Supporting Information

Table of Content

Supporting Texts

Text S1. The EvoEF energy function.

EvoEF is a physics-based energy function designed to describe the atomic interactions in proteins for design scoring and was first implemented in our protein design protocol EvoDesign (Pearce, et al., 2019). In general, it consists of five energy terms:

$$E_{EvoEF} = E_{VDW} + E_{ELEC} + E_{HB} + E_{DESOLV} - E_{REF} \tag{S1}$$

$$E_{VDW} = \sum_{i,j} w_{vdw} E_{vdw}(i,j) \tag{S2}$$

$$E_{ELEC} = \sum_{i,j} w_{elec} E_{elec}(i,j) \tag{S3}$$

$$E_{HB} = \sum_{i,j} w_{hb} E_{hb}(i,j) \tag{S4}$$

$$E_{DESOLV} = \sum_{i,j} w_{desolv} E_{desolv}(i,j) \tag{S5}$$

$$E_{REF} = \sum_{l=1}^{L} E_{ref}(aa_l) \tag{S6}$$

Here, E_{VDW}, E_{ELEC}, E_{HB}, E_{DESOLV}, and E_{REF} are the total van der Waals, electrostatic, hydrogen bonding, desolvation and reference energies for a protein. $E_{vdw}(i,j)$, $E_{elec}(i,j)$, $E_{hb}(i,j)$, and $E_{desolv}(i,j)$ are the pairwise interactions between non-bonded atoms i and j, and w_{vdw}, w_{elec}, w_{hb} and w_{desolv} are the relative weights for the corresponding energy terms. $E_{ref}(aa_l)$ is the amino acid-specific reference energy used to model the energy of an amino acid in the unfolded state, and the reference energy of a protein (E_{REF}) is assumed to be the summation of the amino acid-specific $E_{ref}(aa_l)$ values at position l across the whole protein chain with length L, where $E_{ref}(aa_l)$ is a parameter that was determined for each amino acid aa.

$E_{vdw}(i,j)$, is the van der Waals energy between atoms i and j, which is modified from the Lennard-Jones (LJ) 12-6 potential (Jones, 1924; Jones, 1924), which describes atomic packing interactions:

$$E_{vdw}(i,j) = \begin{cases} min\left\{5.0\varepsilon_{ij}, \varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right]\right\}, & if\ d_{ij} < 0.8909\sigma_{ij} \\ \varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{\sigma_{ij}}{d_{ij}}\right)^{6}\right], & if\ 0.8909\sigma_{ij} \le d_{ij} < 5.0 \\ A * d_{ij}^3 + B * d_{ij}^2 + C * d_{ij} + D, & if\ 5.0 \le d_{ij} < 6.0 \\ 0, & if\ d_{ij} \ge 6.0 \end{cases} \tag{S7}$$

$$\begin{cases} A = -0.4\varepsilon_{ij}\left(\frac{\sigma_{ij}}{5.0}\right)^{12} - 1.6\varepsilon_{ij}\left(\frac{\sigma_{ij}}{5.0}\right)^{6} \\ B = 7.8\varepsilon_{ij}\left(\frac{\sigma_{ij}}{5.0}\right)^{12} + 25.2\varepsilon_{ij}\left(\frac{\sigma_{ij}}{5.0}\right)^{6} \\ C = -50.4\varepsilon_{ij}\left(\frac{\sigma_{ij}}{5.0}\right)^{12} + 129.6\varepsilon_{ij}\left(\frac{\sigma_{ij}}{5.0}\right)^{6} \\ D = 108\varepsilon_{ij}\left(\frac{\sigma_{ij}}{5.0}\right)^{12} + 216\varepsilon_{ij}\left(\frac{\sigma_{ij}}{5.0}\right)^{6} \end{cases} \tag{S8}$$

where d_{ij} is the distance between the two atoms i and j, $\sigma_{ij} = \sigma_i + \sigma_j$ is the sum of their van der Waals atomic radii and ε_{ij} is the combined well-depth for atoms i and j, which is taken from the CHARMM19

force field (Brooks, et al., 1983). The attractive and repulsive components of the van der Waals energy are split at $d_{ij} = 0.8909\sigma_{ij}$. A maximum distance cutoff of 6.0 Å is set to increase the computational efficiency of EvoEF, and a cubic function is used to continuously connect the LJ energy from its value at 5.0 Å to zero at the cutoff distance (6.0 Å). For the repulsive component of the LJ potential, the maximum energy cutoff is set to $5.0\varepsilon_{ij}$, which helps alleviate possible clashes, while not overly penalizing them due to the discrete rotameric conformations used in protein design. The weights for the attractive and repulsive energies were separately determined.

$E_{elec}(i,j)$ is used to determine the electrostatic interactions between partially charged, non-bonded atoms i and j in a protein:

$$E_{elec}(i,j) = \begin{cases} \dfrac{C_0 q_i q_j}{\varepsilon(0.8\sigma_{ij})}\dfrac{1}{0.8\sigma_{ij}}, & if\ d_{ij} < 0.8\sigma_{ij} \\ \dfrac{C_0 q_i q_j}{\varepsilon(d_{ij})}\dfrac{1}{d_{ij}}, & if\ 0.8\sigma_{ij} < d_{ij} < 6.0 \\ 0, & if\ d_{ij} \geq 6.0 \end{cases} \quad (S9)$$

where q_i and q_j are the CHARMM19 atomic charges (Brooks, et al., 1983). Furthermore, $C_0 = 332$ Å kcal mol$^{-1}e^{-2}$, where e is the elementary charge, and $\varepsilon(d_{ij})$ is the distance-dependent dielectric constant, defined as $\varepsilon(d_{ij}) = 40d_{ij}$. The distance d_{ij} is set to $0.8\,\sigma_{ij}$ if d_{ij} is less than $0.8\,\sigma_{ij}$, when calculating the electrostatics term and dielectric constant. This avoids an infinite electrostatic energy value when d_{ij} is close to zero. Again, for the sake of computational efficiency, the electrostatics energy is set to zero if d_{ij} is beyond the maximum distance cutoff of 6.0 Å.

$E_{hb}(i,j)$ is used to calculate the hydrogen-bonding interactions between potential hydrogen bond donor/acceptor pairs for atoms i and j, one of which should be a polar hydrogen. $E_{hb}(i,j)$ is a linear combination of three energy terms that depend on the hydrogen-acceptor distance (d_{ij}^{HA}), the angle between the donor atom, hydrogen and acceptor (θ_{ij}^{DHA}), and the angle between the hydrogen, acceptor and base atom (φ_{ij}^{HAB}):

$$E_{hb}(i,j) = w_{d_{HA}} E(d_{ij}^{HA}) + w_{\theta_{DHA}} E(\theta_{ij}^{DHA}) + w_{\varphi_{HAB}} E(\varphi_{ij}^{HAB}) \quad (S10)$$

where:

$$\begin{cases} E(d_{ij}^{HA}) = \begin{cases} -cos\left[\dfrac{\pi}{2}(d_{ij}^{HA} - 1.9)/(1.9 - d_{min})\right], d_{min} \leq d_{HA} \leq 1.9 \\ -0.5\,cos\left[\pi\,(d_{ij}^{HA} - 1.9)/(d_{max} - 1.9)\right] - 0.5,\ \ 1.9\ \text{Å} < d_{HA} \leq d_{max} \\ 0 \qquad\qquad\qquad\qquad , \quad otherwise \end{cases} \\ E(\theta_{ij}^{DHA}) = -cos^4(\theta_{ij}^{DHA}),\ \theta_{ij}^{DHA} \geq 90° \\ E(\varphi_{ij}^{HAB}) = \begin{cases} -cos^4(\varphi_{ij}^{HAB} - 150°), \varphi_{ij}^{HAB} \geq 80° & for\ BBHB\ and\ for\ sp^2\ in\ SBHB\ or\ SSHB \\ -cos^4(\varphi_{ij}^{HAB} - 135°), \varphi_{ij}^{HAB} \geq 80° & for\ sp^3\ in\ SBHB\ or\ SSHB \end{cases} \end{cases} \quad (S11)$$

The optimal distance between the hydrogen and its acceptor is set to 1.9 Å, which is taken from Kortemme *et al.* (Kortemme, et al., 2003). Additionally, $d_{min} = 1.4$ Å and $d_{max} = 3.0$ Å are the lower and upper bounds on the distance between the hydrogen-acceptor pair. The optimal φ_{ij}^{HAB} value is set to either 150 °or 135 °, depending on the acceptor hybridization (sp^2 or sp^3) and the locations of the donor and acceptor atoms (BBHB: Backbone-Backbone Hydrogen Bond; SBHB: Sidechain-Backbone Hydrogen Bond; SSHB: Sidechain-Sidechain Hydrogen Bond).

$E_{desolv}(i,j)$ is used to describe the desolvation energy using the pairwise gaussian volume-excluded implicit solvation model developed by Lazaridis and Karplus (Lazaridis and Karplus, 1999). The pairwise $E_{desolv}(i,j)$ is calculated as the sum of the desolvation energies for atom i desolvating j ($f_{desov}(i,j)$) and for j desolvating i ($f_{desov}(j,i)$), as shown in Eq. S12:

$$E_{desolv}(i,j) = f_{desov}(i,j) + f_{desov}(j,i) \quad (S12)$$

$$f_{desolv}(i,j) = -V_j \frac{\Delta G_i^{free}}{2\pi^{\frac{3}{2}}\lambda_i d_{ij}^2} \exp\left[-\left(\frac{d_{ij}-\sigma_i}{\lambda_i}\right)^2\right] \quad (S13)$$

$$f_{desolv}(j,i) = -V_i \frac{\Delta G_j^{free}}{2\pi^{\frac{3}{2}}\lambda_j d_{ij}^2} \exp\left[-\left(\frac{d_{ij}-\sigma_j}{\lambda_j}\right)^2\right] \quad (S14)$$

where $V_{i,j}$, $\Delta G_{i,j}^{free}$, and $\lambda_{i,j}$ are the volumes, reference solvation energies, and correlation lengths for atoms i and j, respectively. All types of carbon and sulfur atoms are considered as nonpolar, while oxygen and nitrogen atoms are polar. The desolvation energy for hydrogen atoms is ignored in the Larzaridis-Karplus model, while the desolvation energy for other polar and nonpolar atoms are separately calculated and weighted. Specifically, $E_{DESOLV} = \sum w_{desolvPolar} f_{desolvPolar} + \sum w_{desolvNonP} f_{desolvNonP}$.

The parameter optimization and benchmark procedure of EvoEF was described in detail in the previous study (Pearce, et al., 2019). Briefly, EvoEF is not only benchmarked on 3,989 non-redundant protein monomer stability change data from 201 proteins taken from the FoldX and STRUM datasets (Guerois, et al., 2002; Quan, et al., 2016), but also on 2,204 non-redundant protein-protein binding affinity change data from 177 dimeric complexes taken from the SKEMPI v2.0 database (Jankauskaite, et al., 2018). Each dataset was randomly split in half into training and test sets. Out test results showed that EvoEF slightly outperformed FoldX in both tests. With respect to the protein stability change prediction, EvoEF obtained a PCC/RMSE of 0.472/1.751 kcal/mol, while FoldX obtained a PCC/RMSE of 0.465/2.010 kcal/mol; for protein-protein binding affinity change prediction, EvoEF obtained a PCC/RMSE of 0.514/2.109 kcal/mol, while FoldX obtained a PCC/RMSE of 0.490/2.248 kcal/mol.

Text S2. The EvoEF2 energy function.

In EvoEF2, the five original terms from EvoEF are preserved, but the weights and reference energies were determined differently (see Text S3). Moreover, four new terms were introduced to make EvoEF2 capable of tackling more difficult design cases and to fully utilize the structural information present in a given protein backbone. First, disulfide bonds exist in many proteins, but in EvoEF there is no term to model the possible formation of disulfide bonds. Since the length of a disulfide bond is around 2 Å, which is much less than the sum of the van der Waals radii of two sulfur atoms, possible disulfide-bond configurations in EvoEF may incur large clash penalties. Hence, we considered explicitly modeling disulfide bonds in EvoEF2. Second, the 20 canonical amino acids have different side-chain groups, and for the same amino acid, there may exist different rotamers. The different amino acids and rotamers exhibit distinct propensities and may occur at various frequencies at different protein backbone positions. To model this propensity, we introduced amino acid propensity, Ramachandran and Dunbrack rotamer probability terms into EvoEF2. The complete EvoEF2 energy function is written as:

$$E_{EvoEF2} = E_{VDW} + E_{ELEC} + E_{HB} + E_{DESOLV} + E_{SS} + E_{AAPP} + E_{RAMA} + E_{ROT} - E_{REF} \quad \text{(S15)}$$

Here, E_{SS} describes the disulfide-bonding interactions which are modeled as follows:

$$\begin{aligned} E_{SS} = \sum_{i,j} w_{SS} \Bigg[& 0.8 \times \left(1 - e^{-10\left(d_{ij}^{S_{\gamma 1}S_{\gamma 2}} - 2.03\right)}\right)^2 + 0.005 \times \left(\theta_{ij}^{C_{\beta 1}S_{\gamma 1}S_{\gamma 2}} - 105°\right)^2 + 0.005 \\ & \times \left(\theta_{ij}^{C_{\beta 2}S_{\gamma 2}S_{\gamma 1}} - 105°\right)^2 + \left(cos\left(2\chi_{ij}^{C_{\beta 1}S_{\gamma 1}S_{\gamma 2}C_{\beta 2}}\right) + 1\right) + 1.25 \\ & \times sin\left(\chi_{ij}^{C_{\alpha 1}C_{\beta 1}S_{\gamma 1}S_{\gamma 2}} + 120°\right) - 1.75 + 1.25 \times sin\left(\chi_{ij}^{C_{\alpha 2}C_{\beta 2}S_{\gamma 2}S_{\gamma 1}} + 120°\right) \\ & - 1.75 \Bigg] \end{aligned} \quad (S16)$$

where w_{ss} is the weight factor, atom pair i and j represent the two S_γ atoms from two different cysteines, $d_{ij}^{S_{\gamma 1}S_{\gamma 2}}$ is the distance between them, $\theta_{ij}^{C_{\beta 1}S_{\gamma 1}S_{\gamma 2}}$ is the angle between atoms $C_{\beta 1}$, $S_{\gamma 1}$, and $S_{\gamma 2}$, $\theta_{ij}^{C_{\beta 2}S_{\gamma 2}S_{\gamma 1}}$ is the angle between atoms $C_{\beta 2}$, $S_{\gamma 2}$ and $S_{\gamma 1}$, $\chi_{ij}^{C_{\beta 1}S_{\gamma 1}S_{\gamma 2}C_{\beta 2}}$ is the torsion angle between atoms $C_{\beta 1}$, $S_{\gamma 1}$, $S_{\gamma 2}$ and $C_{\beta 2}$, $\chi_{ij}^{C_{\alpha 1}C_{\beta 1}S_{\gamma 1}S_{\gamma 2}}$ is the torsional angle between atoms $C_{\alpha 1}$, $C_{\beta 1}$, $S_{\gamma 1}$ and $S_{\gamma 2}$, and $\chi_{ij}^{C_{\alpha 2}C_{\beta 2}S_{\gamma 2}S_{\gamma 1}}$ is the torsional angle between atoms $C_{\alpha 2}$, $C_{\beta 2}$, $S_{\gamma 2}$ and $S_{\gamma 1}$. The distance $d_{ij}^{S_{\gamma 1}S_{\gamma 2}}$ must be within [1.95, 2.15] in order to calculate $E_{ss}(i,j)$, and $E_{ss}(i,j)$=0 if $E_{ss}(i,j)$>0 or $d_{ij}^{S_{\gamma 1}S_{\gamma 2}} \notin [1.95, 2.15]$.

E_{AAPP} represents the energy for calculating amino acid propensities at given backbone (ϕ/ψ) angles by

$$E_{AAPP} = \sum_{1 \le l \le L} -w_{aapp} \ln \frac{P(aa_l|\phi_l, \psi_l)}{P(aa_l)} \quad \text{(S17)}$$

where w_{aapp} is the weight parameter, l is the design position, aa_l and (ϕ_l, ψ_l) are the amino acid type and backbone torsional angles, respectively, at position l. $P(aa_l|\phi_l, \psi_l)$ and $P(aa_l)$ are the probabilities of observing amino acid aa_l at a given (ϕ_l, ψ_l) and any backbone torsional angle, respectively. The statistics were obtained from the Top8000 dataset (Chen, et al., 2010) using a grid step of 10° for ϕ and ψ. Following a similar strategy to (Krivov, et al., 2009), for the N-terminal residue and other residues whose ϕ angle cannot be determined by backbone (due to missing backbone atoms), ϕ is set to -60 °, and similarly, for the C-terminal residue and other residues whose ψ angle cannot be determined by backbone (due to missing backbone atoms), ψ is set to 60 °. This is also applicable to the calculation of E_{RAMA} and E_{ROT}.

E_{RAMA} is the Ramachandran term for choosing specific backbone angles (ϕ/ψ) given a particular amino acid:

$$E_{RAMA} = \sum_{1 \le l \le L} -w_{rama} \ln P(\phi_l, \psi_l | aa_l) \tag{S18}$$

where w_{rama} is the weight parameter. This term is used to calculate how suitable the backbone (ϕ_l, ψ_l) is given aa_l. The statistics were obtained using the same dataset and grid step mentioned above (Chen, et al., 2010).

Finally, E_{ROT} is the energy term for modeling the rotamer probabilities from the rotamer library:

$$E_{ROT} = \sum_{1 \le l \le L} -w_{rot} \ln P(rot_l) \tag{S19}$$

where w_{rot} is the corresponding weight. $P(rot_l)$ is the probability of seeing a rotamer rot_l at position l for an amino acid type, which is directly taken from Dunbrack's 2010 backbone-dependent rotamer library (Shapovalov and Dunbrack, 2011) without further modification.

Text S3. Optimization of weights and reference energies in EvoEF2.

EvoEF was optimized and benchmarked by using two large sets of thermodynamic change data upon mutation, where the parameterization procedure of EvoEF was previously described in detail (Pearce, et al., 2019). We optimized EvoEF2 using a procedure similar to the one-at-a-time approach used by Rosetta (Kuhlman and Baker, 2000; Leaver-Fay, et al., 2013) and Medusa (Ding and Dokholyan, 2006). Specifically, the weights for each energy term and the 20 reference energies were determined by maximizing the product of $e^{-E_{min}(AA_{nat}|w)}/\sum_i e^{-E_{min}(AA_i|w)}$ across all of the residues positions for a training set of 222 monomers using a gradient descent optimization procedure, where $E_{min}(AA_{nat}|w)$ was the energy of the best rotamer for the native amino acid AA_{nat} given the weight set (w), $E_{min}(AA_i|w)$ was the energy of the best rotamer for amino acid AA_i using the same weight set, w, and the partition function was over all 20 amino acids at each position. The rotamers were taken from Dunbrack's backbone-dependent 2010 rotamer library (Shapovalov and Dunbrack, 2011). The energies for the rotamers at each position were calculated in the context of fixed surrounding residues. The determined weights and reference energies were then refined based on the results of complete sequence design for the same training proteins to reduce the deviation of the 20 amino acid distributions between the designed sequences and the native sequences. Similarly, optimization of the weights for inter-chain interactions was first determined by maximizing the product of $e^{-E_{min}(AA_{nat}|w)}/\sum_i e^{-E_{min}(AA_i|w)}$ over the interface residues positions for a training set of 132 dimers, where the weights for the monomeric energy terms and reference energies were fixed. The weights were also refined by complete PPI sequence design simulations for the 132 dimers.

Text S4. Protein design procedure.

We extended the EvoDesign Monte Carlo (MC) pipeline (Pearce, et al., 2019) to test the ability of EvoEF and EvoEF2 to perform protein design. Starting from a random sequence, a simulated annealing Monte Carlo (SAMC) (Kirkpatrick, et al., 1983) procedure was used to explore the sequence space for fixed protein backbones, where the MC movement consisted of exchanging one rotamer for another at a randomly chosen position. All amino acid types were considered at each design position and their side-chain conformations were taken from Dunbrack's 2010 backbone-dependent rotamer library (Shapovalov and Dunbrack, 2011). Rarely seen rotamers with probabilities < 3% were excluded (Kuhlman and Baker, 2000), yielding about a total of 165 rotamers for each amino acid at a given site. An MC move was accepted or rejected according to the Metropolis rule, where the acceptance probability for an unfavorable energy increase, ΔE, was $e^{-\Delta E/T}$. The temperature T varied from $T_{high} = 5$ to $T_{low} = 0.01$ with a decrease factor of 0.8, and three SAMC cycles were performed for convergence. We did not use a protein length-dependent temperature because ΔE did not exhibit strong length dependency. The number of MC steps at each temperature was determined as the sum of the number of rotamers at all designed positions, approximately equal to $165 \times L$ for a protein with L amino acids. This SAMC procedure was very fast, and for instance, it took less than 15 minutes to completely design a protein that was about 200 amino acids long. Because MC-based methods do not guarantee the global optimum solution, for a given structure, we performed ten independent simulation trajectories starting from different random sequences and selected the lowest-energy sequence as the final design and used it for comparison with the native sequence. In fact, the SAMC simulation procedure converged well and in almost all cases the sequences obtained from different runs shared > 85% sequence identity and had similar energies, and the designed sequences in the core region were nearly identical.

Supporting Tables

Table S1. Native sequence recapitulation results from designing 148 test set monomers using EvoEF and EvoEF2. '#nat' is the number of native residues, '#des' is the number of designed residues and '#id' is the number of residues with recapitulated identities (the same designations are used in Table S2 and Table S7-9).

Residue	EvoEF					EvoEF2				
	#nat	#des	#id	#id/#nat	#id/#des	#nat	#des	#id	#id/#nat	#id/#des
All residues										
Ala	1945	832	337	0.173	0.405	1945	2010	722	0.371	0.359
Cys	373	906	37	0.099	0.041	373	384	71	0.190	0.185
Asp	1333	1999	180	0.135	0.090	1333	1298	368	0.276	0.284
Glu	1670	1660	141	0.084	0.085	1670	1606	281	0.168	0.175
Phe	951	371	106	0.111	0.286	951	939	309	0.325	0.329
Gly	1806	719	613	0.339	0.853	1806	1773	1338	0.741	0.755
His	531	263	32	0.060	0.122	531	548	68	0.128	0.124
Ile	1313	4890	542	0.413	0.111	1313	1310	555	0.423	0.424
Lys	1353	167	16	0.012	0.096	1353	1310	198	0.146	0.151
Leu	2219	5361	837	0.377	0.156	2219	2275	1177	0.530	0.517
Met	429	351	8	0.019	0.023	429	436	59	0.138	0.135
Asn	969	573	50	0.052	0.087	969	1071	160	0.165	0.149
Pro	1072	534	195	0.182	0.365	1072	1132	558	0.521	0.493
Gln	871	148	12	0.014	0.081	871	827	81	0.093	0.098
Arg	1257	463	51	0.041	0.110	1257	1240	170	0.135	0.137
Ser	1415	581	136	0.096	0.234	1415	1334	262	0.185	0.196
Thr	1304	723	87	0.067	0.120	1304	1252	279	0.214	0.223
Val	1735	2235	450	0.259	0.201	1735	1770	812	0.468	0.459
Trp	350	154	14	0.040	0.091	350	364	49	0.140	0.135
Tyr	838	804	141	0.168	0.175	838	855	201	0.240	0.235
Total	**23734**	**23734**	**3985**	**0.168**	**0.168**	**23734**	**23734**	**7718**	**0.325**	**0.325**
Core residues										
Ala	775	526	253	0.326	0.481	775	901	468	0.604	0.519
Cys	198	343	21	0.106	0.061	198	239	52	0.263	0.218
Asp	119	885	47	0.395	0.053	119	150	39	0.328	0.260
Glu	120	600	23	0.192	0.038	120	193	31	0.258	0.161
Phe	438	137	57	0.130	0.416	438	369	177	0.404	0.480
Gly	434	313	255	0.588	0.815	434	346	315	0.726	0.910
His	117	115	11	0.094	0.096	117	100	15	0.128	0.150
Ile	620	776	256	0.413	0.330	620	696	338	0.545	0.486

Lys	88	11	0	0.000	0.000	88	129	22	0.250	0.171
Leu	996	838	371	0.372	0.443	996	1147	682	0.685	0.595
Met	167	45	2	0.012	0.044	167	220	39	0.234	0.177
Asn	133	175	19	0.143	0.109	133	82	25	0.188	0.305
Pro	198	222	94	0.475	0.423	198	176	121	0.611	0.688
Gln	102	27	3	0.029	0.111	102	63	13	0.127	0.206
Arg	138	42	9	0.065	0.214	138	108	19	0.138	0.176
Ser	278	292	66	0.237	0.226	278	112	39	0.140	0.348
Thr	270	240	40	0.148	0.167	270	252	92	0.341	0.365
Val	825	646	248	0.301	0.384	825	933	514	0.623	0.551
Trp	139	44	8	0.058	0.182	139	85	23	0.165	0.271
Tyr	342	220	65	0.190	0.295	342	196	90	0.263	0.459
Total	**6497**	**6497**	**1848**	**0.284**	**0.284**	**6497**	**6497**	**3114**	**0.479**	**0.479**
Surface residues										
Ala	626	72	11	0.018	0.153	626	542	76	0.121	0.140
Cys	40	262	1	0.025	0.004	40	33	1	0.025	0.030
Asp	827	397	50	0.060	0.126	827	762	211	0.255	0.277
Glu	1079	376	59	0.055	0.157	1079	774	144	0.133	0.186
Phe	157	73	4	0.025	0.055	157	235	24	0.153	0.102
Gly	808	143	126	0.156	0.881	808	990	628	0.777	0.634
His	210	43	6	0.029	0.140	210	244	23	0.110	0.094
Ile	243	2711	101	0.416	0.037	243	104	26	0.107	0.250
Lys	804	97	10	0.012	0.103	804	647	85	0.106	0.131
Leu	411	2886	159	0.387	0.055	411	260	75	0.182	0.288
Met	124	175	4	0.032	0.023	124	56	2	0.016	0.036
Asn	561	181	12	0.021	0.066	561	767	94	0.168	0.123
Pro	606	130	33	0.054	0.254	606	687	284	0.469	0.413
Gln	473	69	4	0.008	0.058	473	492	41	0.087	0.083
Arg	630	285	22	0.035	0.077	630	629	67	0.106	0.107
Ser	714	101	24	0.034	0.238	714	953	160	0.224	0.168
Thr	549	260	14	0.026	0.054	549	571	80	0.146	0.140
Val	336	899	44	0.131	0.049	336	226	53	0.158	0.235
Trp	63	36	0	0.000	0.000	63	147	5	0.079	0.034
Tyr	163	228	12	0.074	0.053	163	305	23	0.141	0.075
Total	**9424**	**9424**	**696**	**0.074**	**0.074**	**9424**	**9424**	**2102**	**0.223**	**0.223**

Table S2. Native sequence recapitulation results from designing 222 training set monomers using EvoEF and EvoEF2.

Residue	EvoEF					EvoEF2				
	#nat	#des	#id	#id/#nat	#id/#des	#nat	#des	#id	#id/#nat	#id/#des
All residues										
Ala	2683	1192	434	0.162	0.364	2683	2832	994	0.370	0.351
Cys	555	1207	57	0.103	0.047	555	497	77	0.139	0.155
Asp	1902	2837	272	0.143	0.096	1902	1959	537	0.282	0.274
Glu	2286	2379	196	0.086	0.082	2286	2437	388	0.170	0.159
Phe	1309	489	148	0.113	0.303	1309	1378	467	0.357	0.339
Gly	2511	960	830	0.331	0.865	2511	2511	1856	0.739	0.739
His	774	336	30	0.039	0.089	774	819	93	0.120	0.114
Ile	1888	7206	780	0.413	0.108	1888	1881	827	0.438	0.440
Lys	2009	170	19	0.009	0.112	2009	1962	302	0.150	0.154
Leu	3198	7704	1229	0.384	0.160	3198	3126	1691	0.529	0.541
Met	616	456	11	0.018	0.024	616	606	79	0.128	0.130
Asn	1426	819	81	0.057	0.099	1426	1387	213	0.149	0.154
Pro	1520	819	284	0.187	0.347	1520	1566	769	0.506	0.491
Gln	1389	246	23	0.017	0.093	1389	1281	143	0.103	0.112
Arg	1770	644	70	0.040	0.109	1770	1823	241	0.136	0.132
Ser	2062	754	134	0.065	0.178	2062	1949	401	0.194	0.206
Thr	1893	1071	118	0.062	0.110	1893	1789	395	0.209	0.221
Val	2433	3181	627	0.258	0.197	2433	2416	1095	0.450	0.453
Trp	431	231	32	0.074	0.139	431	482	64	0.148	0.133
Tyr	1196	1150	230	0.192	0.200	1196	1150	270	0.226	0.235
Total	**33851**	**33851**	**5605**	**0.166**	**0.166**	**33851**	**33851**	**10902**	**0.322**	**0.322**
Core residues										
Ala	1056	714	321	0.304	0.450	1056	1208	614	0.581	0.508
Cys	283	444	28	0.099	0.063	283	345	61	0.216	0.177
Asp	209	1181	72	0.344	0.061	209	212	71	0.340	0.335
Glu	142	867	27	0.190	0.031	142	283	35	0.246	0.124
Phe	578	195	81	0.140	0.415	578	472	241	0.417	0.511
Gly	588	404	340	0.578	0.842	588	477	406	0.690	0.851
His	184	150	16	0.087	0.107	184	164	30	0.163	0.183
Ile	893	1129	349	0.391	0.309	893	1023	529	0.592	0.517
Lys	124	15	2	0.016	0.133	124	180	25	0.202	0.139
Leu	1446	1135	550	0.380	0.485	1446	1559	985	0.681	0.632
Met	234	83	5	0.021	0.060	234	320	49	0.209	0.153
Asn	170	229	19	0.112	0.083	170	83	24	0.141	0.289

Pro	230	276	108	0.470	0.391	230	220	142	0.617	0.645
Gln	161	59	5	0.031	0.085	161	122	22	0.137	0.180
Arg	237	50	17	0.072	0.340	237	135	34	0.143	0.252
Ser	361	367	51	0.141	0.139	361	146	45	0.125	0.308
Thr	367	352	44	0.120	0.125	367	369	114	0.311	0.309
Val	1114	962	353	0.317	0.367	1114	1295	693	0.622	0.535
Trp	168	79	19	0.113	0.241	168	108	25	0.149	0.231
Tyr	430	284	92	0.214	0.324	430	254	107	0.249	0.421
Total	**8975**	**8975**	**2499**	**0.278**	**0.278**	**8975**	**8975**	**4252**	**0.474**	**0.474**
Surface residues										
Ala	888	118	12	0.014	0.102	888	809	122	0.137	0.151
Cys	67	380	8	0.119	0.021	67	38	1	0.015	0.026
Asp	1188	585	82	0.069	0.140	1188	1162	304	0.256	0.262
Glu	1493	573	70	0.047	0.122	1493	1227	208	0.139	0.170
Phe	202	90	4	0.020	0.044	202	331	36	0.178	0.109
Gly	1154	213	192	0.166	0.901	1154	1386	901	0.781	0.650
His	296	63	2	0.007	0.032	296	356	22	0.074	0.062
Ile	324	3938	149	0.460	0.038	324	127	36	0.111	0.283
Lys	1214	87	11	0.009	0.126	1214	962	146	0.120	0.152
Leu	554	4245	200	0.361	0.047	554	340	94	0.170	0.276
Met	157	209	4	0.025	0.019	157	49	4	0.025	0.082
Asn	794	276	24	0.030	0.087	794	1000	118	0.149	0.118
Pro	873	248	67	0.077	0.270	873	947	399	0.457	0.421
Gln	751	81	9	0.012	0.111	751	768	67	0.089	0.087
Arg	872	383	28	0.032	0.073	872	1027	121	0.139	0.118
Ser	1118	138	17	0.015	0.123	1118	1415	251	0.225	0.177
Thr	871	373	35	0.040	0.094	871	815	134	0.154	0.164
Val	500	1277	75	0.150	0.059	500	301	76	0.152	0.252
Trp	74	43	1	0.014	0.023	74	171	9	0.122	0.053
Tyr	249	319	23	0.092	0.072	249	408	46	0.185	0.113
Total	**13639**	**13639**	**1013**	**0.074**	**0.074**	**13639**	**13639**	**3095**	**0.227**	**0.227**

Table S3. Summary of the weights for each EvoEF and EvoEF2 energy term. The extended terms in EvoEF2 are highlighted in bold.

Classification	Energy terms	EvoEF weight	EvoEF2 weight
Intra-residue interactions	Van der Waals attractive	0.0000	0.43
	Van der Waals repulsive	0.1200	0.06
	Coulomb's electrostatics	0.0000	0.29
	desolvP	0.0000	0.00
	desolvH	0.0000	0.34
	HBsb_dist	0.0000	0.83
	HBsb_theta	0.0400	0.28
	HBsb_phi	0.0012	0.00
	Amino_acid_propensity	-	**0.59**
	Ramachandran	-	**0.42**
	DunbrackRot	-	**0.35**
Inter-residue interactions in the same chain	Van der Waals attractive	0.6076	1.21
	Van der Waals repulsive	0.4968	1.28
	Coulomb's electrostatics	0.0000	2.31
	desolvP	0.3008	0.75
	desolvH	0.0322	4.59
	SSbond	-	**2.72**
	HBbb_dist	0.3554	1.02
	HBbb_theta	0.0000	1.01
	HBbb_phi	0.0000	1.07
	HBsb_dist	0.5452	0.85
	HBsb_theta	0.2080	0.91
	HBsb_phi	0.1600	0.17
	HBss_dist	0.3036	1.19
	HBss_theta	0.0800	0.71
	HBss_phi	0.0000	0.00
Inter-residue interactions in different chains	Van der Waals attractive	0.6384	1.06
	Van der Waals repulsive	0.7904	0.80
	Coulomb's electrostatics	0.0000	2.44
	desolvP	0.4048	0.68
	desolvH	0.3432	4.79
	SSbond	-	**1.07**
	HBbb_dist	0.0000	1.01
	HBbb_theta	0.7600	1.01
	HBbb_phi	0.5000	1.02
	HBsb_dist	0.0000	0.94
	HBsb_theta	0.6552	0.70
	HBsb_phi	0.5236	0.32
	HBss_dist	0.7200	0.94
	HBss_theta	0.6720	0.97
	HBss_phi	0.5040	0.00

Table S4. Summary of the EvoEF and EvoEF2 reference energies.

Amino acid	EvoEF reference energy	EvoEF2 reference energy
ALA	1.200	-0.408
CYS	0.600	-0.111
ASP	1.200	-0.802
GLU	1.000	-1.225
PHE	2.000	0.679
GLY	2.000	-2.093
HIS	2.200	-0.295
ILE	-0.120	2.330
LYS	1.200	-1.250
LEU	0.000	1.613
MET	1.000	0.759
ASN	1.000	-2.155
PRO	0.960	-0.647
GLN	1.400	-1.936
ARG	0.720	-1.322
SER	1.200	-1.978
THR	0.800	-0.416
VAL	0.240	1.700
TRP	2.600	2.004
TYR	1.600	0.700

Table S5. Folding assessment results obtained by I-TASSER for the 148 test set monomers.

PDB ID	Sequence identity (%)	TM-score	RMSD (Å)
1AM2	28.2	0.983	0.949
1AQB	29.7	0.991	0.475
1BDO	33.8	0.962	0.722
1BK7	32.1	0.993	0.442
1C3D	37.1	0.980	0.972
1CTF	38.2	0.952	0.706
1EVS	28.8	0.931	0.860
1EW4	29.2	0.973	0.672
1IC6	42.3	0.998	0.306
1KUH	37.9	0.997	0.252
1LTU	26.5	0.933	1.630
1MD6	34.4	0.965	0.953
1NOA	40.7	0.958	0.864
1NWA	39.5	0.972	0.901
1OPD	41.2	0.981	0.468
1OW1	32.9	0.979	0.783
1P3C	36.7	0.997	0.313
1PGV	25.7	0.942	1.273
1PPN	33.5	0.995	0.393
1PS4	34.4	0.989	0.613
1QQP	30.3	0.979	0.759
1R12	33.1	0.994	0.470
1RLJ	34.8	0.992	0.381
1RW	33.2	0.987	0.701
1S7K	26.6	0.880	1.410
1SEI	32.3	0.962	0.930
1SMX	40.2	0.889	1.586
1T0F	26.5	0.999	0.160
1TC5	35.8	0.998	0.226
1THV	37.2	0.992	0.619
1UAI	36.8	0.998	0.272
1UEB	38.0	0.998	0.240
1V05	37.6	0.973	0.594
1V30	32.1	0.968	0.751
1V6T	40.2	0.997	0.322
1V7Q	33.1	0.964	1.001
1V8E	41.5	0.989	0.592
1V8I	35.3	0.905	2.263
1VIE	28.3	0.942	0.881
1WV8	39.1	0.941	0.726
1WVH	31.8	0.932	1.389
1WY6	32.7	0.996	0.302
1X1E	41.8	0.992	0.550
1XDZ	24.4	0.991	0.565
1XFS	24.8	0.872	2.483
1XS5	41.7	0.998	0.286
1YCK	26.9	0.988	0.554
1YHH	33.2	0.990	0.619
1YPF	35.3	0.932	2.537

1YU5	28.4	0.929	1.338
1YW5	33.3	0.998	0.227
1YZX	29.8	0.988	0.633
1Z0C	42.5	0.958	1.202
1ZEQ	31.2	0.961	0.682
1ZUH	35.8	0.932	2.497
1ZZK	33.8	0.851	2.856
2A8F	27.6	0.978	0.547
2AHE	33.3	0.961	1.258
2AVU	24.4	0.996	0.285
2AYH	35.5	0.989	0.635
2B0A	35.5	0.998	0.230
2B2A	25.2	0.957	2.357
2BK8	28.9	0.886	1.689
2BS2	28.5	0.994	0.459
2BV9	32.7	0.999	0.216
2CAY	25.8	0.993	0.357
2CBA	36.8	0.995	0.516
2CG7	35.6	0.966	0.652
2CWC	41.2	0.993	0.677
2CWR	39.2	0.948	1.118
2CWZ	31.6	0.904	1.656
2CZS	27.1	0.985	0.364
2D4P	37.6	0.988	0.447
2DRI	31.4	0.981	0.877
2E8F	23.0	0.982	0.509
2E8G	34.6	0.999	0.168
2FBQ	29.1	0.999	0.201
2FI9	47.5	0.989	0.450
2FJZ	25.4	0.906	0.955
2FL4	31.3	0.949	1.196
2FQ3	27.1	0.984	0.431
2FRG	30.2	0.986	0.467
2G69	35.4	0.925	1.222
2G7S	25.7	0.911	2.044
2H2R	28.1	0.813	3.653
2H5N	25.4	0.998	0.167
2HP7	25.1	0.972	0.985
2HU9	36.2	0.996	0.264
2I3F	37.4	0.999	0.199
2I6V	33.3	0.806	1.969
2IBL	26.9	0.684	3.288
2IIA	31.5	0.970	1.216
2IL5	22.8	0.781	2.750
2J6A	34.6	0.996	0.268
2NS9	30.4	0.985	0.574
2OEB	31.6	0.993	0.375
2OFC	25.5	0.996	0.291
2OHW	35.9	0.995	0.292
2OL7	32.8	0.896	2.080
2OSA	28.1	0.980	0.867

2OVO	32.1	0.720	2.030
2P5D	27.6	0.969	0.936
2PBP	31.8	0.939	1.925
2PET	31.2	0.997	0.316
2PMR	25.0	0.992	0.273
2PND	31.1	0.973	0.706
2PTV	41.7	0.908	1.518
2Q7A	28.3	0.998	0.230
2QHK	36.1	0.867	2.010
2QKH	21.3	0.883	2.009
2QQ4	37.7	0.996	0.291
2QR3	37.2	0.890	1.714
2QSI	34.6	0.955	0.847
2QV5	38.2	0.998	0.278
2RIK	31.7	0.854	2.723
2TGI	27.7	0.956	0.943
2VC8	31.9	0.976	0.543
2VQ2	32.3	0.973	1.057
2W2U	24.0	0.951	0.942
2YXF	30.3	0.869	3.048
2Z37	34.8	0.994	0.451
3B7H	22.4	0.847	2.233
3BN6	34.2	0.986	0.579
3DEE	22.2	0.939	1.508
3E35	36.1	0.963	1.514
3FNI	30.1	0.916	1.719
3FYB	25.0	0.933	1.145
3GK6	28.2	0.976	0.791
3H0D	27.2	0.873	2.195
3HWP	30.8	0.959	1.429
3IR4A	34.1	0.969	1.070
3KNV	30.9	0.996	0.270
3M7O	35.0	0.933	1.353
3MQZ	32.3	0.989	0.586
3NRW	31.4	0.987	0.457
3NUL	36.2	0.896	1.693
3PFG	36.1	0.986	0.789
3PO0	31.0	0.988	0.373
3RHT	33.3	0.983	0.809
3RJT	29.8	0.917	1.877
3RKV	23.8	0.979	0.666
3RMQ	30.8	0.993	0.331
3RNV	29.3	0.997	0.213
3RPC	26.6	0.949	1.496
3RPE	30.3	0.997	0.282
3RTL	28.1	0.990	0.407
3VUB	33.7	0.990	0.366
4PTI	24.1	0.929	0.995

Table S6. 29 X-ray/NMR structure pairs used for protein sequence design.

PDB ID		X-ray resolution(Å)	Number of NMR structures	Number of residues	Sequence identity (%) of the designed sequence based on X-ray structure to the native Sequence
X-ray	NMR				
1AGI	1GIO	1.5	10	125	27.2
1BED	2IJY	2.0	22	181	23.8
1BP2	1BVM	1.7	20	123	22.8
1C44	1QND	1.5	20	123	27.6
1CHN	1DJM	1.78	27	126	36.5
1EKG	1LY7	1.8	15	119	31.9
1FKJ	1FKR	1.7	20	102	44.9
1GNU	1KOT	1.75	15	117	37.6
1GPR	1AX3	1.9	16	158	41.1
1GSV	3PHY	1.75	26	122	32.0
1HCV	1G9E	1.85	20	116	34.5
1IFB	1AEL	1.96	20	131	26.7
1IFR	1IVT	1.4	15	111	31.0
1J2A	1CLH	1.8	12	159	34.9
1JF4	1VRE	1.4	29	140	32.0
1KF5	2AAS	1.15	32	124	34.7
1KM8	1BC4	1.9	15	105	29.1
1LDS	1JNJ	1.8	20	97	25.8
1MG4	1UF0	1.5	20	101	32.7
1OPC	2JPB	1.95	20	99	39.4
1RRO	2NLN	1.3	20	108	34.3
1UOH	1TR4	2.0	20	223	39.5
1WHO	1BMW	1.9	38	89	31.9
1YV6	1PU3	1.78	20	103	27.2
2CWR	2CZN	1.7	38	97	30.9
2E1F	2DGZ	2.0	20	94	33.0
2GRC	2H60	1.5	11	105	30.6
2SAK	1SSN	1.8	20	121	29.8
2V75	2JPS	1.8	45	89	32.6

Table S7. Native sequence recapitulation results from designing 88 test set dimers using EvoEF2-mon. EvoEF2-mon is the EvoEF2 energy function that uses all of the EvoEF2 energy terms but with weights determined by designing monomer proteins (see main text).

Residue	EvoEF2-mon					EvoEF2-mon (88 split monomers)				
	#nat	#des	#id	#id/#nat	#id/#des	#nat	#des	#id	#id/#nat	#id/#des
All residues										
Ala	1040	1311	381	0.366	0.291	1040	1098	346	0.333	0.315
Cys	266	315	34	0.128	0.108	266	229	40	0.150	0.175
Asp	786	702	186	0.237	0.265	786	768	195	0.248	0.254
Glu	908	779	130	0.143	0.167	908	939	129	0.142	0.137
Phe	549	575	198	0.361	0.344	549	568	178	0.324	0.313
Gly	1014	1239	798	0.787	0.644	1014	1128	770	0.759	0.683
His	331	416	47	0.142	0.113	331	372	40	0.121	0.108
Ile	739	722	273	0.369	0.378	739	679	270	0.365	0.398
Lys	799	630	107	0.134	0.170	799	772	101	0.126	0.131
Leu	1242	1230	600	0.483	0.488	1242	1169	595	0.479	0.509
Met	245	276	27	0.110	0.098	245	231	28	0.114	0.121
Asn	599	602	104	0.174	0.173	599	622	100	0.167	0.161
Pro	709	626	299	0.422	0.478	709	700	335	0.472	0.479
Gln	570	543	50	0.088	0.092	570	500	39	0.068	0.078
Arg	651	576	92	0.141	0.160	651	738	99	0.152	0.134
Ser	822	781	154	0.187	0.197	822	847	164	0.200	0.194
Thr	792	702	169	0.213	0.241	792	722	146	0.184	0.202
Val	978	1084	412	0.421	0.380	978	949	383	0.392	0.404
Trp	203	185	31	0.153	0.168	203	213	28	0.138	0.131
Tyr	495	444	97	0.196	0.218	495	494	93	0.188	0.188
Total	**13738**	**13738**	**4189**	**0.305**	**0.305**	**13738**	**13738**	**4079**	**0.297**	**0.297**
Core residues										
Ala	348	416	181	0.520	0.435	348	404	198	0.569	0.490
Cys	132	136	22	0.167	0.162	132	127	30	0.227	0.236
Asp	70	72	19	0.271	0.264	70	75	29	0.414	0.387
Glu	45	101	5	0.111	0.050	45	100	10	0.222	0.100
Phe	218	170	92	0.422	0.541	218	184	93	0.427	0.505
Gly	200	194	151	0.755	0.778	200	178	148	0.740	0.831
His	63	72	11	0.175	0.153	63	63	12	0.190	0.190
Ile	322	330	153	0.475	0.464	322	349	169	0.525	0.484
Lys	49	57	9	0.184	0.158	49	58	10	0.204	0.172
Leu	501	506	295	0.589	0.583	501	550	335	0.669	0.609
Met	82	105	9	0.110	0.086	82	120	16	0.195	0.133
Asn	73	41	15	0.205	0.366	73	39	19	0.260	0.487

Pro	102	83	52	0.510	0.627	102	93	59	0.578	0.634
Gln	38	39	3	0.079	0.077	38	44	2	0.053	0.045
Arg	58	37	10	0.172	0.270	58	51	14	0.241	0.275
Ser	140	72	18	0.129	0.250	140	67	26	0.186	0.388
Thr	134	163	43	0.321	0.264	134	130	38	0.284	0.292
Val	413	498	223	0.540	0.448	413	454	232	0.562	0.511
Trp	73	48	13	0.178	0.271	73	40	10	0.137	0.250
Tyr	161	82	29	0.180	0.354	161	96	26	0.161	0.271
Total	**3222**	**3222**	**1353**	**0.420**	**0.420**	**3222**	**3222**	**1476**	**0.458**	**0.458**
Surface residues										
Ala	281	296	44	0.157	0.149	281	238	19	0.068	0.080
Cys	26	31	1	0.038	0.032	26	18	1	0.038	0.056
Asp	369	305	75	0.203	0.246	369	327	79	0.214	0.242
Glu	477	282	53	0.111	0.188	477	339	58	0.122	0.171
Phe	53	100	9	0.170	0.090	53	90	10	0.189	0.111
Gly	345	517	284	0.823	0.549	345	469	288	0.835	0.614
His	82	120	6	0.073	0.050	82	116	5	0.061	0.043
Ile	76	50	4	0.053	0.080	76	35	0	0.000	0.000
Lys	368	238	36	0.098	0.151	368	293	40	0.109	0.137
Leu	138	117	24	0.174	0.205	138	99	24	0.174	0.242
Met	39	21	0	0.000	0.000	39	14	1	0.026	0.071
Asn	229	323	37	0.162	0.115	229	325	29	0.127	0.089
Pro	294	268	106	0.361	0.396	294	287	119	0.405	0.415
Gln	250	234	17	0.068	0.073	250	211	15	0.060	0.071
Arg	219	268	28	0.128	0.104	219	297	28	0.128	0.094
Ser	360	431	72	0.200	0.167	360	430	76	0.211	0.177
Thr	289	231	38	0.131	0.165	289	248	36	0.125	0.145
Val	131	104	18	0.137	0.173	131	88	12	0.092	0.136
Trp	17	44	3	0.176	0.068	17	52	2	0.118	0.038
Tyr	55	118	6	0.109	0.051	55	122	7	0.127	0.057
Total	**4098**	**4098**	**861**	**0.210**	**0.210**	**4098**	**4098**	**849**	**0.207**	**0.207**
Interface residues										
Ala	159	281	76	0.478	0.270	159	175	35	0.220	0.200
Cys	45	83	7	0.156	0.084	45	31	3	0.067	0.097
Asp	174	137	45	0.259	0.328	174	178	39	0.224	0.219
Glu	181	163	40	0.221	0.245	181	215	18	0.099	0.084
Phe	133	133	41	0.308	0.308	133	113	22	0.165	0.195
Gly	182	263	149	0.819	0.567	182	229	121	0.665	0.528
His	92	105	13	0.141	0.124	92	79	11	0.120	0.139

Ile	138	148	50	0.362	0.338	138	91	34	0.246	0.374
Lys	130	115	21	0.162	0.183	130	195	16	0.123	0.082
Leu	222	231	107	0.482	0.463	222	155	63	0.284	0.406
Met	53	75	10	0.189	0.133	53	24	5	0.094	0.208
Asn	140	123	26	0.186	0.211	140	158	27	0.193	0.171
Pro	151	127	64	0.424	0.504	151	154	64	0.424	0.416
Gln	123	102	15	0.122	0.147	123	126	8	0.065	0.063
Arg	167	88	24	0.144	0.273	167	168	15	0.090	0.089
Ser	136	122	24	0.176	0.197	136	202	23	0.169	0.114
Thr	152	137	41	0.270	0.299	152	139	23	0.151	0.165
Val	170	201	72	0.424	0.358	170	140	45	0.265	0.321
Trp	60	46	11	0.183	0.239	60	66	11	0.183	0.167
Tyr	150	78	25	0.167	0.321	150	120	30	0.200	0.250
Total	**2758**	**2758**	**861**	**0.312**	**0.312**	**2758**	**2758**	**613**	**0.222**	**0.222**

Table S8. Native sequence recapitulation results from designing 88 test set dimers using EvoEF and the final optimized EvoEF2 energy function.

Residue	EvoEF					EvoEF2				
	#nat	#des	#id	#id/#nat	#id/#des	#nat	#des	#id	#id/#nat	#id/#des
All residues										
Ala	1040	821	206	0.198	0.251	1040	1333	388	0.373	0.291
Cys	266	791	31	0.117	0.039	266	328	31	0.117	0.095
Asp	786	1173	110	0.140	0.094	786	693	184	0.234	0.266
Glu	908	924	88	0.097	0.095	908	785	127	0.140	0.162
Phe	549	230	59	0.107	0.257	549	565	182	0.332	0.322
Gly	1014	565	382	0.377	0.676	1014	1209	805	0.794	0.666
His	331	265	19	0.057	0.072	331	391	46	0.139	0.118
Ile	739	1951	214	0.290	0.110	739	764	284	0.384	0.372
Lys	799	141	17	0.021	0.121	799	665	122	0.153	0.183
Leu	1242	2116	330	0.266	0.156	1242	1191	589	0.474	0.495
Met	245	302	2	0.008	0.007	245	276	25	0.102	0.091
Asn	599	515	49	0.082	0.095	599	617	107	0.179	0.173
Pro	709	390	125	0.176	0.321	709	622	312	0.440	0.502
Gln	570	200	18	0.032	0.090	570	523	69	0.121	0.132
Arg	651	343	26	0.040	0.076	651	548	86	0.132	0.157
Ser	822	523	70	0.085	0.134	822	755	149	0.181	0.197
Thr	792	646	78	0.098	0.121	792	718	167	0.211	0.233
Val	978	1352	245	0.251	0.181	978	1110	427	0.437	0.385
Trp	203	99	13	0.064	0.131	203	190	28	0.138	0.147
Tyr	495	391	63	0.127	0.161	495	455	112	0.226	0.246
Total	**13738**	**13738**	**2145**	**0.156**	**0.156**	**13738**	**13738**	**4240**	**0.309**	**0.309**
Core residues										
Ala	348	323	107	0.307	0.331	348	446	192	0.552	0.430
Cys	132	225	13	0.098	0.058	132	138	19	0.144	0.138
Asp	70	367	20	0.286	0.054	70	78	21	0.300	0.269
Glu	45	229	6	0.133	0.026	45	93	8	0.178	0.086
Phe	218	59	26	0.119	0.441	218	171	80	0.367	0.468
Gly	200	185	126	0.630	0.681	200	179	147	0.735	0.821
His	63	76	7	0.111	0.092	63	63	8	0.127	0.127
Ile	322	333	103	0.320	0.309	322	362	161	0.500	0.445
Lys	49	11	1	0.020	0.091	49	51	12	0.245	0.235
Leu	501	321	122	0.244	0.380	501	500	296	0.591	0.592
Met	82	33	0	0.000	0.000	82	108	10	0.122	0.093
Asn	73	127	9	0.123	0.071	73	37	15	0.205	0.405

Pro	102	107	37	0.363	0.346	102	97	63	0.618	0.649
Gln	38	38	2	0.053	0.053	38	39	2	0.053	0.051
Arg	58	23	0	0.000	0.000	58	44	9	0.155	0.205
Ser	140	171	16	0.114	0.094	140	68	18	0.129	0.265
Thr	134	171	21	0.157	0.123	134	139	45	0.336	0.324
Val	413	324	113	0.274	0.349	413	476	221	0.535	0.464
Trp	73	24	8	0.110	0.333	73	43	11	0.151	0.256
Tyr	161	75	23	0.143	0.307	161	90	28	0.174	0.311
Total	**3222**	**3222**	**760**	**0.236**	**0.236**	**3222**	**3222**	**1366**	**0.424**	**0.424**
Surface residues										
Ala	281	100	10	0.036	0.100	281	323	33	0.117	0.102
Cys	26	198	2	0.077	0.010	26	40	1	0.038	0.025
Asp	369	216	30	0.081	0.139	369	313	72	0.195	0.230
Glu	477	211	29	0.061	0.137	477	233	45	0.094	0.193
Phe	53	45	1	0.019	0.022	53	98	8	0.151	0.082
Gly	345	69	50	0.145	0.725	345	508	290	0.841	0.571
His	82	63	4	0.049	0.063	82	137	7	0.085	0.051
Ile	76	819	13	0.171	0.016	76	46	2	0.026	0.043
Lys	368	78	9	0.024	0.115	368	265	41	0.111	0.155
Leu	138	900	32	0.232	0.036	138	95	24	0.174	0.253
Met	39	133	0	0.000	0.000	39	22	2	0.051	0.091
Asn	229	119	9	0.039	0.076	229	340	36	0.157	0.106
Pro	294	79	22	0.075	0.278	294	271	111	0.378	0.410
Gln	250	71	7	0.028	0.099	250	247	34	0.136	0.138
Arg	219	180	12	0.055	0.067	219	229	27	0.123	0.118
Ser	360	95	12	0.033	0.126	360	424	70	0.194	0.165
Thr	289	171	20	0.069	0.117	289	252	44	0.152	0.175
Val	131	426	18	0.137	0.042	131	112	21	0.160	0.188
Trp	17	22	0	0.000	0.000	17	41	3	0.176	0.073
Tyr	55	103	3	0.055	0.029	55	102	7	0.127	0.069
Total	**4098**	**4098**	**283**	**0.069**	**0.069**	**4098**	**4098**	**878**	**0.214**	**0.214**
Interface residues										
Ala	159	191	50	0.314	0.262	159	257	73	0.459	0.284
Cys	45	175	8	0.178	0.046	45	60	5	0.111	0.083
Asp	174	271	35	0.201	0.129	174	157	46	0.264	0.293
Glu	181	202	27	0.149	0.134	181	192	37	0.204	0.193
Phe	133	51	14	0.105	0.275	133	121	39	0.293	0.322
Gly	182	151	88	0.484	0.583	182	245	150	0.824	0.612
His	92	58	3	0.033	0.052	92	83	16	0.174	0.193

Ile	138	328	43	0.312	0.131	138	153	56	0.406	0.366
Lys	130	19	2	0.015	0.105	130	128	23	0.177	0.180
Leu	222	340	70	0.315	0.206	222	246	104	0.468	0.423
Met	53	56	2	0.038	0.036	53	71	6	0.113	0.085
Asn	140	123	15	0.107	0.122	140	104	28	0.200	0.269
Pro	151	94	35	0.232	0.372	151	109	58	0.384	0.532
Gln	123	37	5	0.041	0.135	123	91	13	0.106	0.143
Arg	167	55	6	0.036	0.109	167	111	25	0.150	0.225
Ser	136	133	26	0.191	0.195	136	133	30	0.221	0.226
Thr	152	129	16	0.105	0.124	152	125	31	0.204	0.248
Val	170	242	47	0.276	0.194	170	222	76	0.447	0.342
Trp	60	20	2	0.033	0.100	60	44	10	0.167	0.227
Tyr	150	83	18	0.120	0.217	150	106	36	0.240	0.340
Total	**2758**	**2758**	**512**	**0.186**	**0.186**	**2758**	**2758**	**862**	**0.313**	**0.313**

Table S9. Native sequence recapitulation results from designing 132 test set dimers using EvoEF and the final optimized EvoEF2 energy function.

Residue	EvoEF					EvoEF2				
	#nat	#des	#id	#id/#nat	#id/#des	#nat	#des	#id	#id/#nat	#id/#des
All residues										
Ala	1728	1320	365	0.211	0.277	1728	2251	685	0.396	0.304
Cys	382	1334	44	0.115	0.033	382	549	56	0.147	0.102
Asp	1232	1667	175	0.142	0.105	1232	999	296	0.240	0.296
Glu	1513	1437	147	0.097	0.102	1513	1176	232	0.153	0.197
Phe	820	342	70	0.085	0.205	820	812	256	0.312	0.315
Gly	1549	858	601	0.388	0.700	1549	1704	1129	0.729	0.663
His	467	366	23	0.049	0.063	467	540	62	0.133	0.115
Ile	1222	3026	362	0.296	0.120	1222	1319	487	0.399	0.369
Lys	1330	244	37	0.028	0.152	1330	938	162	0.122	0.173
Leu	1967	3356	574	0.292	0.171	1967	2047	1015	0.516	0.496
Met	465	474	14	0.030	0.030	465	473	47	0.101	0.099
Asn	859	868	55	0.064	0.063	859	860	138	0.161	0.160
Pro	934	564	176	0.188	0.312	934	816	406	0.435	0.498
Gln	750	329	21	0.028	0.064	750	782	80	0.107	0.102
Arg	1008	463	48	0.048	0.104	1008	801	126	0.125	0.157
Ser	1164	790	107	0.092	0.135	1164	1148	204	0.175	0.178
Thr	1203	968	93	0.077	0.096	1203	1153	212	0.176	0.184
Val	1711	2026	433	0.253	0.214	1711	1873	778	0.455	0.415
Trp	239	133	12	0.050	0.090	239	289	45	0.188	0.156
Tyr	664	642	97	0.146	0.151	664	677	141	0.212	0.208
Total	**21207**	**21207**	**3454**	**0.163**	**0.163**	**21207**	**21207**	**6557**	**0.309**	**0.309**
Core residues										
Ala	595	530	178	0.299	0.336	595	797	361	0.607	0.453
Cys	198	440	30	0.152	0.068	198	255	32	0.162	0.125
Asp	122	558	37	0.303	0.066	122	116	38	0.311	0.328
Glu	101	345	19	0.188	0.055	101	140	21	0.208	0.150
Phe	296	88	24	0.081	0.273	296	220	104	0.351	0.473
Gly	381	303	209	0.549	0.690	381	293	246	0.646	0.840
His	82	100	7	0.085	0.070	82	74	14	0.171	0.189
Ile	543	533	145	0.267	0.272	543	627	263	0.484	0.419
Lys	79	12	1	0.013	0.083	79	100	13	0.165	0.130
Leu	784	524	216	0.276	0.412	784	813	483	0.616	0.594
Met	139	69	4	0.029	0.058	139	166	22	0.158	0.133
Asn	92	216	14	0.152	0.065	92	56	14	0.152	0.250

Pro	138	153	41	0.297	0.268	138	114	74	0.536	0.649
Gln	78	66	3	0.038	0.045	78	60	10	0.128	0.167
Arg	86	22	2	0.023	0.091	86	45	8	0.093	0.178
Ser	210	271	29	0.138	0.107	210	88	21	0.100	0.239
Thr	253	257	24	0.095	0.093	253	220	58	0.229	0.264
Val	777	605	219	0.282	0.362	777	869	445	0.573	0.512
Trp	87	24	3	0.034	0.125	87	58	15	0.172	0.259
Tyr	198	123	30	0.152	0.244	198	128	39	0.197	0.305
Total	**5239**	**5239**	**1235**	**0.236**	**0.236**	**5239**	**5239**	**2281**	**0.435**	**0.435**
Surface residues										
Ala	459	162	24	0.052	0.148	459	524	58	0.126	0.111
Cys	24	325	2	0.083	0.006	24	68	1	0.042	0.015
Asp	587	288	43	0.073	0.149	587	439	118	0.201	0.269
Glu	801	345	42	0.052	0.122	801	422	88	0.110	0.209
Phe	67	66	0	0.000	0.000	67	154	10	0.149	0.065
Gly	521	111	84	0.161	0.757	521	748	399	0.766	0.533
His	119	84	4	0.034	0.048	119	185	7	0.059	0.038
Ile	100	1212	35	0.350	0.029	100	78	17	0.170	0.218
Lys	659	133	16	0.024	0.120	659	380	66	0.100	0.174
Leu	214	1406	58	0.271	0.041	214	153	34	0.159	0.222
Met	86	208	3	0.035	0.014	86	49	2	0.023	0.041
Asn	387	238	12	0.031	0.050	387	508	68	0.176	0.134
Pro	416	134	45	0.108	0.336	416	364	148	0.356	0.407
Gln	311	98	5	0.016	0.051	311	376	29	0.093	0.077
Arg	332	243	15	0.045	0.062	332	359	35	0.105	0.097
Ser	465	137	12	0.026	0.088	465	635	94	0.202	0.148
Thr	400	262	17	0.043	0.065	400	371	50	0.125	0.135
Val	205	594	30	0.146	0.051	205	195	29	0.141	0.149
Trp	25	29	1	0.040	0.034	25	67	4	0.160	0.060
Tyr	77	180	8	0.104	0.044	77	180	15	0.195	0.083
Total	**6255**	**6255**	**456**	**0.073**	**0.073**	**6255**	**6255**	**1272**	**0.203**	**0.203**
Interface residues										
Ala	278	321	86	0.309	0.268	278	405	123	0.442	0.304
Cys	55	237	6	0.109	0.025	55	114	10	0.182	0.088
Asp	251	350	48	0.191	0.137	251	199	65	0.259	0.327
Glu	258	333	41	0.159	0.123	258	248	52	0.202	0.210
Phe	212	91	32	0.151	0.352	212	196	70	0.330	0.357
Gly	258	241	152	0.589	0.631	258	316	201	0.779	0.636
His	135	85	6	0.044	0.071	135	131	23	0.170	0.176

Ile	218	504	68	0.312	0.135	218	246	86	0.394	0.350
Lys	247	38	6	0.024	0.158	247	177	39	0.158	0.220
Leu	389	602	132	0.339	0.219	389	458	215	0.553	0.469
Met	114	81	5	0.044	0.062	114	134	13	0.114	0.097
Asn	168	186	14	0.083	0.075	168	131	24	0.143	0.183
Pro	186	125	46	0.247	0.368	186	181	97	0.522	0.536
Gln	163	55	5	0.031	0.091	163	131	23	0.141	0.176
Arg	273	68	15	0.055	0.221	273	144	35	0.128	0.243
Ser	206	183	38	0.184	0.208	206	201	37	0.180	0.184
Thr	252	214	32	0.127	0.150	252	245	53	0.210	0.216
Val	304	350	81	0.266	0.231	304	343	144	0.474	0.420
Trp	69	31	4	0.058	0.129	69	69	13	0.188	0.188
Tyr	181	122	29	0.160	0.238	181	148	38	0.210	0.257
Total	**4217**	**4217**	**846**	**0.201**	**0.201**	**4217**	**4217**	**1361**	**0.323**	**0.323**

Supporting Figures

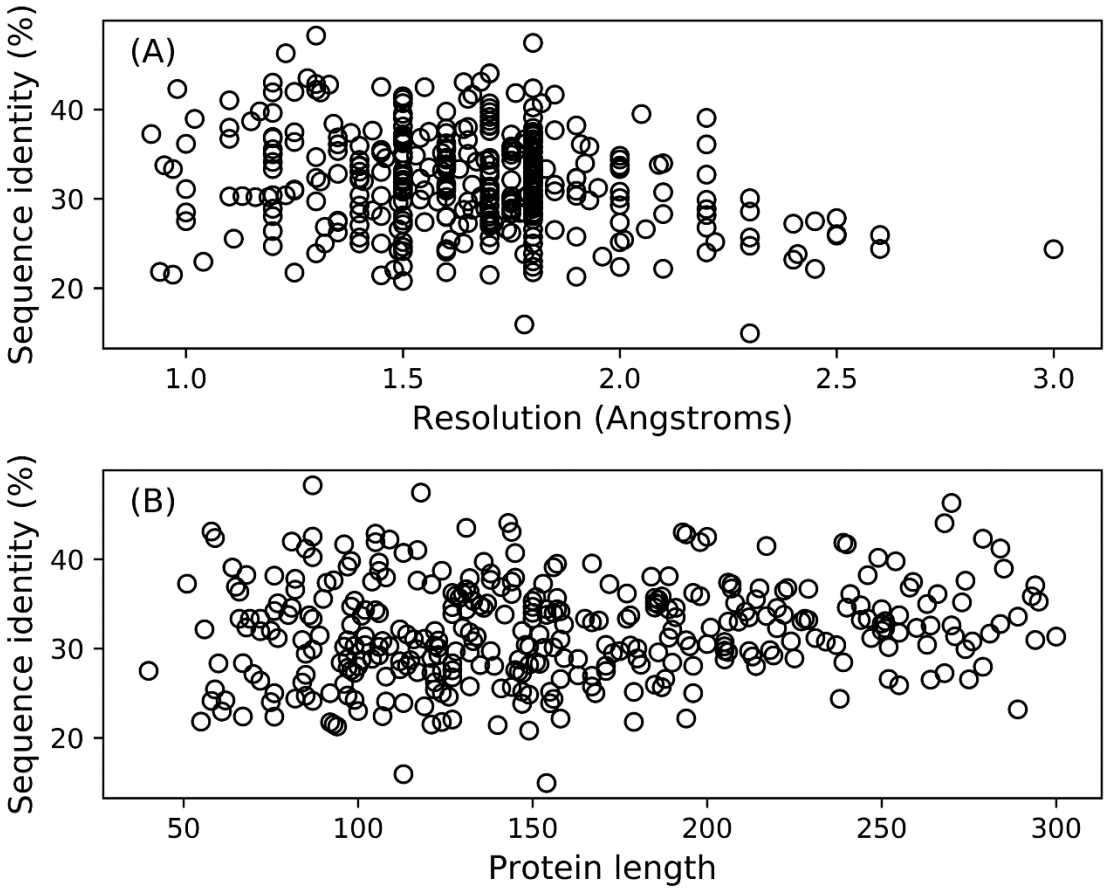

Fig. S1. Sequence identity between the native and designed sequences using EvoEF2 as a function of crystal structure resolution (A) and protein length (B).

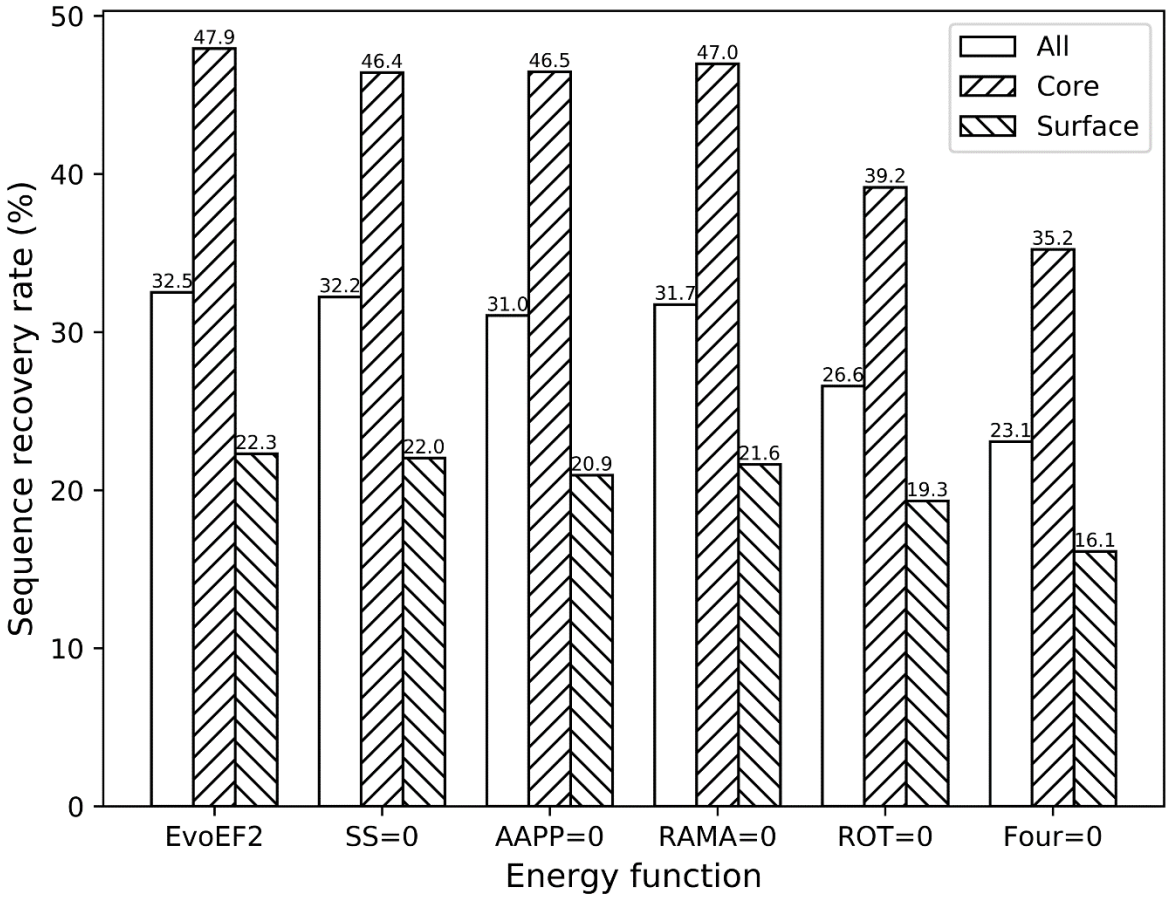

Fig. S2. Sequence recovery rates for the 148 test set monomers using different EvoEF2 energy functions. The columns marked with 'EvoEF2', 'SS=0', 'AAPP=0', 'RAMA=0', 'ROT=0' and 'Four=0' signify designs performed using the complete EvoEF2 energy function or disabling the disulfide bond term, amino acid propensity term, Ramachandran energy term, Dunbrack rotamer probability term, or each of the four new terms, respectively.

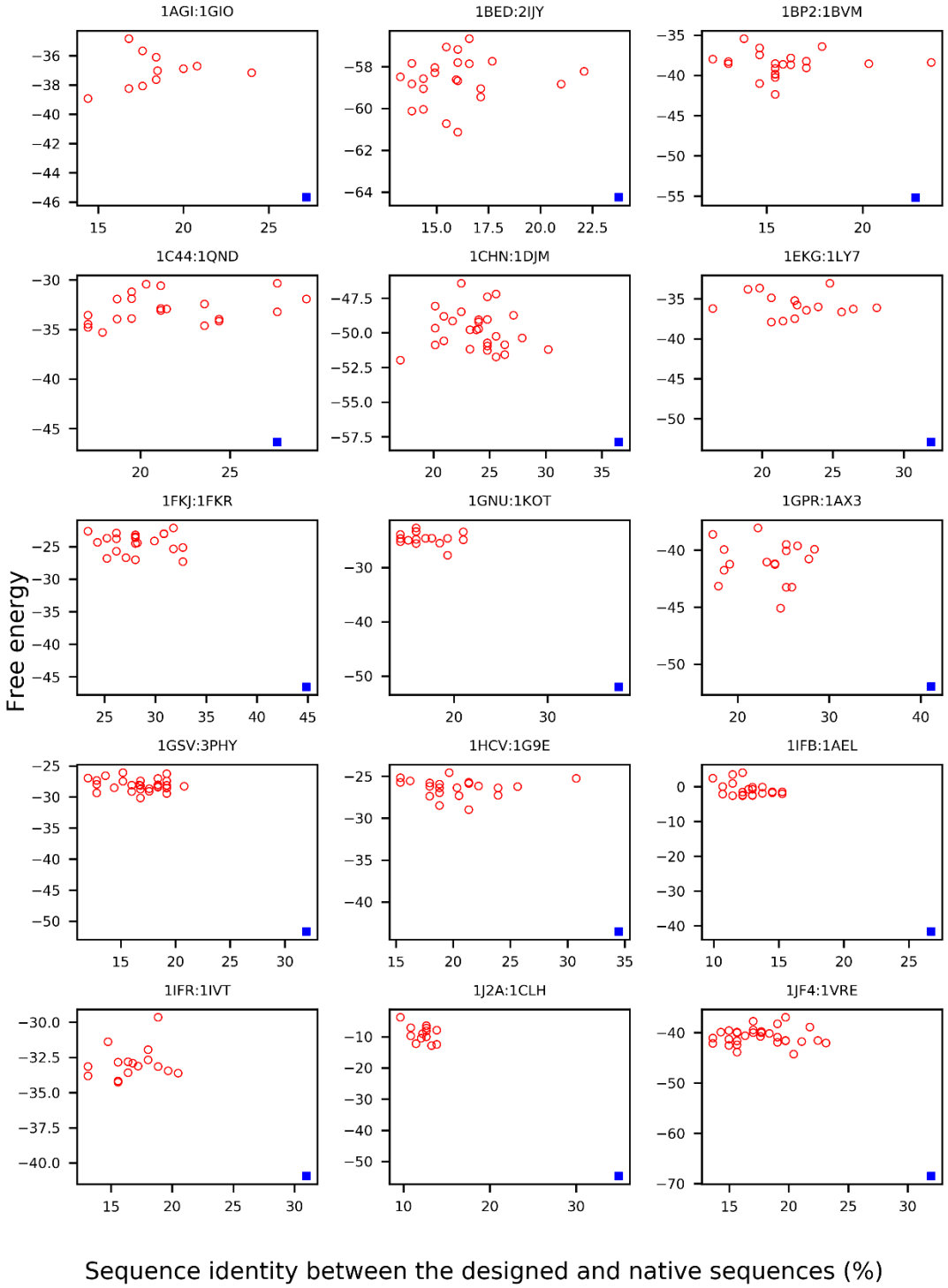

Fig. S3. (continued on next page)

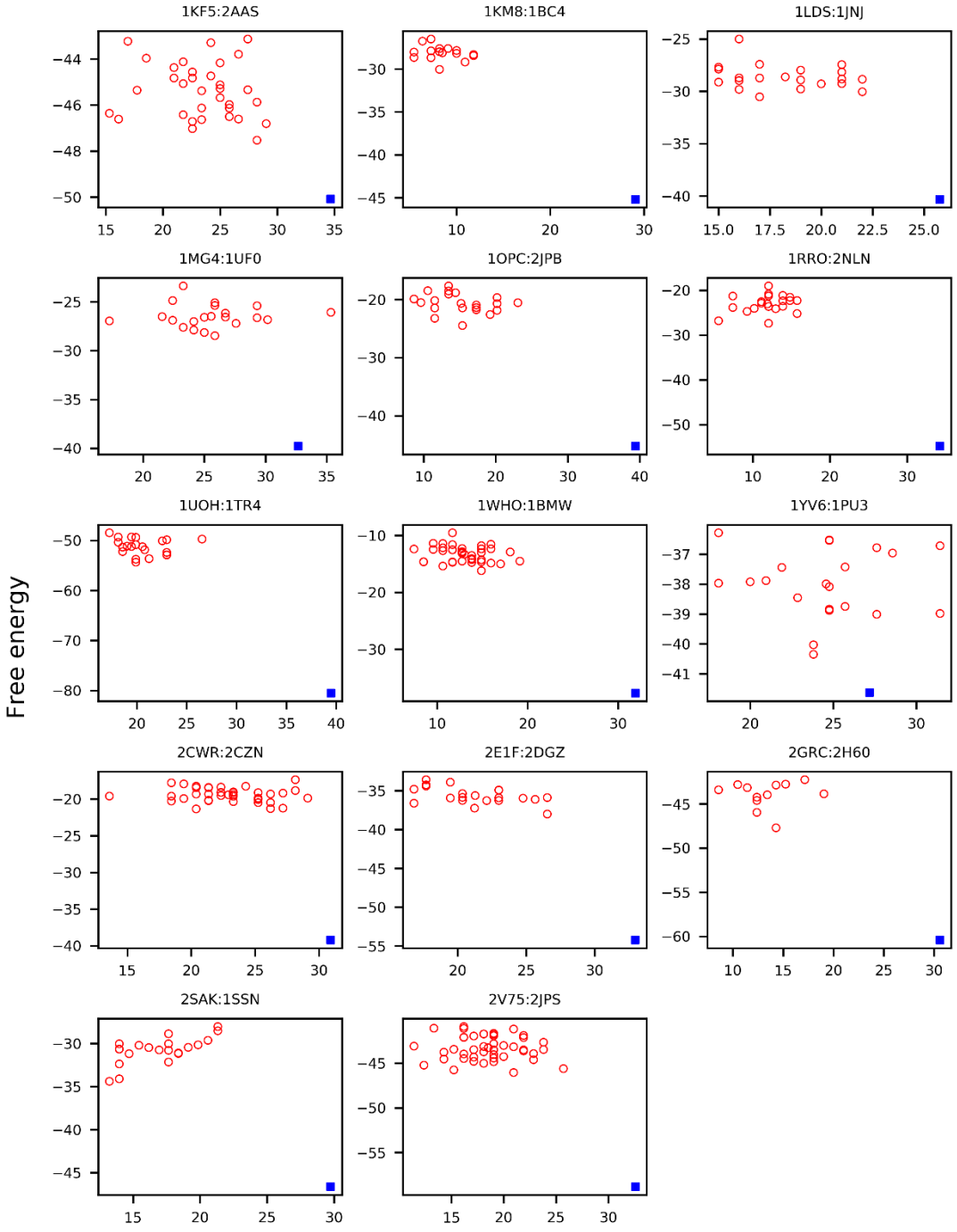

Fig. S3. Sequence design for 29 X-ray/NMR structure pairs. The free energy (in EvoEF2 energy units) is plotted against the sequence identity (over all residues) between the designed and native sequences. In each subplot, the first PDB ID is the X-ray structure code and the second is the NMR PDB code. The results for the X-ray crystal structures are shown as filled blue squares, while the results for the NMR structures are shown as open red circles.

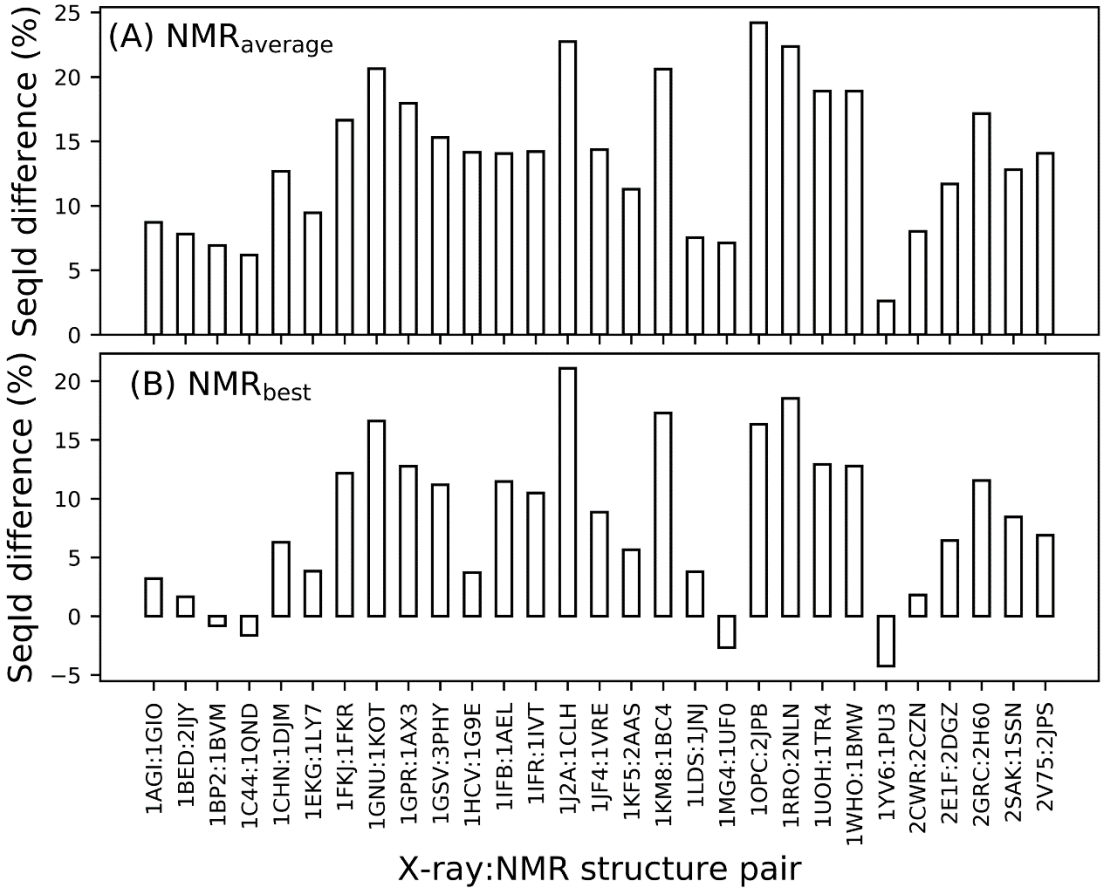

Fig. S4. Sequence design performance on X-ray and NMR templates using EvoEF2. The difference of sequence identity on the Y axis is calculated as SequenceIdentity$_{(X\text{-}ray)}$ - SequenceIdentity$_{(NMR)}$. (A) Performance on the X-ray template is compared to the average performance over all NMR templates in the corresponding ensemble. (B) Performance on the X-ray template is compared to performance on the best performing NMR template.

References

Brooks, B.R., *et al.* (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, *J. Comput. Chem.*, **4**, 187-217.

Chen, V.B., *et al.* (2010) MolProbity: all-atom structure validation for macromolecular crystallography, *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 12-21.

Ding, F. and Dokholyan, N.V. (2006) Emergence of protein fold families through rational design, *PLoS Comput. Biol.*, **2**, e85.

Guerois, R., Nielsen, J.E. and Serrano, L. (2002) Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations, *J. Mol. Biol.*, **320**, 369-387.

Jankauskaite, J., *et al.* (2018) SKEMPI 2.0: An updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation, *Bioinformatics*, **35**, 462-469.

Jones, J.E. (1924) On the determination of molecular fields. I. From the variation of the viscosity of a gas with temperature, *Proc. R. Soc. Lond. A*, **106**, 441-462.

Jones, J.E. (1924) On the determination of molecular fields. II. From the equation of state of a gas, *Proc. R. Soc. Lond. A*, **106**, 463-477.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983) Optimization by simulated annealing, *Science*, **220**, 671-680.

Kortemme, T., Morozov, A.V. and Baker, D. (2003) An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein–Protein Complexes, *J. Mol. Biol.*, **326**, 1239-1259.

Krivov, G.G., Shapovalov, M.V. and Dunbrack, R.L., Jr. (2009) Improved prediction of protein side-chain conformations with SCWRL4, *Proteins*, **77**, 778-795.

Kuhlman, B. and Baker, D. (2000) Native protein sequences are close to optimal for their structures, *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 10383-10388.

Lazaridis, T. and Karplus, M. (1999) Effective energy function for proteins in solution, *Proteins*, **35**, 133-152.

Leaver-Fay, A., *et al.* (2013) Scientific benchmarks for guiding macromolecular energy function improvement, *Methods Enzymol.*, **523**, 109-143.

Pearce, R., *et al.* (2019) EvoDesign: Designing protein-protein binding interactions using evolutionary interface profiles in conjunction with an optimized physical energy function, *J. Mol. Biol.*, **431**, 2467-2476.

Quan, L., Lv, Q. and Zhang, Y. (2016) STRUM: structure-based prediction of protein stability changes upon single-point mutation, *Bioinformatics*, **32**, 2936-2946.

Shapovalov, M.V. and Dunbrack, R.L., Jr. (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions, *Structure*, **19**, 844-858.