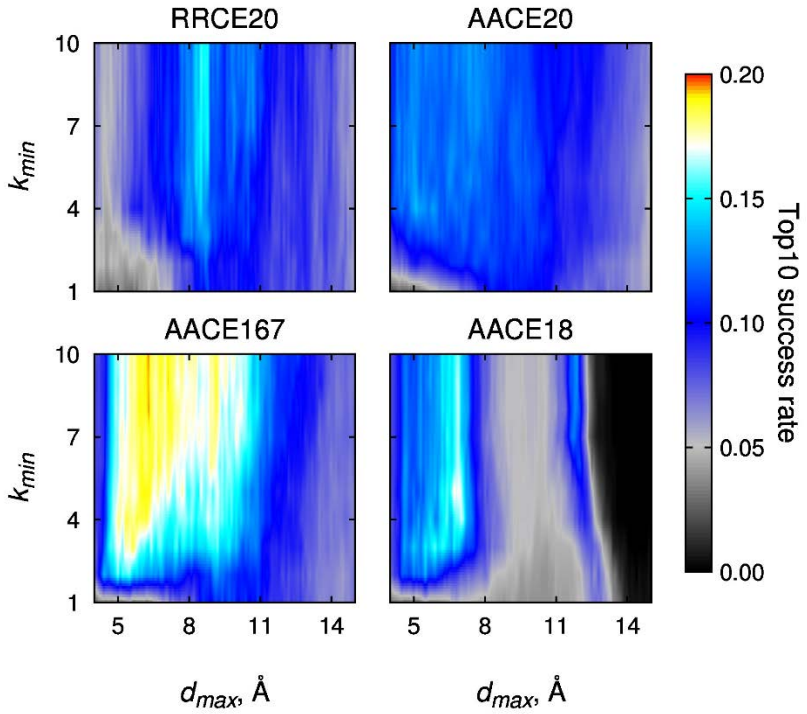

Figure S5: *Performance of the residue-residue and atom-atom contact potentials in near-native complex discrimination from low-resolution docking decoys*. Statistical potentials derived at different values of sequence separation k_{min} and distance cut-off d_{max} were used to score 100,000 unclustered matches for each of the 394 protein-protein complexes from DOCKGROUND Benchmark 4.0. Performance is measured in terms of the top-10 docking success rate (the fraction of complexes that have at least one near-native solution - acceptable or better quality according to CAPRI - among 10 best-scored models).

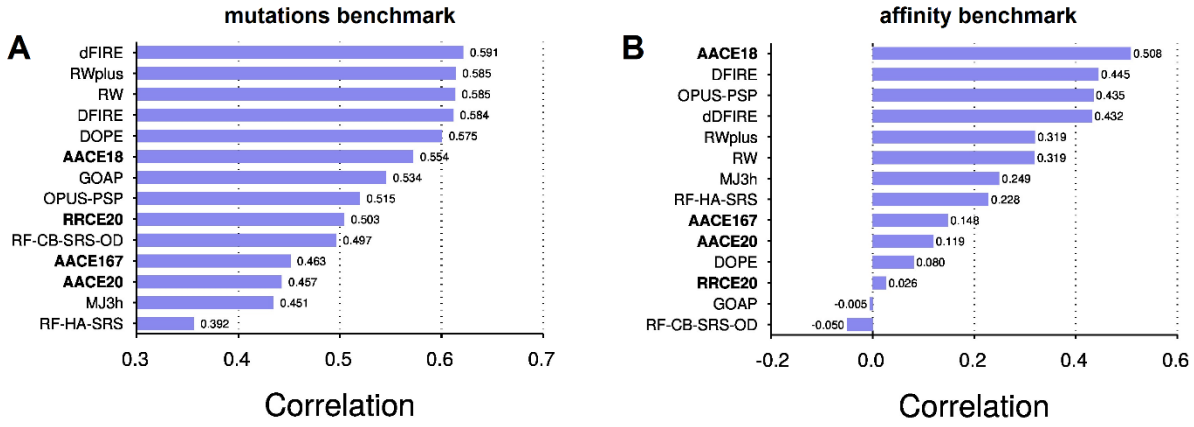

Figure S6: *Correlation of experimentally determined and calculated free energies.* (A) Pearson's correlation coefficient *r* between experimentally measured ($\Delta\Delta G_{\text{exp}}$) and calculated ($\Delta\Delta G_{calc}$) changes in folding free energies caused by point mutations over a set of 2,684 mutations for different knowledge-based energy functions. (B) The same scoring functions tested on their ability to recapitulate experimentally measured binding free energies (ΔG_{exp}) of 92 rigid-body complexes from Affinity Benchmark 2.0. The plot shows correlation coefficient *r* between ΔG_{exp} and ΔG_{calc} (see Methods). The RRCE20, AACE20, AACE167 and AACE18 potentials were derived at d_{max} = 8.0 Å and k_{min} = 3. Scoring functions on both panels are sorted by their performance according to *r*.

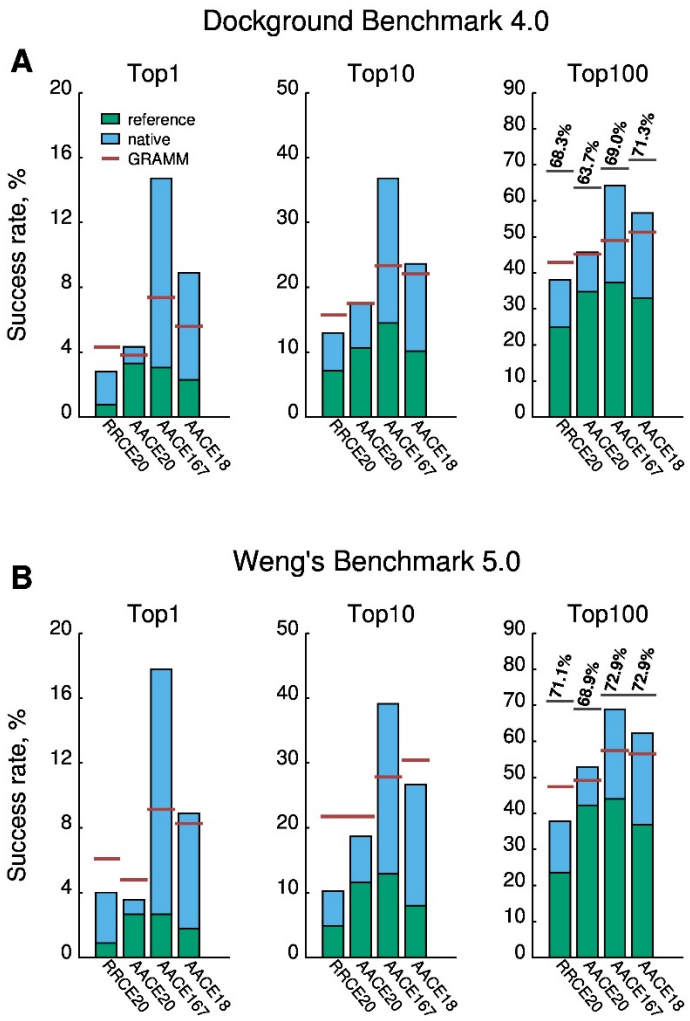

Figure S7: *Ranking of the native and reference structures in low-resolution docking decoys*. After scoring and clustering of top 100,000 matches from GRAMM (see Methods and caption to Fig. 6 in the main text for details), we checked whether the native (bound conformation, blue bars) and reference (unbound superimposed onto bound, green bars) is scored higher than any of the top 1,10 and 100 docking clusters. The fraction of such cases is plotted for (A) DOCKGROUND Benchmark 4 and (B) Weng's Benchmark 5. For comparison, docking success rates from Fig. 6 are shown by horizontal red lines. The top100 plots also show the maximal achievable docking success rates: black lines show the fraction of cases for which at least one docking cluster is of acceptable or better quality (see Table 1), regardless of its score.