JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design

Rebecca F. Alford,[†] Andrew Leaver-Fay,[‡] Jeliazko R. Jeliazkov,[§] Matthew J. O'Meara,[‖] Frank P. DiMaio,[⊥] Hahnbeom Park,[#] Maxim V. Shapovalov,[∇] P. Douglas Renfrew,[○,◆] Vikram K. Mulligan,[#] Kalli Kappel,[¶] Jason W. Labonte,[†] Michael S. Pacella,[+] Richard Bonneau,[○,◆] Philip Bradley,[Δ] Roland L. Dunbrack, Jr.,[∇] Rhiju Das,[¶] David Baker,[#,∞] Brian Kuhlman,[‡] Tanja Kortemme,[◇] and Jeffrey J. Gray[*,†,§]

[†]Department of Chemical and Biomolecular Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, United States

[‡]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, 120 Mason Farm Road, Chapel Hill, North Carolina 27599, United States

[§]Program in Molecular Biophysics, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, United States

[‖]Department of Pharmaceutical Chemistry, University of California at San Francisco, 1700 Fourth Street, San Francisco, California 94158, United States

[⊥]Department of Biochemistry, University of Washington, J-Wing, Health Sciences Building, Box 357350, Seattle, Washington 98195, United States

[#]Department of Biochemistry, University of Washington, Molecular Engineering and Sciences, Box 351655, 3946 West Stevens Way NE, Seattle, Washington 98195, United States

[∇]Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, Pennsylvania 19111, United States

[○]Department of Biology, Center for Genomics and Systems Biology, New York University, 100 Washington Square East, New York, New York 10003, United States

[◆]Center for Computational Biology, Flatiron Institute, Simons Foundation, 162 Fifth Avenue, New York, New York 10010, United States

[¶]Biophysics Program, Stanford University, 450 Serra Mall, Stanford, California 94305, United States

[+]Department of Biomedical Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, United States

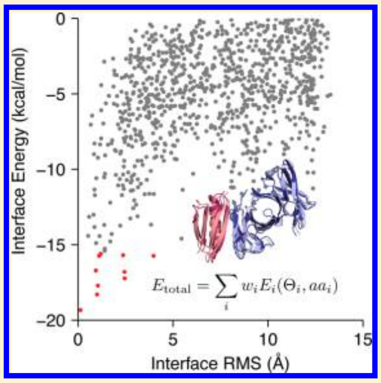
[Δ]Computational Biology Program, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, United States

[◇]Department of Bioengineering and Therapeutic Sciences, University of California at San Francisco, San Francisco, California 94158, United States

[∞]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, United States

Supporting Information

ABSTRACT: Over the past decade, the Rosetta biomolecular modeling suite has informed diverse biological questions and engineering challenges ranging from interpretation of low-resolution structural data to design of nanomaterials, protein therapeutics, and vaccines. Central to Rosetta's success is the energy function: a model parametrized from small-molecule and X-ray crystal structure data used to approximate the energy associated with each biomolecule conformation. This paper describes the mathematical models and physical concepts that underlie the latest Rosetta energy function, called the Rosetta Energy Function 2015 (REF15). Applying these concepts, we explain how to use Rosetta energies to identify and analyze the features of biomolecular models. Finally, we discuss the latest advances in the energy function that extend its capabilities from soluble proteins to also include membrane proteins, peptides containing noncanonical amino acids, small molecules, carbohydrates, nucleic acids, and other macromolecules.



Received: February 6, 2017
Published: April 21, 2017

DOI: 10.1021/acs.jctc.7b00125

INTRODUCTION

Proteins adopt diverse three-dimensional conformations to carry out the complex mechanisms of life. Their structures are constrained by the underlying amino acid sequence[1] and stabilized by a fine balance between enthalphic and entropic contributions to non-covalent interactions.[2] Energy functions that seek to approximate the energies of these interactions are fundamental to computational modeling of biomolecular structures. The goal of this paper is to describe the energy calculations used by the Rosetta macromolecular modeling program:[3] we explain the underlying physical concepts, mathematical models, latest advances, and application to biomolecular simulations.

Energy functions are based on Anfinsen's hypothesis that native-like protein conformations represent unique, low-energy, thermodynamically stable conformations.[4] These folded states reside in minima on the energy landscape, and they have a net favorable change in Gibbs free energy, which is the sum of contributions from both enthalpy (ΔH) and entropy ($T\Delta S$), relative to the unfolded state. To follow these heuristics, macromolecular modeling programs require a mathematical function that can discriminate between the unfolded, folded, and native-like conformations. Typically, these functions are linear combinations of terms that compute energies as a function of various degrees of freedom.

The earliest macromolecular energy functions combined a Lennard-Jones potential for van der Waals interactions[5−7] with harmonic torsional potentials[8] that were parametrized using force constants from vibrational spectra of small molecules.[9−11] These formulations were first applied to investigations of the structures of hemolysin,[12] trypsin inhibitor,[13] and hemoglobin[14] and have now diversified into a large family of commonly used energy functions such as AMBER,[15] DREIDING,[16] OPLS,[17] and CHARMM.[18,19] Many of these energy functions also rely on new terms and parametrizations. For example, faster computers have enabled the derivation of parameters from ab initio quantum calculations.[20] The maturation of X-ray crystallography and NMR protein structure determination methods has enabled the development of statistical potentials derived from per-residue, inter-residue, secondary-structure, and whole-structure features.[21−28] Additionally, there are alternate models of electrostatics and solvation, such as the generalized Born approximation of the Poisson−Boltzmann equation[29] and polarizable electrostatic terms that accommodate varying charge distributions.[30]

The first version of the Rosetta energy function was developed for proteins by Simons et al.[31] Initially, it used statistical potentials describing individual residue environments and frequent residue-pair interactions derived from the Protein Data Bank (PDB).[32] Later the authors added terms for packing of van der Waals spheres and hydrogen-bonding, secondary-structure, and van der Waals interactions to improve the performance of ab initio structure prediction.[33] These terms were for low-resolution modeling, meaning that the scores were dependent on only the coordinates of the backbone atoms and that interactions between the side chains were treated implicitly.

To enable higher-resolution modeling, in the early 2000s Kuhlman and Baker[34] implemented an all-atom energy function that emphasized atomic packing, hydrogen bonding, solvation, and protein torsion angles commonly found in folded proteins. This energy function first included a Lennard-Jones term,[35] a pairwise-additive implicit solvation model,[36] a statistically derived electrostatics term, and a term for backbone-dependent rotamer preferences.[37] Shortly thereafter, several terms were added, including an orientation-dependent hydrogen-bonding term,[38] in agreement with electronic structure calculations.[39] This combination of traditional molecular mechanics energies and statistical torsion potentials enabled Rosetta to reach several milestones in structure prediction and design, including accurate ab initio structure prediction,[40] hot-spot prediction,[41,42] protein−protein docking,[43] small-molecule docking,[44] and specificity redesign[45] as well as the first de novo designed protein backbone not found in nature[46] and the first computationally designed new protein−protein interface.[47]

The Rosetta energy function has changed dramatically since it was last described in complete detail by Rohl et al.[48] in 2004. It has undergone significant advances ranging from improved models of hydrogen bonding[49] and solvation[50] to updated evaluation of backbone[51] and rotamer conformations.[52] Along the way, these developments have enabled Rosetta to address new biomolecular modeling problems, including the refinement of low-resolution X-ray structures and use of sparse data[53,54] and the design of vaccines,[55] biomineralization peptides,[56] self-assembling materials,[57] and enzymes that perform new functions.[58,59] Instead of arbitrary units, the energy function is now also fitted to estimate energies in kilocalories per mole. The details of the energy function advances are distributed across code comments, methods development papers, application papers, and individual experts, making it challenging for Rosetta developers and users in both academia and industry to learn the underlying concepts. Moreover, members of the Rosetta community are actively working to generalize the all-atom energy function for use in different contexts[60,61] and for all macromolecules, including RNA,[62] DNA,[63,64] small-molecule ligands,[65,66] noncanonical amino acids and backbones,[67−69] and carbohydrates,[70] further encouraging us to reexamine the underpinnings of the energy function. Thus, there is a need for an up-to-date description of the current energy function.

In this paper, we describe the new default energy function, called the Rosetta Energy Function 2015 (REF15). Our discussion aims to expose the physical and mathematical details of the energy function required for rigorous understanding. In addition, we explain how to apply the computed energies to analyze structural models produced by Rosetta simulations. We hope this paper will provide critically needed documentation of the energy methods as well as an educational resource to help students and scientists interpret the results of these simulations.

COMPUTING THE TOTAL ROSETTA ENERGY

The Rosetta energy function approximates the energy of a biomolecule conformation. This quantity, called ΔE_{total}, is computed as a linear combination of energy terms E_i that are calculated as functions of geometric degrees of freedom (Θ) and chemical identities (aa) and scaled by a weight on each term (w_i), as shown in eq 1:

$$\Delta E_{\text{total}} = \sum_i w_i E_i(\Theta_i, \text{aa}_i) \tag{1}$$

Here we explain the Rosetta energy function term by term. First, we describe the energies of interactions between nonbonded atom pairs, which are important for atomic packing, electrostatics, and solvation. Second, we explain the

Table 1. Summary of Terms in REF15 for Proteins

term	description	weight	units	ref(s)
`fa_atr`	attractive energy between two atoms on different residues separated by a distance d	1.0	kcal/mol	5, 6
`fa_rep`	repulsive energy between two atoms on different residues separated by a distance d	0.55	kcal/mol	5, 6
`fa_intra_rep`	repulsive energy between two atoms on the same residue separated by a distance d	0.005	kcal/mol	5, 6
`fa_sol`	Gaussian exclusion implicit solvation energy between protein atoms in different residues	1.0	kcal/mol	36
`lk_ball_wtd`	orientation-dependent solvation of polar atoms assuming ideal water geometry	1.0	kcal/mol	50, 71
`fa_intra_sol`	Gaussian exclusion implicit solvation energy between protein atoms in the same residue	1.0	kcal/mol	36
`fa_elec`	energy of interaction between two nonbonded charged atoms separated by a distance d	1.0	kcal/mol	50
`hbond_lr_bb`	energy of short-range hydrogen bonds	1.0	kcal/mol	38, 49
`hbond_sr_bb`	energy of long-range hydrogen bonds	1.0	kcal/mol	38, 49
`hbond_bb_sc`	energy of backbone−side-chain hydrogen bonds	1.0	kcal/mol	38, 49
`hbond_sc`	energy of side-chain−side-chain hydrogen bonds	1.0	kcal/mol	38, 49
`dslf_fa13`	energy of disulfide bridges	1.25	kcal/mol	49
`rama_prepro`	probability of backbone ϕ, ψ angles given the amino acid type	(0.45 kcal/mol)/kT	kT	50, 51
`p_aa_pp`	probability of amino acid identity given backbone ϕ, ψ angles	(0.4 kcal/mol)/kT	kT	51
`fa_dun`	probability that a chosen rotamer is native-like given backbone ϕ, ψ angles	(0.7 kcal/mol)/kT	kT	52
`omega`	backbone-dependent penalty for cis ω dihedrals that deviate from 0° and trans ω dihedrals that deviate from 180°	(0.6 kcal/mol)/AU	AU[a]	72
`pro_close`	penalty for an open proline ring and proline ω bonding energy	(1.25 kcal/mol)/AU	AU	51
`yhh_planarity`	sinusoidal penalty for nonplanar tyrosine χ_3 dihedral angle	(0.625 kcal/mol)/AU	AU	49
`ref`	reference energies for amino acid types	(1.0 kcal/mol)/AU	AU	1, 51

[a] AU = arbitrary units.

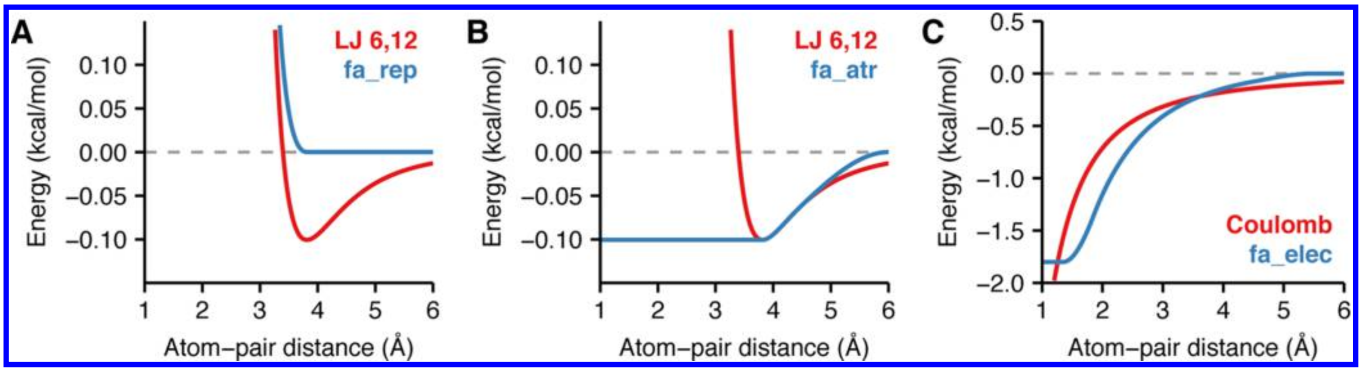


Figure 1. Van der Waals and electrostatics energies. Shown are comparisons between the pairwise energies of nonbonded atoms computed by Rosetta and the forms computed by traditional molecular mechanics force fields. Here the interactions between the backbone nitrogen and carbon are used as examples. (A) Lennard-Jones van der Waals energy with well-depths $\varepsilon_{\mathrm{Nbb}}$ = 0.162 kcal/mol and $\varepsilon_{\mathrm{Cbb}}$ = 0.063 kcal/mol and atomic radii r_{Nbb} = 1.763 Å and r_{Cbb} = 2.011 Å (red) and Rosetta `fa_rep` (blue). (B) Lennard-Jones van der Waals energy (red) and Rosetta `fa_atr` (blue). As the atom-pair distance approaches 6.0 Å, the `fa_atr` term smoothly approaches zero and deviates slightly from the original Lennard-Jones potential. (C) Coulomb electrostatic energy with a dielectric constant ϵ = 10 and partial charges q_{Nbb} = −0.604e and q_{Cbb} = 0.090e (red) compared with Rosetta `fa_elec` (blue). The `fa_elec` model is shifted to reach zero at the cutoff distance of 6.0 Å.

empirical potentials used to model hydrogen and disulfide bonds. Next, we explain the statistical potentials used to describe backbone and side-chain torsional preferences in proteins. Then we explain a set of terms that accommodate features that are not explicitly captured yet are important for native structural feature recapitulation. Finally, we discuss how the energy terms are combined into a single function used to approximate the energy of biomolecules. For reference, items in the fixed-width font are names of energy terms in the Rosetta code. The energy terms are summarized in Table 1.

Terms for Atom-Pair Interactions. *Van der Waals Interactions.* Van der Waals interactions are short-range attractive and repulsive forces that vary with atom-pair distance. Whereas attractive forces result from the cross-correlated motions of electrons in neighboring nonbonded atoms, repulsive forces occur because electrons cannot occupy the same orbitals by the Pauli exclusion principle. To model van der Waals interactions, Rosetta uses the Lennard-Jones (LJ) 6−12 potential,[5,6] which calculates the interaction energy of atoms i and j in different residues given the sum of their atomic radii, $\sigma_{i,j}$ ($\sigma_{i,j}$ in Rosetta has the same definition as $r_{i,j}^{\min}$ in CHARMM), the atom-pair distance, $d_{i,j}$, and the geometric mean of their well depths, $\varepsilon_{i,j}$ (eq 2):

$$E_{\mathrm{vdw}}(i,j) = \varepsilon_{i,j}\left[\left(\frac{\sigma_{i,j}}{d_{i,j}}\right)^{12} - 2\left(\frac{\sigma_{i,j}}{d_{i,j}}\right)^{6}\right] \tag{2}$$

The atomic radii and well depths are derived from small-molecule liquid-phase data optimized in the context of the energy model.[50]

Rosetta splits the LJ potential at the function's minimum ($d_{i,j} = \sigma_{i,j}$) into two components that can be weighted separately: attractive (`fa_atr`) and repulsive (`fa_rep`). By decomposing the function this way, we can alter component weights without changing the minimum-energy distance or introducing any derivative discontinuities. Many conformational sampling protocols in Rosetta take advantage of this splitting by slowly increasing the weight of the repulsive component to traverse rugged energy landscapes and to prevent structures from unfolding during sampling.[73]

The repulsive van der Waals energy, `fa_rep`, varies as a function of atom-pair distance. At short distances, atomic overlap results in strong forces that lead to large changes in the energy. The steep $1/d_{i,j}{}^{12}$ term can cause poor performance in minimization routines and overall structure prediction and design calculations.[74,75] To alleviate this problem, we weaken the repulsive component by replacing the $1/d_{i,j}{}^{12}$ term with a softer linear term for $d_{i,j} \leq 0.6\sigma_{i,j}$. The term is computed using the atom-type-specific parameters $m_{i,j}$ and $b_{i,j}$, which are fit to ensure derivative continuity at $d_{i,j} = 0.6\sigma_{i,j}$. After the linear component, the function transitions smoothly to the 6−12 form until $d_{i,j} = \sigma_{i,j}$, where it reaches zero and remains zero (eq 3 and Figure 1A):

$$E_{\text{fa\_rep}}(i, j) = \sum_{i,j} w_{i,j}^{\text{conn}} \begin{cases} m_{i,j} d_{i,j} + b_{i,j} & d_{i,j} \leq 0.6\sigma_{i,j} \\ \varepsilon_{i,j}\left[\left(\dfrac{\sigma_{i,j}}{d_{i,j}}\right)^{12} - 2\left(\dfrac{\sigma_{i,j}}{d_{i,j}}\right)^{6} + 1\right] & 0.6\sigma_{i,j} < d_{i,j} \leq \sigma_{i,j} \\ 0 & \sigma_{i,j} < d_{i,j} \end{cases} \tag{3}$$

Rosetta also includes an intraresidue version of the repulsive component, `fa_intra_rep`, with the same functional form as the `fa_rep` term (eq 3). We include this term because the knowledge-based rotamer energy (`fa_dun`; see below) underestimates intraresidue collisions.

The attractive van der Waals energy, `fa_atr`, has a value of $-\varepsilon_{i,j}$ for $d_{i,j} \leq \sigma_{i,j}$ and then transitions to the 6−12 potential as the distance increases (eq 4 and Figure 1B):

$$E_{\text{fa\_atr}} = \sum_{i,j} w_{i,j}^{\text{conn}} \begin{cases} -\varepsilon_{i,j} & d_{i,j} \leq \sigma_{ij} \\ \varepsilon_{i,j}\left[\left(\dfrac{\sigma_{i,j}}{d_{i,j}}\right)^{12} - 2\left(\dfrac{\sigma_{i,j}}{d_{i,j}}\right)^{6}\right] & \sigma_{i,j} < d_{i,j} \leq 4.5\text{ Å} \\ f_{\text{poly}}(d_{i,j}) & 4.5\text{ Å} < d_{i,j} \leq 6.0\text{ Å} \\ 0 & 6.0\text{ Å} < d_{i,j} \end{cases} \tag{4}$$

For speed, we truncate the LJ term beyond $d_{i,j} = 6.0$ Å, where the van der Waals forces are small. To avoid derivative discontinuities, we use a cubic polynomial function, $f_{\text{poly}}(d_{i,j})$, for $d_{i,j} > 4.5$ Å to transition the standard Lennard-Jones functional form smoothly to zero. These smooth derivatives are necessary to ensure that bumps do not accumulate in the distributions of structural features at inflection points in the energy landscape during conformational sampling with gradient-based minimization (Sheffler, unpublished, 2006).

Both terms are multiplied by a connectivity weight, $w_{i,j}^{\text{conn}}$, to exclude the large repulsive energetic contributions that would otherwise be calculated for atoms separated by fewer than four chemical bonds (eq 5).

$$w_{i,j}^{\text{conn}} = \begin{cases} 0 & n_{i,j}^{\text{bonds}} \leq 3 \\ 0.2 & n_{i,j}^{\text{bonds}} = 4 \\ 1 & n_{i,j}^{\text{bonds}} \geq 5 \end{cases} \tag{5}$$

To ensure the connectivity rules include both atoms in a dipole, if an atom is part of a strong dipole, then we count bonds ($n_{i,j}^{\text{bonds}}$) to include both atoms in the dipole. Such weights are common to molecular force fields that assume that covalent bonds are not formed or broken during a simulation. Rosetta uses four chemical bonds as the "crossover" separation when $w_{i,j}^{\text{conn}}$ transitions from 0 to 1 (rather than three chemical bonds as used by traditional force fields) to limit the effects of double counting due to knowledge-based torsional potentials.

The comparison between eq 2 and the modified LJ potential (eqs 3 and 4) is shown in Figure 1A,B.

Electrostatics. Nonbonded electrostatic interactions arise from forces between fully and partially charged atoms. To evaluate these interactions, Rosetta uses Coulomb's law with partial charges originally taken from CHARMM and adjusted via a group optimization scheme (Table S3 in the Supporting Information).[50] Coulomb's law is a pairwise term commonly expressed in terms of $d_{i,j}$, the dielectric constant, ϵ, the partial atomic charges for each atom, q_i and q_j, and Coulomb's constant, C_0 (=322 Å kcal/mol e^{-2}, where e is the elementary charge) (eq 6):

$$E_{\text{Coulomb}}(i, j) = \frac{C_0 q_i q_j}{\epsilon}\frac{1}{d_{i,j}} \tag{6}$$

To approximate electrostatic interactions in biomolecules, we modify the potential to account for the difference in dielectric constant between the protein core and the solvent-exposed surface.[76] Specifically, we substitute the constant ϵ in eq 6 with a sigmoidal function $\epsilon(d_{i,j})$ that increases from $\epsilon_{\text{core}} = 6$ to $\epsilon_{\text{solvent}} = 80$ when $d_{i,j}$ is between 0 and 4 Å (eqs 7 and 8):

$$\epsilon(d_{i,j}) = g\left(\frac{d_{i,j}}{4}\right)\epsilon_{\text{core}} + \left[1 - g\left(\frac{d_{i,j}}{4}\right)\right]\epsilon_{\text{solvent}} \tag{7}$$

$$g(x) = \left(1 + x + \frac{x^2}{2}\right)\exp(-x) \tag{8}$$

As with the van der Waals term, we make several heuristic approximations to adapt this calculation for simulations of biomolecules. To avoid strong repulsive forces at short distances, we replace the steep gradient with the constant $E_{\text{elec}}(d_{\text{min}})$ for $d_{i,j} < 1.45$ Å. Next, since the distance-dependent dielectric assumption results in dampened long-range electrostatics, for speed we truncate the potential at $d_{\text{max}} = 5.5$ Å and modify the Coulomb curve by subtracting $1/d_{\text{max}}{}^2$ to shift the potential to zero at $d_{i,j} = d_{\text{max}}$ (eq 9).

$$E_{\text{elec}}(i, j, d_{i,j}) = \frac{C_0 q_i q_j}{\epsilon(d_{i,j})}\begin{cases} \dfrac{1}{d_{i,j}{}^2} - \dfrac{1}{d_{\text{max}}{}^2} & d_{i,j} \leq d_{\text{max}} \\ 0 & d_{\text{max}} < d_{i,j} \end{cases} \tag{9}$$

We use the cubic polynomials $f_{\text{poly}}^{\text{elec,low}}(d_{i,j})$ and $f_{\text{poly}}^{\text{elec,high}}(d_{i,j})$ to smooth between the traditional form and our adjustments while avoiding derivative discontinuities. The energy is also multiplied by the connectivity weight, $w_{i,j}^{\text{conn}}$ (eq 5). The final

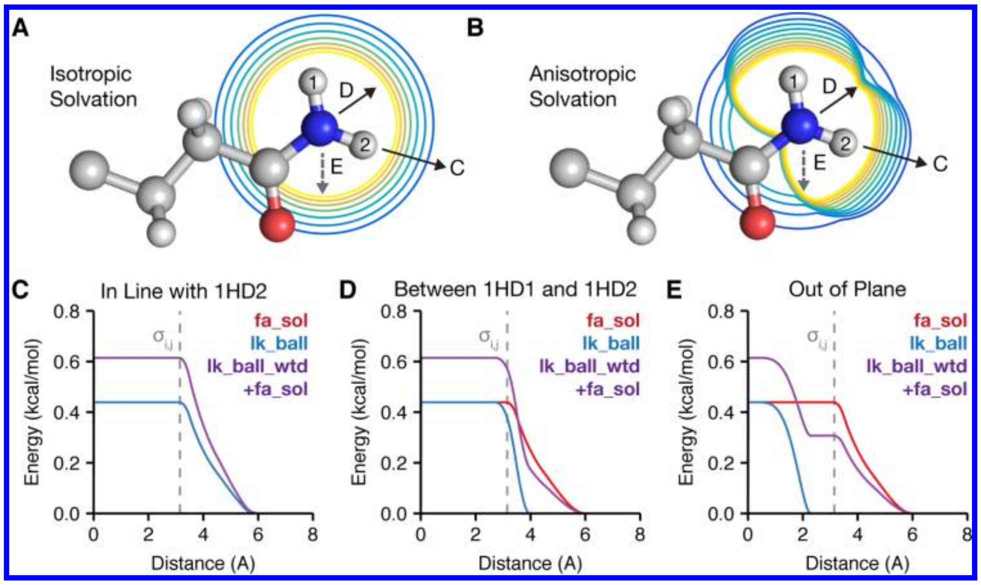


Figure 2. Two-component Lazaridis−Karplus solvation model. Rosetta uses two energy terms to evaluate the desolvation of protein side chains: an isotropic term (`fa_sol`) and an anisotropic term (`lk_ball_wtd`). (A) and (B) demonstrate the difference between isotropic and anisotropic solvation of the NH_2 group by CH_3 on the asparagine side chain. The contours vary from low energy (blue) to high energy (yellow). The arrows represent the approach vectors for the pair potentials shown in (C−E), where we compare the `fa_sol`, `lk_ball`, and `lk_ball_wtd` + `fa_sol` energies for the solvation of the NH_2 group on asparagine for three different approach angles: (C) in line with the 1HD2 atom, (D) along the bisector of the angle between 1HD1 and 1HD2, and (E) vertically down from above the plane of the hydrogens (out of plane).

modified electrostatic potential is given by eq 10 and compared with the standard form in Figure 1C.

$$E_{\text{fa\_elec}} = \sum_{i,j} w_{i,j}^{\text{conn}} \begin{cases} E_{\text{elec}}(i, j, d_{\min}) & d_{i,j} < 1.45\ \text{Å} \\ f_{\text{poly}}^{\text{elec,low}}(d_{i,j}) & 1.45\ \text{Å} \le d_{i,j} < 1.85\ \text{Å} \\ E_{\text{elec}}(i, j, d_{i,j}) & 1.85\ \text{Å} \le d_{i,j} < 4.5\ \text{Å} \\ f_{\text{poly}}^{\text{elec,high}}(d_{i,j}) & 4.5\ \text{Å} \le d_{i,j} < 5.5\ \text{Å} \\ 0 & 5.5\ \text{Å} \le d_{i,j} \end{cases} \tag{10}$$

Solvation. Native-like protein conformations minimize the exposure of hydrophobic side chains to the surrounding polar solvent. Unfortunately, explicitly modeling all of the interactions between solvent and protein atoms is computationally expensive. Instead, Rosetta represents the solvent as bulk water based upon the Lazaridis−Karplus (LK) implicit Gaussian exclusion model.[36] Rosetta's solvation model has two components: an isotropic solvation energy, called `fa_sol`, which assumes that bulk water is uniformly distributed around the atoms (Figure 2A), and an anisotropic solvation energy, called `lk_ball_wtd`, which accounts for specific waters near polar atoms that form the solvation shell (Figure 2B).

The isotropic (LK) model[36] is based on the function f_{desolv} that describes the energy required to desolvate (remove contacting water) an atom i when it is approached by a neighboring atom j. In Rosetta, we exclude the LK ΔG^{ref} term because we implement our own reference energy (discussed later). The energy of the atom-pair interaction varies with the separation distance, $d_{i,j}$, the experimentally determined vapor-to-water transfer free energy, ΔG_i^{free}, the sum of the atomic radii, $\sigma_{i,j}$, the correlation length, λ_i, and the atomic volume of the desolvating atom, V_j (eq 11):

$$f_{\text{desolv}} = -V_j \frac{\Delta G_i^{\text{free}}}{2\pi^{3/2}\lambda_i \sigma_i^2} \exp\left[-\left(\frac{d_{i,j} - \sigma_{i,j}}{\lambda_i}\right)^2\right] \tag{11}$$

At short distances, `fa_rep` prevents atoms from overlapping; however, many protocols briefly downweight or disable the `fa_rep` term. To avoid scenarios where f_{desolv} encourages atom-pair overlap in the absence of `fa_rep`, we smoothly increase the value of the function to a constant at short distances where the van der Waals spheres overlap ($d_{i,j} = \sigma_{i,j}$). At large distances, the function asymptotically approaches zero; therefore, we truncate the function at 6.0 Å for speed. We also transition between the constants at short and long distances using the distance-dependent cubic polynomials $f_{\text{poly}}^{\text{solv,low}}$ and $f_{\text{poly}}^{\text{solv,high}}$ with constants c_0 = 0.3 Å and c_1 = 0.2 Å that define a window for smoothing. The overall desolvation function is given by eq 12:

$$g_{\text{desolv}} = \begin{cases} f_{\text{desolv}}(i, j, \sigma_{i,j}) & d_{i,j} \le \sigma_{i,j} - c_0 \\ f_{\text{poly}}^{\text{solv,low}}(i, j, d_{i,j}) & \sigma_{i,j} - c_0 < d_{i,j} \le \sigma_{i,j} + c_1 \\ f_{\text{desolv}}(i, j, d_{i,j}) & \sigma_{i,j} + c_1 < d_{i,j} \le 4.5\ \text{Å} \\ f_{\text{poly}}^{\text{solv,high}}(i, j, d_{i,j}) & 4.5\ \text{Å} < d_{i,j} \le 6.0\ \text{Å} \\ 0 & 6.0\ \text{Å} < d_{i,j} \end{cases} \tag{12}$$

The total isotropic solvation energy, $E_{\text{fa\_sol}}$, is computed as a sum of the energies for atom j desolvating atom i and vice versa, scaled by the previously defined connectivity weight (eq 13):

$$E_{\text{fa\_sol}} = \sum_{i,j} w_{i,j}^{\text{conn}}[g_{\text{desolv}}(i, j) + g_{\text{desolv}}(j, i)] \tag{13}$$

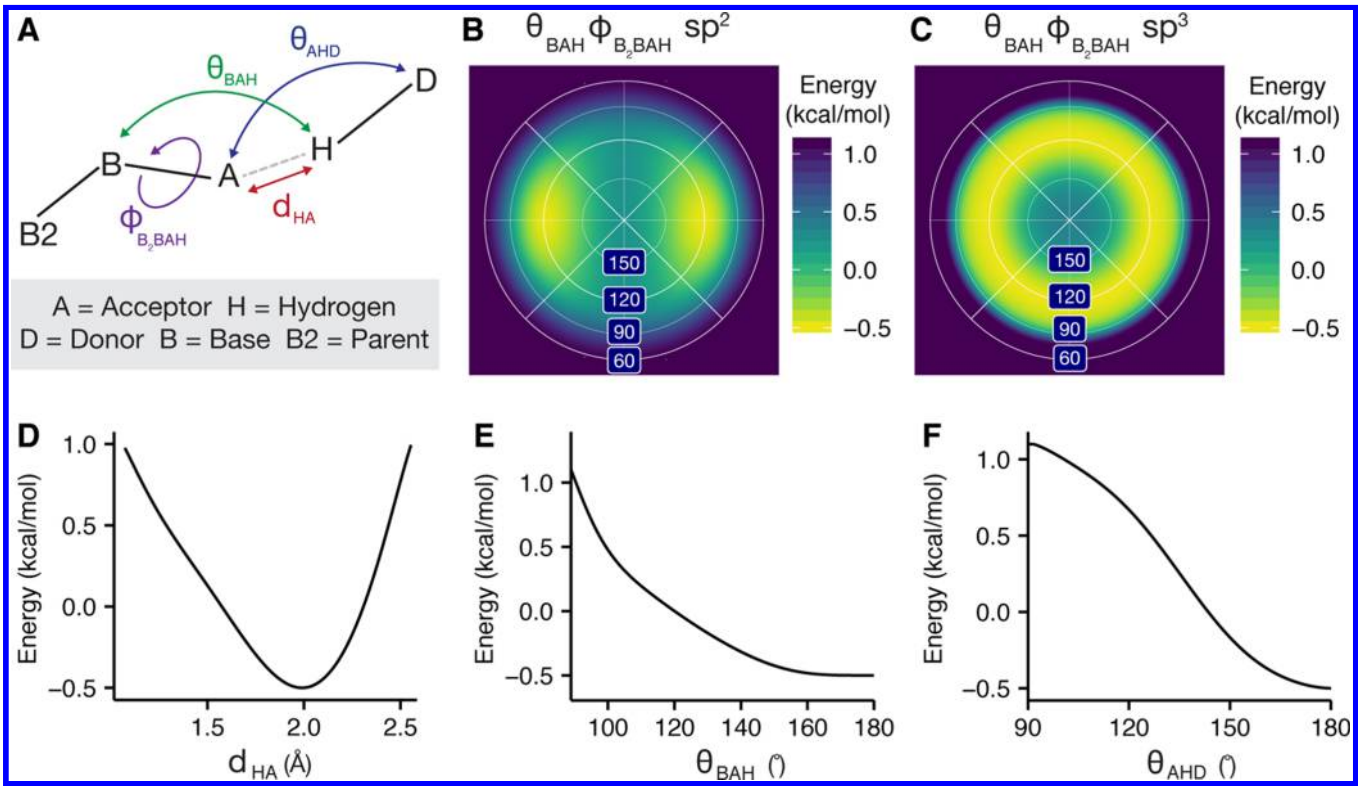


Figure 3. Orientation-dependent hydrogen bonding model. (A) Degrees of freedom evaluated by the hydrogen-bonding term: the acceptor−donor distance, d_{HA}; the angle between the base, acceptor, and hydrogen, θ_{BAH}; the angle between the acceptor, hydrogen, and donor, θ_{AHD}; and the dihedral angle corresponding to rotation around the base−acceptor bond, ϕ_{B_2BAH}. (B) Lambert azimuthal projection of the $E_{\text{hbond}}^{B_2BAH}$ energy landscape for an sp^2-hybridized acceptor.[49] (C) $E_{\text{hbond}}^{B_2BAH}$ energy landscape for an sp^3-hybridized acceptor. (D−F) Example energies for hydrogen bonding of the histidine imidazole ring acceptor with a protein backbone amide: (D) energy E_{hbond}^{HA} vs the acceptor−donor distance d_{HA}; (E) energy E_{hbond}^{BAH} vs the base−acceptor−hydrogen angle θ_{BAH}; (F) energy E_{hbond}^{AHD} vs the acceptor−hydrogen−donor angle θ_{AHD}.

Rosetta also includes an intraresidue version of the isotropic solvation energy, `fa_intra_sol`, with the same functional form as the `fa_sol` term (eq 13).

A recent innovation (2016) is the addition of the energy term `lk_ball_wtd` to model the orientation-dependent solvation of polar atoms. This anisotropic model increases the desolvation penalty for occluding polar atoms near sites where waters may form hydrogen-bonding interactions. For polar atoms, we subtract off part of the isotropic energy given by eq 13 and then add the anisotropic energy to account for the position of the desolvating atom relative to hypothesized water positions.

To compute the anisotropic energy, we first calculate the set of ideal water sites around atom i, $\mathcal{W}_i = \{\boldsymbol{\nu}_{i1}, \boldsymbol{\nu}_{i2}, ...\}$. This set contains one to three water sites, depending on the atom type of atom i. Each site is 2.65 Å from atom i and has an optimal hydrogen-bonding geometry, and we consider the potential overlap of a desolvating atom j with each water. The overlap is considered negligible until the van der Waals sphere of the desolvating atom j (with radius σ_j) touches the van der Waals sphere of the water (with radius σ_w) at site k, and then the term smoothly increases over a zone of partial overlap of approximately 0.5 Å. Thus, for each water site k with coordinates $\boldsymbol{\nu}_{i,k}$, we compute an occlusion measure $d_k^{\ 2}$ to capture the gap between the hypothetical water and the desolvating atom j, using the offset $\Omega = 3.7$ Å^2 to provide the ramp-up buffer (eq 14):

$$d_k^{\ 2} = \| \mathbf{r}_j - \boldsymbol{\nu}_{i,k} \|^2 - (\sigma_w + \sigma_j)^2 + \Omega \qquad (14)$$

Next, we find the soft minimum of $d_k^{\ 2}$ over all water sites in $\mathcal{W}_i$ by computing the logarithmic average:

$$d_{\min}^2(i, j) = -\ln\left[\sum_{k\in\mathcal{W}_i} \exp(-d_k^{\ 2})\right] \qquad (15)$$

Then $d_{\min}^{\ 2}$ and Ω are used to compute a damping function, f_{lkfrac}, that varies from 0 when the desolvating atom is at least a van der Waals distance from any preferred water site to 1 when the desolvating atom overlaps a water site by more than ∼0.5 Å (eq 16):

$$f_{\text{lkfrac}}(i, j) = \begin{cases} 1 & d_{\min}^2(i, j) < 0 \\ \left[1 - \left(\dfrac{d_{\min}^2(i, j)}{\Omega}\right)\right]^2 & 0 \le d_{\min}^2(i, j) < \Omega \\ 0 & \Omega \le d_{\min}^2(i, j) \end{cases} \qquad (16)$$

We calculate the anisotropic energy for desolvating a polar atom, $E_{\text{lk\_ball}}$, by scaling the desolvation function g_{desolv} by the damping function f_{lkfrac} and an atom-type-specific weight, w_{aniso}, which is typically ∼0.7 (eq 17):

$$E_{\text{lk\_ball}}(i, j) = w_{\text{aniso},i}\, g_{\text{desolv}}(i, j) f_{\text{lkfrac}}(i, j) \qquad (17)$$

The amount of isotropic solvation energy subtracted is g_{desolv} multiplied by w_{iso}, where w_{iso} is an atom-type-specific weight, typically ∼0.3 (eq 18):

$$E_{\text{lk\_ball\_iso}}(i, j) = -w_{\text{iso},i} g_{\text{desolv}}(i, j) \tag{18}$$

The total weight on the isotropic contribution through both the `fa_sol` and `lk_ball_wtd` terms is thus ∼0.7. The isotropic and anisotropic components are then summed to yield a new desolvation function, h_{desolv} (eq 19):

$$h_{\text{desolv}}(i, j) = E_{\text{lk\_ball\_iso}}(i, j) + E_{\text{lk\_ball}}(i, j) \tag{19}$$

Like `fa_sol`, the energies for desolvation of atom i by atom j and desolvation of atom j by atom i are summed to yield the overall `lk_ball_wtd` energy, but counting only the desolvation of polar, hydrogen-bonding heavy atoms (O, N), defined as the set $\mathcal{P}$ (eq 20):

$$E_{\text{lk\_ball\_wtd}} = \sum_{i \in \mathcal{P}} w_{i,j}^{\text{conn}} h_{\text{desolv}}(i, j) + \sum_{j \in \mathcal{P}} w_{i,j}^{\text{conn}} h_{\text{desolv}}(j, i) \tag{20}$$

Figure 2 shows comparisons of `fa_sol`, `lk_ball` (eq 17), and the sum of `fa_sol` and `lk_ball_wtd` for the example of an asparagine NH_2 desolvated from three different approach angles. As the approach angle varies, the sum of `lk_ball_wtd` and `fa_sol` creates a larger desolvation penalty when water sites are occluded and a smaller penalty otherwise, relative to `fa_sol` alone.

Hydrogen Bonding. Hydrogen bonds are partially covalent interactions that form when a nucleophilic heavy atom donates electron density to a polar hydrogen.[77] At short ranges (<2.5 Å), they exhibit geometries that maximize orbital overlap.[78] The interactions between hydrogen-bonding groups are also partially described by electrostatics. While this hybrid covalent−electrostatic character is complex, it is crucial for capturing the structural specificity that underlies protein folding, function, and interactions.

Rosetta calculates the energies of hydrogen bonds using `fa_elec` and a hydrogen-bonding model that evaluates energies on the basis of orientation preferences of hydrogen bonds found in high-resolution crystal structures.[38,49] To derive this model, we curated intraprotein polar contacts from ∼8000 high-resolution crystal structures (the Top8000 data set[79]) and identified features using adaptive density estimation. We then empirically fit the functional form of the energy such that the Rosetta-generated polar contacts mimic the distributions from Top8000. The resulting hydrogen-bonding energy is evaluated for all pairs of donor hydrogens, H, and acceptors, A, as a function of four degrees of freedom (Figure 3A): (1) the distance between the donor and acceptor, d_{HA}; (2) the angle θ_{AHD} formed by the acceptor, the donor H, and the donor heavy atom, D; (3) the angle θ_{BAH} formed by the acceptor's parent atom ("base"), B, the acceptor, and the donor; and (4) the torsion angle $\phi_{\text{B}_2\text{BAH}}$ formed by the donor H, the acceptor, and two subsequent parent atoms B and B_2. B, the parent atom of A, is the first atom on the shortest path to the root atom (e.g., C_α). The B_2 atom of A is the parent atom of B (e.g., the sp^2 plane is defined by B_2, B, and A). For convenience, the hydrogen-bonding energy is subdivided into four separate terms: long-range backbone hydrogen bonds (`hbond_lr_bb`), short-range backbone hydrogen bonds (`hbond_sr_bb`), hydrogen bonds between backbone and side-chain atoms (`hbond_bb_sc`), and hydrogen bonds between side-chain atoms (`hbond_sc`).

To avoid overcounting, side chain to backbone hydrogen bonds are excluded if the backbone group is already involved in a hydrogen bond to prevent formation of too many i to i-4 or i to i-3 hydrogen bonds for helical serines and threonines. For speed, the component terms have simple analytic functional forms (Figure 3B−F and eqs S1−S7 in the Supporting Information). The term is also multiplied by two atom-type-specific weights, w_{H} and w_{A}, that account for the varying strength of hydrogen bonds. The overall model is given by eq 21, in which the $E_{\text{hbond}}^{\text{B}_2\text{BAH}}$ term depends on the orbital hybridization of the acceptor, ρ, and the function $f(x)$ (eq 22) is used for smoothing to avoid derivative discontinuities and ensure that edge-case hydrogen bonds are considered:

$$\begin{aligned} E_{\text{hbond}} = \sum_{\text{H,A}} & w_{\text{H}} w_{\text{A}} f(E_{\text{hbond}}^{\text{HA}}(d_{\text{HA}}) + E_{\text{hbond}}^{\text{AHD}}(\theta_{\text{AHD}}) \\ & + E_{\text{hbond}}^{\text{BAH}}(\theta_{\text{BAH}}) + E_{\text{hbond}}^{\text{B}_2\text{BAH}}(\rho, \phi_{\text{B}_2\text{BAH}}, \theta_{\text{BAH}})) \end{aligned} \tag{21}$$

$$f(x) = \begin{cases} x & x < -0.1 \\ -0.025 + \dfrac{x}{2} - 2.5x^2 & -0.1 \le x < 0.1 \\ 0 & 0.1 \le x \end{cases} \tag{22}$$

Disulfide Bonding. Disulfide bonds are covalent interactions that link sulfur atoms in cysteine residues. In Rosetta we typically rely on a tree-based kinematic system[3,80] to keep bond lengths and angles fixed so that we may sample the conformation space by changing only torsions. For this reason, we do not generally need terms that evaluate bond-length and bond-angle energetics. However, with disulfide bonds and proline (discussed below), the extra bonds cannot be represented with a tree (since a tree graph is acyclic) and thus must be treated explicitly. Therefore, disulfide bonds are a special case of inter-residue covalent contacts that requires a representation with more degrees of freedom. To evaluate disulfide-bonding interactions, Rosetta identifies pairs of cysteines that have covalent bonds linking the S_γ atoms and computes the energies of these interactions using an orientation-dependent model called `dslf_fa13`.[49] The model was derived by curating intraprotein disulfide bonds from Top8000 and identifying features using kernel density estimates. For speed, the feature distributions are modeled using skewed Gaussian functions and a mixture of one, two, and three von Mises functions (eqs S8−S11).

The overall disulfide energy is computed as a function of six degrees of freedom (Figure 4) that map to four component energies. First, the component due to the sulfur−sulfur distance, d_{SS}, is evaluated as $E_{\text{dslf}}^{\text{SS}}(d_{\text{SS}})$. Second, the components due to the angles formed by $C_{\beta 1}$ and $C_{\beta 2}$ with the S−S bond, $\theta_{\text{C}_{\beta 1}\text{SS}}$ and $\theta_{\text{C}_{\beta 2}\text{SS}}$, respectively, are evaluated as $E_{\text{dslf}}^{\text{CSS}}(\theta_{\text{C}_\beta\text{SS}})$. Third, the components due to the dihedral angles formed by $C_{\alpha 1}C_{\beta 1}$ and $C_{\alpha 2}C_{\beta 2}$ with the S−S bond, $\phi_{\text{C}_{\alpha 1}\text{C}_{\beta 1}\text{SS}}$ and $\phi_{\text{C}_{\alpha 2}\text{C}_{\beta 2}\text{SS}}$, respectively, are evaluated as $E_{\text{dslf}}^{\text{C}_\alpha\text{C}_\beta\text{SS}}(\phi_{\text{C}_\alpha\text{C}_\beta\text{SS}})$. Finally, the dihedral angle formed by $C_{\beta 1}$, $C_{\beta 2}$, and the S−S bond, $\phi_{\text{C}_{\beta 1}\text{SSC}_{\beta 2}}$, is evaluated as $E_{\text{dslf}}^{\text{C}_\beta\text{SSC}_\beta}(\phi_{\text{C}_{\beta 1}\text{SSC}_{\beta 2}})$. The complete disulfide bonding energy evaluated for all S−S pairs is given by eq 23:

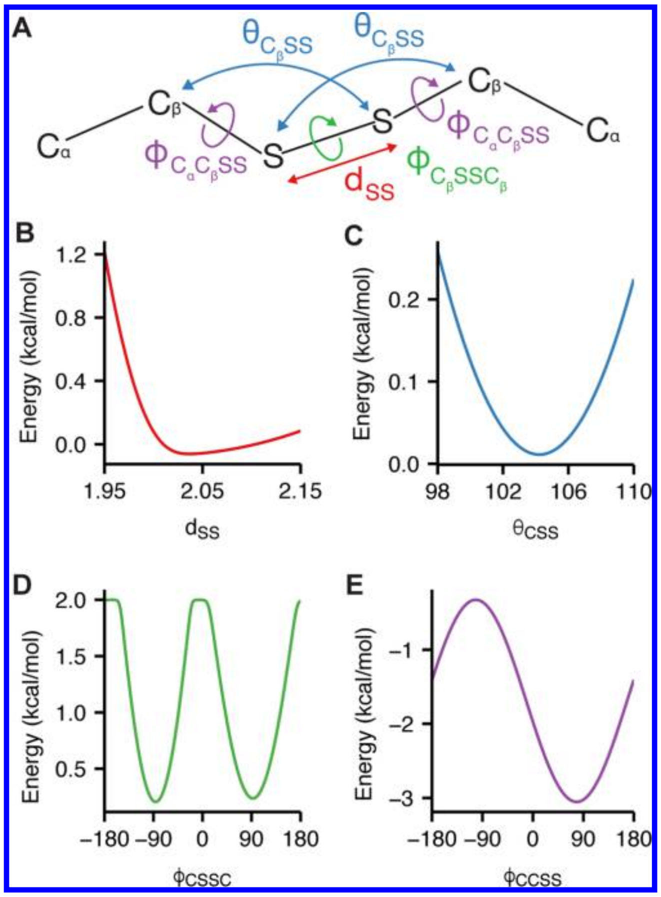


Figure 4. Orientation-dependent disulfide bonding model. (A) Degrees of freedom evaluated by the disulfide bonding energy: the sulfur−sulfur distance, d_{SS}; the angle formed by C_β and the two sulfur atoms, $\theta_{C_\beta SS}$; the dihedral angle corresponding to rotation about the C_β−sulfur bond, $\phi_{C_\alpha C_\beta SS}$; and the dihedral angle corresponding to rotation about the S−S bond, $\phi_{C_\beta SSC_\beta}$. (B−E) Plots of the energy terms (B) $E^{SS}_{dslf}(d_{SS})$, (C) $E^{CSS}_{dslf}(\theta_{C_\beta SS})$, (D) $E^{C_\beta SSC_\beta}_{dslf}(\phi_{C_\beta SSC_\beta})$, and (E) $E^{C_\alpha C_\beta SS}_{dslf}(\phi_{C_\alpha C_\beta SS})$.

$$E_{\text{dslf\_fa13}} = \sum_{S_1,S_2} E^{SS}_{\text{dslf}}(d_{SS}) + E^{CSS}_{\text{dslf}}(\theta_{C_{\beta1}SS}) + E^{CSS}_{\text{dslf}}(\theta_{C_{\beta2}SS}) + E^{C_\alpha C_\beta SS}_{\text{dslf}}(\phi_{C_{\alpha1}C_{\beta1}SS}) + E^{C_\alpha C_\beta SS}_{\text{dslf}}(\phi_{C_{\alpha2}C_{\beta2}SS}) + E^{C_\beta SSC_\beta}_{\text{dslf}}(\phi_{C_{\beta1}SSC_{\beta2}}) \tag{23}$$

Terms for Protein Backbone and Side-Chain Torsions. Rosetta evaluates backbone and side-chain conformations in torsion space to greatly reduce the search domain and increase the computational efficiency. Traditional molecular mechanics force fields describe torsional energies in terms of sines and cosines, which have at times performed poorly at reproducing the observed backbone dihedral distributions in unstructured regions.[81] Instead, Rosetta uses several knowledge-based terms for torsion angles that are fast approximations of quantum effects and more accurately model the preferred conformations of protein backbones and side chains.

Ramachandran Maps. To evaluate the backbone ϕ and ψ angles, we define an energy term called `rama_prepro` based on Ramachandran maps for each amino acid using torsions from 3985 protein chains with a resolution of ≤1.8 Å, *R* factor of ≤0.22 and sequence identity of ≤50%.[82] Amino acids with low electron density (in the bottom 25th percentile of each residue type) were removed from the data set. The resulting ~581 000 residues were used in adaptive kernel density estimates[52] of Ramachandran maps with a grid step of 10° for both ϕ and ψ. Residues preceding proline are also treated separately because they exhibit distinct ϕ, ψ preferences due to steric interactions with the proline C_δ.[83] The energy, called $E_{\text{rama\_prepro}}$, is then computed by converting the probabilities to energies at the grid points via the inverted Boltzmann relation[84] (eq 24 and Figure 5):

$$E_{\text{rama\_prepro}} = \sum_i \begin{cases} -\ln[P_{\text{reg}}(\phi_i, \psi_i|\text{aa}_i)] & \text{C-terminus, } i+1 \neq \text{Pro} \\ -\ln[P_{\text{prepro}}(\phi_i, \psi_i|\text{aa}_i)] & i+1 = \text{Pro} \end{cases} \tag{24}$$

The energies are then evaluated using bicubic interpolation. The Supporting Information includes a detailed discussion of why interpolation is performed on the backbone torsional

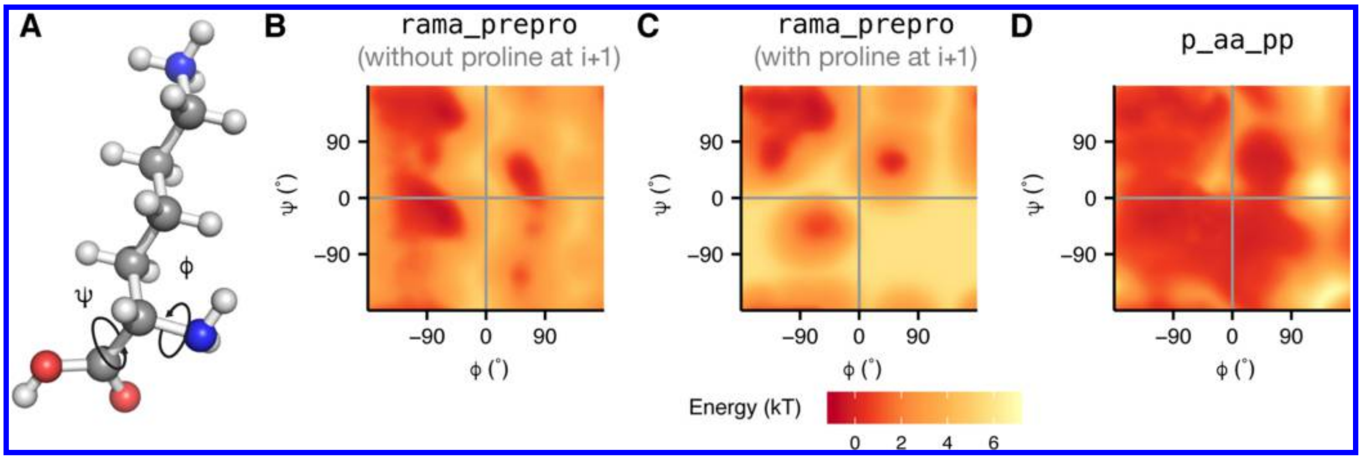


Figure 5. Backbone torsion energies. (A) The angle ϕ is defined by the backbone atoms C_{i-1}−N−C_α−C, and the angle ψ is defined by N−C_α−C−N_{i+1}. (B, C) Backbone-dependent torsion energies ($E_{\text{rama\_prepro}}$) for the lysine residue (B) without a proline at *i* + 1 and (C) with a proline at *i* + 1. (D) $E_{\text{p\_aa\_pp}}$ of lysine.

energies rather than the *probabilities* (Figure S3 and eqs S12 and S13).

Backbone Design Term. Rosetta also computes the likelihood of placing a specific amino acid side chain given an existing ϕ, ψ backbone conformation. This term, called `p_aa_pp`, represents the propensity to observe an amino acid relative to the other 19 canonical amino acids.[85] The knowledge-based propensity, $P(\text{aa}|\phi, \psi)$ (eq 25), was derived using the adaptive kernel density estimates for $P(\phi, \psi|\text{aa})$ and Bayes' rule, and the equation for $E_{\text{p\_aa\_pp}}$ is given in eq 26 (Figure 5D):

$$P(\text{aa}|\phi, \psi) = \frac{P(\phi, \psi|\text{aa})P(\text{aa})}{\sum_{\text{aa}'} P(\phi, \psi|\text{aa}')} \tag{25}$$

$$E_{\text{p\_aa\_pp}} = \sum_r -\ln\left[\frac{P(\text{aa}_r|\phi_r, \psi_r)}{P(\text{aa}_r)}\right] \tag{26}$$

Side-Chain Conformations. Protein side chains mostly occupy discrete conformations (rotamers) separated by large energy barriers. To evaluate rotamer conformations, Rosetta derives probabilities from the 2010 backbone-dependent rotamer library (http://dunbrack.fccc.edu/bbdep2010/), which contains the frequencies, means, and standard deviations of individual χ angles for each χ angle k of each rotamer of each amino acid type.[52] The probability has three components: (1) observing a specific rotamer given the backbone dihedral angles; (2) observing specific χ angles given the rotamer; and (3) observing the terminal χ angle distribution, which is either Gaussian-like or continuous when the terminal χ angle is sp^2-hybridized (eq 27):

$$P(\chi|\phi, \psi, \text{aa}) = P(\text{rot}|\phi, \psi, \text{aa})\left[\prod_{k<T} P(\chi_k|\phi, \psi, \text{rot}, \text{aa})\right] P(\chi_T|\phi, \psi, \text{rot}, \text{aa}) \tag{27}$$

where T is the number of rotameric χ angles plus 1.

The 2010 rotamer library distinguishes between rotameric and nonrotameric torsions. A torsion is rotameric when the third of the four atoms defining the torsion is sp^3-hybridized (i.e., preferring ~60°, ~180°, and ~ −60° with steep energy barriers between the wells), If the last χ torsion is rotameric, the probability $p(\chi_T|\phi, \psi, \text{rot}, \text{aa})$ is fixed at 1. On the other hand, a torsion is nonrotameric if its third atom is sp^2-hybridized: the library describes its probability distribution continuously instead. The category of semirotameric amino acids with both rotameric and nonrotameric dihedrals encompasses eight amino acids: Asp, Asn, Gln, Glu, His, Phe, Tyr, and Trp.[86]

The probability of each rotamer, $P(\text{rot}|\phi, \psi, \text{aa})$, is derived from the same data set as the Ramachandran maps described above. The probabilities were identified using adaptive kernel density estimation, and the same data set is used to estimate the mean μ_{χ_k} and standard deviation σ_{χ_k} for each χ dihedral in the rotamer as functions of the backbone dihedrals, allowing us to compute a probability for the χ values using eq 28:

$$P(\chi_k|\phi_k, \psi_k, \text{rot}) = \exp\left[-\frac{1}{2}\left(\frac{\chi_k - \mu_{\chi_k}(\phi, \psi|\text{rot}, \text{aa})}{\sigma_{\chi_k}(\phi, \psi|\text{rot}, \text{aa})}\right)^2\right] \tag{28}$$

This formulation is reminiscent of the Gaussian distribution, except that it is missing the normalization coefficient of $[2\pi\sigma_{\chi_k}(\phi, \psi|\text{rot}, \text{aa})]^{-1/2}$. Taking the logarithm of this probability gives a term resembling Hooke's law with the spring constant given by $\sigma_{\chi_k}^{-2}(\phi, \psi|\text{rot}, \text{aa})$.

The full form of $E_{\text{fa\_dun}}$ is given by eq 29 as a sum over all residues r. The difference between the rotameric and semirotameric models is also shown in Figure 6.

$$\begin{aligned} E_{\text{fa\_dun}} = {} & \sum_r -\ln[P(\text{rot}_r|\phi_r, \psi_r, \text{aa}_r)] \\ & + \sum_{k<T_r} \frac{1}{2}\left(\frac{\chi_{k,r} - \mu_{\chi_k}(\phi_r, \psi_r|\text{rot}_r, \text{aa}_r)}{\sigma_{\chi_k}(\phi_r, \psi_r|\text{rot}_r, \text{aa}_r)}\right)^2 \\ & - \ln[P(\chi_{T_r,r}|\phi_r, \psi_r, \text{rot}_r, \text{aa}_r)] \end{aligned} \tag{29}$$

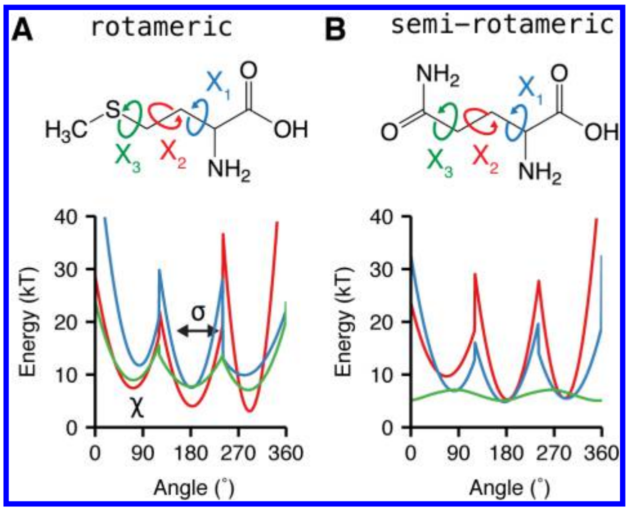


Figure 6. Energies for side-chain rotamer conformations. The Dunbrack rotamer energy, $E_{\text{fa\_dun}}$, is dependent on both the ϕ and ψ backbone torsions and the χ side-chain torsions. Here we demonstrate the variation of $E_{\text{fa\_dun}}$ when the backbone is fixed in an α-helical conformation with ϕ = −57° and ψ = −47° and the χ values can vary. χ_1 is shown in blue, χ_2 in red, and χ_3 in green. (A) χ-dependent Dunbrack energy of methionine with an sp^3-hybridized terminus. (B) χ-dependent energy of glutamine with an sp^2-hybridized χ_3 terminus. χ_1, χ_2, and χ_3 of methionine and χ_1 and χ_2 of glutamine express rotameric behavior, while χ_3 of the latter expresses broad nonrotameric behavior.

The energy from $-\ln[P(\text{rot}_r|\phi_r, \psi_r, \text{aa}_r)]$ is computed using bicubic-spline interpolation; $P(\chi_{T_r,r}|\phi_r, \psi_r, \text{rot}_r, \text{aa}_r)$ is computed using tricubic-spline interpolation. To save memory, $\mu_{\chi_k}(\phi_r, \psi_r|\text{rot}_r, \text{aa}_r)$, and $\sigma_{\chi_k}(\phi_r, \psi_r|\text{rot}_r, \text{aa}_r)$ are computed using bilinear interpolation, though this has the effect of producing derivative discontinuities at the (ϕ, ψ) grid boundaries. These discontinuities, however, do not appear to produce noticeable artifacts.[51]

Terms for Special-Case Torsions. Peptide bond dihedral angles, ω, remain mostly fixed in a cis or trans conformation and depend on the backbone ϕ and ψ angles. Since the electron pair on the backbone nitrogen donates electron density to the electrophilic carbonyl carbon, the peptide bond has partial double-bond character. To model this barrier to rotation, Rosetta implements a backbone-dependent harmonic penalty

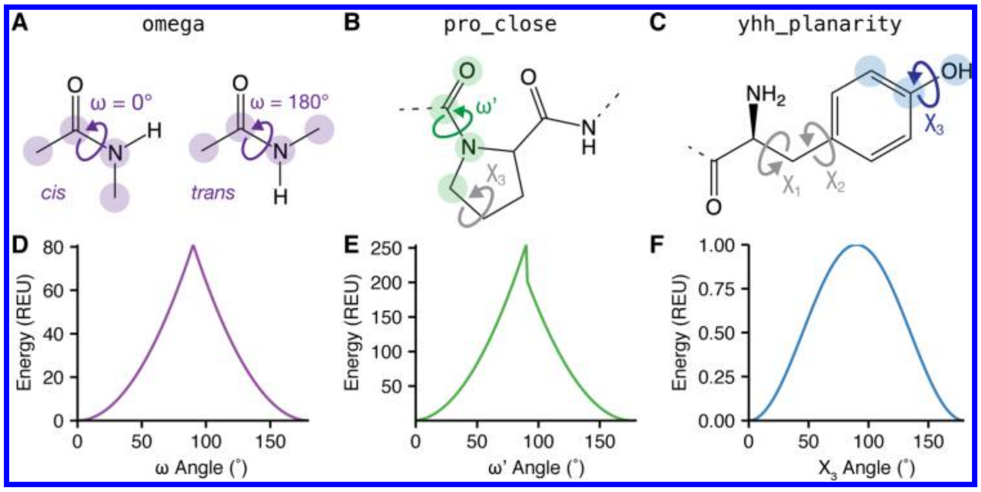


Figure 7. Special case torsion energies. (A−C) Rosetta implements three additional energy terms to model torsional degrees of freedom with acute preferences: (A) omega torsion, corresponding to rotation about C−N; (B) proline secondary omega torsion, corresponding to rotation about C−N related to the C_δ in the ring; (C) tyrosine terminal χ torsion. (D) Omega energy. (E) Proline closure energy. (F) Tyrosine planarity energy.

centered near 0° for cis and 180° for trans (Figure 7A). This energy, called `omega`, is evaluated on all peptide bonds in the biomolecule (eq 30):

$$E_{\text{omega}} = \sum_r \ln\left(\frac{1}{6\sqrt{2\pi}}\right) - \ln\left(\frac{1}{\sigma_\omega(\phi_r, \psi_r|\text{aa}_r)\sqrt{2\pi}}\right) + \frac{[\omega_r - \mu_\omega(\phi_r, \psi_r|\text{aa}_r)]^2}{2\sigma_\omega^2(\phi_r, \psi_r|\text{aa}_r)} \tag{30}$$

The means μ_ω and standard derivations σ_ω are backbone (ϕ, ψ)-dependent, as given by kernel regressions of ω on ϕ and ψ.[72]

Most Rosetta protocols search over only simple torsions within chains and rigid-body degrees of freedom between chains. However, the side chain of proline requires special treatment because its ring cannot be represented by a kinematic tree.[87] Therefore, Rosetta implements a proline closure term, called `pro_close` (Figure 7B). There are two components to this energy, as shown in eq 31. First, there is a torsional potential that operates on the dihedral formed by $O_{r-1}-C_{r-1}-N_r-C_{\delta,r}$ (called ω_r') given the observed mean $\mu_{\omega'}$ and standard deviation $\sigma_{\omega'}$, where r is the residue index. This term keeps the C_δ atom in the peptide plane. Second, to ensure the correct geometry for the two hydrogens bound to C_δ, we build a virtual nitrogen atom, N_v, off C_δ whose coordinate is controlled by χ_3 (Figure 7B). The `pro_close` term seeks to align the virtual atom N_v directly on top of the real backbone nitrogen. The $N-C_\delta-C_\gamma$ bond angle and the $N-C_\delta$ bond length are restrained to their ideal values.

$$E_{\text{pro\_close}} = \sum_{r\in\text{Pro}} \begin{cases} \dfrac{(\omega_r' - \mu_{\omega'})^2}{\sigma_{\omega'}{}^2} + \dfrac{\|N_r - N_{v,r}\|^2}{\sigma_{N,N_v}{}^2} & r \text{ is not the N-terminus} \\[2ex] \dfrac{\|N_r - N_{v,r}\|^2}{\sigma_{N,N_v}{}^2} & r \text{ is the N-terminus} \end{cases} \tag{31}$$

Tyrosine also requires special treatment for its χ_3 angle because the hydroxyl hydrogen prefers to be in the plane of the aromatic ring.[88] To enforce this preference, Rosetta implements a sinusoidal penalty to model the barrier to a χ_3 angle that deviates from planarity. This tyrosine hydroxyl penalty is called `yhh_planarity` (eq 32 and Figure 7C):

$$E_{\text{yhh\_planarity}} = \sum_i \frac{1}{2}[\cos(\pi - 2\chi_{3,i}) + 1] \tag{32}$$

Terms for Modeling Nonideal Bond Lengths and Angles. *Cartesian Bonding Energy.* Recently, modeling Cartesian degrees of freedom during gradient-based minimization has been shown to improve Rosetta's ability to refine low-resolution structures determined by X-ray crystallography and cryogenic electron microscopy[53] as well as its ability to discriminate near-native conformations in the absence of experimental data.[89] These data suggest that capturing nonideal bond lengths and angles can be important for accurate modeling of minimum-energy protein conformations. To accommodate, Rosetta now allows these "nonideal" angles and lengths to be included as additional degrees of freedom in refinement and includes a Cartesian minimization mode in which the atom coordinates are explicit degrees of freedom in optimization.

To evaluate the energetics of nonideal bond lengths, angles, and planar groups, an energy term called `cart_bonded` represents the deviation of these degrees of freedom from ideal using harmonic potentials (eqs 32−34):

$$E_{\text{cart\_length}} = \frac{1}{2}\sum_{i=1}^{n} k_{\text{length},i}(d_i - d_{i,0})^2 \tag{33}$$

$$E_{\text{cart\_angle}} = \frac{1}{2}\sum_{i=1}^{m} k_{\text{angle},i}(\theta_i - \theta_{i,0})^2 \tag{34}$$

$$E_{\text{cart\_torsion}} = \frac{1}{2}\sum_{i=1}^{l} k_{\text{torsion},i}\left[f_{\text{wrap}}\left(\phi_i - \phi_{i,0}, \frac{2\pi}{\rho_i}\right)\right]^2 \tag{35}$$

In these equations, d_i is a bonded-atom-pair distance with $d_{i,0}$ as its ideal distance, θ_i is a bond angle with $\theta_{i,0}$ as its ideal angle, and ϕ_i is a bond torsion or improper torsion with $\phi_{i,0}$ as its ideal value and ρ_i as its periodicity. The ideal bond lengths and angles[90,91] were selected on the basis of their ability to rebuild side chains observed in crystal structures (Kevin Karplus and James J. Havranek, unpublished); they were subsequently modified empirically.[51] The spring constants for the angle and length terms are from CHARMM32.[19] Finally, all planar groups and the C_β "pseudotorsion" are constrained using empirically derived values and spring constants:

The function $f_{\text{wrap}}(x, y)$ wraps x to the range $[0, y)$. To avoid double counting in the case of $E_{\text{cart\_torsion}}$, the spring constant $k_{\text{torsion},i}$ is zero when the torsion ϕ_i is being scored by either the `rama` or `fa_dun` terms.

Terms for Protein Design. *Design Reference Energy.* The terms above are sufficient for comparing different protein conformations with a fixed sequence. However, protein design simulations compare the relative stability of different amino acid sequences given a desired structure to identify models that exhibit a large free energy gap between the folded and unfolded states. Explicit calculations of unfolded-state free energies are computationally expensive and error-prone. Rosetta therefore approximates the relative energies of the unfolded-state ensembles using an unfolded-state reference energy called `ref`.

Rosetta calculates the reference energy as a sum of individual constant unfolded-state reference energies, ΔG_i^{ref}, for each amino acid, aa_i (eq 36):[1]

$$E_{\text{ref}} = \sum_i \Delta G_i^{\text{ref}}(\text{aa}_i) \tag{36}$$

The ΔG_i^{ref} values are empirically optimized by searching for values that maximize native sequence recovery (discussed below) during design simulations on a large set of high-resolution crystal structures.[50,51] During design, this energy term helps normalize the observed frequencies of the different amino acids. When design is turned off, the term contributes a constant offset for a fixed sequence.

Bringing the Energy Terms Together. The Rosetta energy function combines all of the terms using a weighted linear sum to approximate free energies (Table 1). Historically, we have adjusted the weights and parameters to balance the energetic contributions from the various terms. This balance is important because the van der Waals, solvation, and electrostatics energies partially capture torsional preferences and overlap can cause errors as a result of double counting of atomic or residue-specific contributions.[92] More recently, we have fixed the physics-based terms with weights of 1.0 and perturbed the other weights and atomic-level parameters using a Nelder–Mead scheme[93] to optimize the agreement of Rosetta calculations with small-molecule thermodynamic data and high-resolution structural features.[50] The energy function parameters have evolved over the years by optimization of the performance of multiple scientific benchmarks (Table 2).[50,51,94] These benchmarks were chosen to test the recovery of native-like structural features, ranging from individual hydrogen-bond geometries to thermodynamic properties and interface conformations. In addition, and more recently, Song et al.,[95] Conway and DiMaio,[96] and O'Meara et al.[49] have fit intraterm parameters to recover features of the experimentally determined folded conformations. An in-depth review of energy function benchmarking can be found in Leaver-Fay et al.[51] Table S3 lists the Rosetta database files containing the current full set of physical parameters for each score term.

Energy Function Units. Initially, Rosetta energies were expressed in a generic unit called the Rosetta energy unit (REU). This choice was made because some original Rosetta energy terms were not calibrated with experimental data, and the use of statistical potentials convoluted interpretation of the energy. Over time the physical meaning of Rosetta energies has been extensively debated within and outside the community, and several steps have been taken to clarify the interpretation. The most recent energy function (REF15) was parametrized on high-resolution protein structures and small-molecule thermodynamic parameters that were measured in kilocalories per mole.[50] The optimization data show a strong correlation between the experimental data and values predicted by Rosetta ($\Delta\Delta G$ upon mutation, $R = 0.994$; small-molecule ΔH_{vap}, Figure S1). As a result, Rosetta energies are now a stronger approximation of energies in units of kilocalories per mole. Therefore, as is standard practice for molecular force fields such as OPLS, CHARMM, and AMBER, we now also express energies in kilocalories per mole.

Table 2. Common Energy Function Benchmarking Methods

test	description	ref(s)
sequence recovery	percentage of the native sequence recovered after backbone redesign	1, 51
rotamer recovery	percentage of native rotamers recovered after full repacking	51
$\Delta\Delta G$ prediction	prediction of free energy changes upon mutation	97
loop modeling	prediction of loop conformations	98
high-resolution refinement	discrimination of native-like decoys upon refinement of ab initio protein models	99
docking	prediction of protein–protein, protein–peptide, or protein–ligand interfaces	44, 100–102
homology modeling	structure prediction incorporating homologous information from templates	103
thermodynamic properties	recapitulation of thermodynamic properties of protein side-chain analogues	17
recapitulation of crystal structure geometries	recapitulation of features (e.g., atom-pair distance distribution) from high-resolution protein crystal structures	50

ENERGIES IN ACTION: USING INDIVIDUAL ENERGY TERMS TO ANALYZE ROSETTA MODELS

Rosetta energy terms are mathematical models of the physics that governs protein structure, stability, and association. Therefore, the decomposed relative energies of a structure or ensemble of structures can expose important details about the biomolecular model. Now that we have presented the details of each energy term, we here demonstrate how these energies can be applied to detailed interpretations of structural models. In this section, we discuss two common structure calculations: (1) estimating the free energy change ($\Delta\Delta G$) of mutation[97] and (2) modeling the structure of a protein–protein interface.[101]

$\Delta\Delta G$ of Mutation. The first example demonstrates how Rosetta can be used to estimate and rationalize thermodynamic parameters. Here we present an example $\Delta\Delta G$ of mutation calculation for the T193V mutation in the RT-RH-derived peptide bound to HIV-1 protease (PDB entry 1kjg; Figure 8A).[104] The details of this calculation are provided in the Supporting Information.

Rosetta calculates the $\Delta\Delta G$ of the T193V mutation to be −4.95 kcal/mol, and the experimentally measured value is −1.11 kcal/mol.[104] Both the experiment and calculation reveal that T193V is stabilizing, yet these numbers alone do not reveal which specific interactions are responsible for the stabilization. To investigate, we used various analysis tools accessible in PyRosetta[105] to identify important energetic contributions to

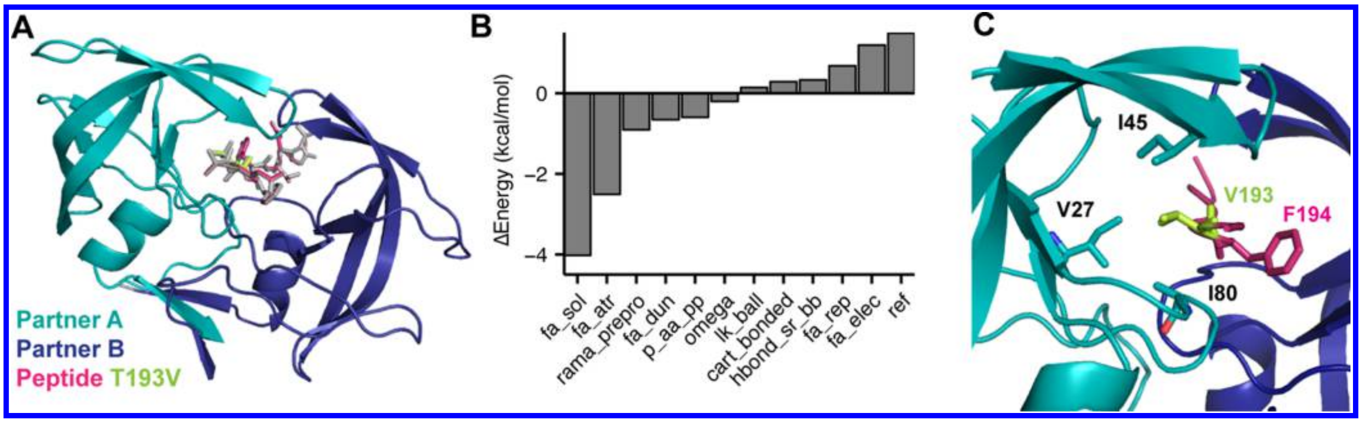


Figure 8. Structural model of the HIV-1 protease bound to the T193V mutant of RT-RH-derived peptide. (A) Structural model of the native HIV-1 peptidase (teal and dark blue) bound to the native peptide (gray) superimposed onto the T193V mutant peptide (magenta). (B) Contributions greater than ±0.1 kcal/mol to the $\Delta\Delta G$ of mutation for T193V. The remaining contributions are `dslf_fa13` = 0 kcal/mol, `hbond_lr_bb` = −0.09 kcal/mol, `hbond_bb_sc` = −0.05, `hbond_sc` = −0.0104, `fa_intra_rep` = 0.01, `fa_intra_sol` = −0.07, and `yhh_planarity` = 0. (C) Hydrophobic patch of residues surrounding position 193 on the RT-RH-derived peptide.

the total $\Delta\Delta G$. First, we decomposed the $\Delta\Delta G$ into individual energy terms and observed the balance of terms, both favorable and unfavorable, that sum to the total (Figure 8B). To decompose the most favorable term, Δ`fa_sol`, we used the `print_residue_pair_energies` function to identify residues that interact with the mutation site (in this case, residue 4) to produce a nonzero residue-pair solvation energy. With the resulting table, we found that a hydrophobic pocket around the mutation site formed by residues V27, I45, G46, and I80 on HIV peptidase and residue F194 on the peptide made a large (>0.05 kcal/mol) and favorable contribution to the change in solvation energy (Figure 8C).

We further investigated this result on the atomic level with the function `print_atom_pair_energy_table` by generating atom-pair energy tables (see the Supporting Information) for residues 5, 27, 45, 46, and 80 against both threonine and valine at residue 193 (an example for residue 80 is shown in Table 3). Here we find that the specific substitution of the polar hydroxyl group on threonine with the nonpolar alkyl group on valine stabilizes the peptide in the hydrophobic protease pocket. This result is consistent with chemical intuition and demonstrates how breaking down the total energies can provide insight into characteristics of the mutated structures.

Table 3. Changes in Atom-Pair Energies (in kcal/mol) between I80 and T193 versus V193

	I80 atoms			
T193→V193 atoms	CB	CG1	CG2	CD1
N	0.000	0.000	0.000	0.000
CA	0.000	0.000	0.000	0.004
C	0.000	0.000	0.000	0.008
O	0.000	0.000	0.000	−0.010
CB	0.000	0.054	0.000	−0.002
OG1 → CG1	0.008	−0.054	−0.316	−0.398
CG2 → CG2′	0.000	0.000	0.001	0.020

Protein−Protein Docking. The second example shows how the Rosetta energies of an ensemble of models can be used to discriminate between models and investigate the characteristics of a protein−protein interface. Below we investigate docked models of West Nile Virus envelope protein and a neutralizing antibody (PDB entry 1ztx; Figure 9A).[106] Calculation details can be found in the Supporting Information.

To evaluate the docked models, we examine the variation of the energies as a function of the root-mean-square deviation (RMSD) between the residues at the interface in each model and the known structure. For our calculation, interface residues are residues with a C_β atom less than 8.0 Å away from the C_β of a residue in the other docking partner. The plot of energies against RMSD values is called a *funnel plot* and is intended to mimic the funnel-like energy landscape of protein folding and binding.

As in the previous example, we decompose the energies to yield information about the nature of interactions at the interface. Here we observe significant changes in the following energy terms upon interface formation relative to the unbound state: `fa_atr`, `fa_rep`, `fa_sol`, `lk_ball_wtd`, `fa_elec`, `hbond_lr_bb`, `hbond_bb_sc`, and `hbond_sc` (Figure 9B). The change in the Lennard-Jones energy upon interface formation is due to the introduction of atom−atom contacts at the interface. As more atoms come into contact near the native conformation (RMSD → 0), the favorable attractive energy (`fa_atr`) decreases whereas the unfavorable repulsive energy (`fa_rep`) increases. The change in the isotropic solvation energy (Δ`fa_sol`) is positive (unfavorable), indicating that polar residues are buried upon interface formation. Balancing the desolvation penalty, the changes in polar solvation energy (Δ`lk_ball_wtd`) and electrostatics (Δ`fa_elec`) are negative because of the formation of polar contacts at the interface. Finally, the changes in the three hydrogen-bonding energies (Δ`hbond_lr_bb`, Δ`hbond_bb_sc`, and Δ`hbond_sc`) reflect the formation of backbone−backbone, backbone−side-chain, and side-chain−side-chain hydrogen bonds at the interface.

DISCUSSION

The Rosetta energy function represents our collaboration's ongoing pursuit to model the rules in nature that govern biomolecular structure, stability, and association. This paper summarizes the latest version, which brings together fundamental physical theories, statistical-mechanical models, and observations of protein structures. This work represents almost 20 years of interdisciplinary collaboration in the Rosetta

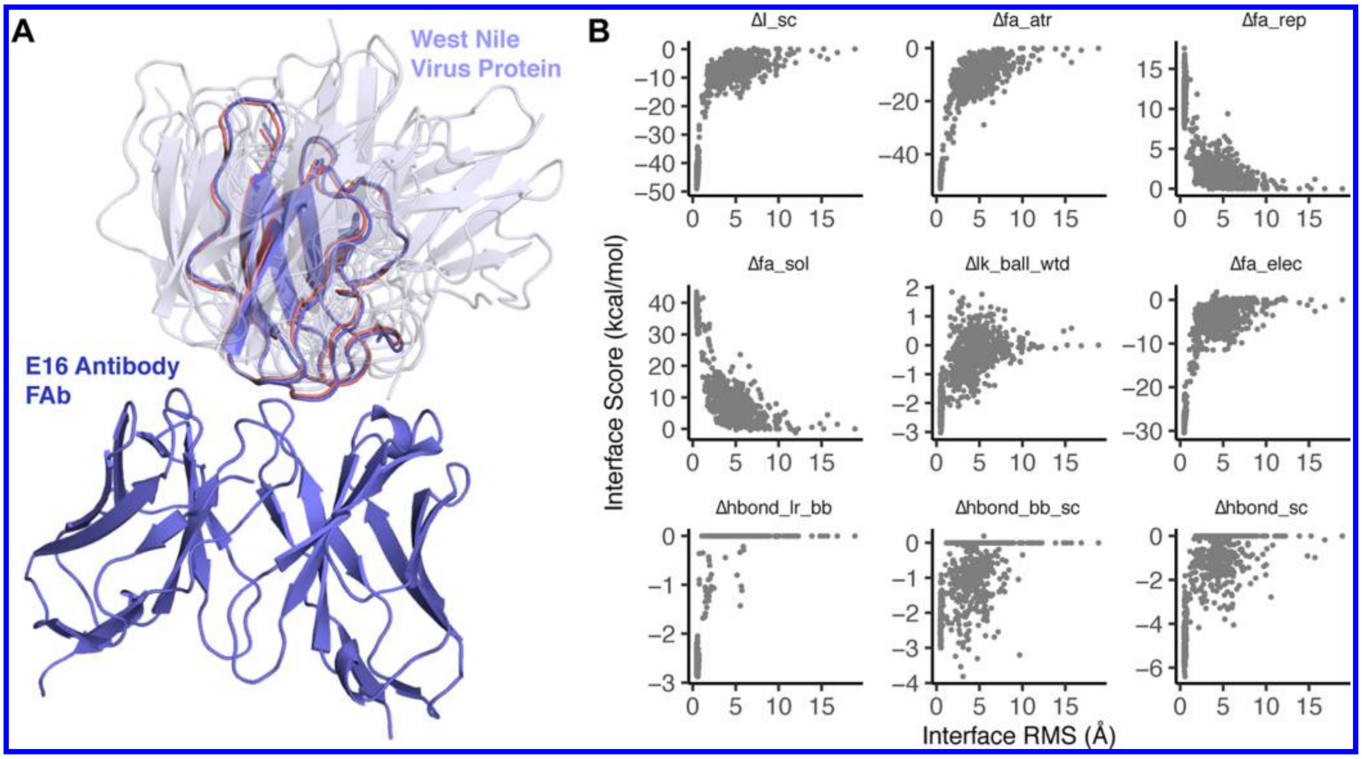


Figure 9. Using energies to discriminate docked models of West Nile Virus and the E16 neutralizing antibody. (A) Comparison of the native E16 antibody (purple) docked to the lowest-RMSD model of the West Nile Virus envelope protein and several other random models of varying energy to show sampling diversity (gray, semitransparent). (B) Change in the interface energy relative to the unbound state vs RMSD relative to the native structure. Models with low RMSD relative to the native interface have a low overall interface energy due to favorable van der Waals contacts, electrostatic interactions, and side-chain hydrogen bonds, as reflected by the Δfa_atr, Δfa_elec, and Δhbond_sc energy terms.

Table 4. New Energy Terms for Biomolecules Other than Proteins

biomolecule	term	description	unit	ref
noncanonical amino acids	`mm_lj_intra_rep`	repulsive van der Waals energy between two atoms from the same residue	kcal/mol	67
	`mm_lj_intra_atr`	attractive van der Waals energy between two atoms from the same residue	kcal/mol	67
	`mm_twist`	molecular mechanics derived torsion term for all proper torsions	kcal/mol	67
	`unfolded`	energy of the unfolded state based on explicit unfolded state model	AU^{a}	67
	`split_unfolded_1b`	one-body component of the two-component reference energy, lowest energy of a side chain in a dipeptide model system	AU	in the SI
	`split_unfolded_2b`	two-body component of the two-component reference energy, median two-body interaction energy based on atom-type composition	AU	in the SI
carbohydrates	`sugar_bb`	energy for glycosidic torsions	kcal/mol	70
DNA	`gb_elec`	generalized Born model of the electrostatics energy	kcal/mol	107
RNA	`fa_stack`	$\pi-\pi$ stacking energy for RNA bases	kT	113
	`stack_elec`	electrostatic energy for stacked RNA bases	kT	114
	`fa_elec_rna_phos`	electrostatic energy (`fa_elec`) between RNA phosphate atoms	kT	62
	`rna_torsion`	knowledge-based torsional potential for RNA	kT	62
	`rna_sugar_close`	penalty for opening an RNA sugar	kT	62

[a] AU = arbitrary units.

community, which in turn builds on and incorporates decades of work outside the community.

After 20 years, we have improved physical theories, structural data, representations, experiments, and computational tools; nevertheless, the energy functions are far from perfect. Compared with the first torsional potentials, the energy functions are also now vastly more complex. There are countless ways to arrive at more accurate energy functions. Here we discuss grand challenges specific to development of the Rosetta energy function in the coming decade.

Modeling of Biomolecules Other than Proteins. The Rosetta energy function was originally developed to predict and design protein structures. A clear artifact of this goal is the energy function's dependence on statistical potentials derived from protein X-ray crystal structures. Today the Rosetta community also pursues goals ranging from the design of synthetic macromolecules to the prediction of interactions and structures of other biomolecules such as glycoproteins and RNA. Accordingly, an active research thrust is to generalize the all-atom energy function for all biomolecules.

Many of the physically derived terms (e.g., van der Waals) have already been made compatible with noncanonical amino acids and nonprotein biomolecules (Table S5). Recently, Bhardwaj, Mulligan, Bahl, and co-workers[69] adapted the

Table 5. Energy Terms for Structure Prediction in Different Contexts

context	term	description	unit	ref(s)
membrane environment	fa_mpsolv	solvation energy dependent on the protein orientation relative to the membrane	kcal/mol	118, 121
	fa_mpenv	one-body membrane environment energy dependent on the protein orientation relative to the membrane	kcal/mol	118, 121
pH	e_pH	likelihood of side-chain protonation given a user-specified pH	kcal/mol	117

rama_prepro, p_aa_pp, fa_dun, pro_close, omega, dslf_fa13, yhh_planarity, and ref terms to be compatible with mixed-chirality peptides. Several of Rosetta's statistical potentials have been validated against quantum-mechanical calculations to evaluate non-protein models (Table 4). Early work by Meiler and Baker[44] on Rosetta Ligand introduced new atom and residue types for non-protein residues. The first non-protein energy terms were added by Havranek et al.[107] and Chen et al.,[108] who modified the hydrogen-bonding potential to capture planar hydrogen bonds between protein side chains and nucleic acid bases. Renfrew and co-workers[67,109] added molecular mechanics torsions and Lennard-Jones terms to model proteins with noncanonical amino acids, oligosaccharides, β-peptides, and oligopeptoids.[68] Labonte et al.[70] implemented Woods' CarboHydrate-Intrinsic (CHI) function,[110,111] which evaluates glycan geometries given the axial−equatorial character of the bonds. Das and co-workers added a set of terms to model Watson−Crick base pairing, $\pi-\pi$ interactions in base stacking, and torsional potentials important for predicting and designing RNA structures.[62,112−114] Bazzoli and Karanicolas[115] recently developed a new polar solvation model that evaluates the penalty associated with displacing waters in the first solvation shell. In addition, Combs[116] tested a small-molecule force field based on electron orbital models. Many of these terms are presented in detail in the Supporting Information.

Expanding Rosetta's chemical library brings new challenges. Currently there are separate energy functions for various types of biomolecules. Typically, these functions mix physically derived terms from the protein energy function with molecule-specific statistical potentials, custom weights, and possibly custom atomic parameters. If nature uses only one energy function, why do we need so many? Some discrepancies may result from features that we do not model explicitly, such as $\pi-\pi$, n−π*, and cation−π interactions. Efforts to converge on a single energy function will therefore pose interesting questions about the set of universal physical determinants of biomolecular structure.

Capturing the Intra- and Extracellular Environments. Rosetta traditionally models the solvent surrounding the protein using the Lazaridis−Karplus (LK) model, which assumes a solvent environment made of pure water. In contrast, biology operates under various conditions influenced by pH, redox potential, temperature, solvent viscosity, chaotropes, kosmotropes, and polarizability. Therefore, modeling more details of the intra- and extracellular environments would enable Rosetta to identify structures that are important in different biological contexts.

Rosetta currently includes two groups of energy terms to model alternate environments (Table 5). Kilambi and Gray[117] implemented a method to account for pH by including a term called e_pH that calculates the likelihood of a protein side chain's protonation state given a user-specified pH; it requires the inclusion of both protonated and deprotonated side chains during side-chain rotamer packing. This model can predict pK_a values with an RMS error under 1 unit,[117] and it improves protein−protein docking, especially under acidic or basic conditions.[60] The accuracy of this model is limited by the distance-dependent Coulomb approximation and sensitivity to fine backbone rearrangements.

In addition, Rosetta implements Lazaridis' implicit membrane model (IMM) for modeling proteins in a lipid bilayer enviornment.[36,118,119] The IMM terms provide a fast approximation of the nonpolar hydrocarbon core of the lipid bilayer and have been successfully applied to membrane protein folding,[120] docking, and early design tasks.[61] This continuum model has a fixed thickness, omitting the detailed chemistry at the membrane interface and any dynamic bilayer rearrangements.

The Origin of Energy Models: Top-Down versus Bottom-Up Development. Traditionally, energy functions are developed using a bottom-up approach: experimental observables serve as building blocks to parametrize physics-based formulas. The advent of powerful optimization techniques and artificial intelligence has recently empowered the top-down category, where numerical methods are used to derive models and/or parameters. Top-down approaches have been used to solve problems in various fields, including structural biology and bioinformatics. Recently, top-down development was also applied to optimize the Lennard-Jones, LK, and Coulomb parameters in the Rosetta energy function (Tables S4−S6).[50,93]

Top-down approaches have enormous potential to improve the accuracy of biomolecular modeling because more parameters can vary and the objective function can be minimized with more benchmarks. These approaches also introduce new challenges. With any computer-derived model there is a risk of overfitting, as validation via structure-prediction data sets reflect observable states, whereas simulations are intended to predict features of states that experiments cannot yet observe. Computer-derived parameters also introduce a unique kind of uncertainty. Consider the following scenario: the performance of scientific benchmarks improves as physical atomic parameters are perturbed away from the measured experimental values. As there is less physical basis for the parameters, are the predictions and interpretations still meaningful?

Top-down development will also provide power to develop more complicated energy functions. Currently, the Rosetta energy function advances by incrementally addressing weaknesses: with each new paper, we modify analytic formulas, add corrective terms, and adjust weights. As this paper demonstrates, the energy function is significantly more complicated than the initial theoretical forms. Given this complexity increase, an interesting approach to leverage the power of top-down development would be to simplify and subtract terms to evaluate their individual benefits.

A Highly Interdisciplinary Endeavor. The Rosetta energy function has advanced rapidly because of the Rosetta Community, a highly interdisciplinary collaboration among scientists with diverse backgrounds located in over 50 laboratories around the world. The many facets of our team enable us to probe different aspects of the energy function. For example, expert computer scientists and applied mathematicians have implemented algorithms to speed up calculations. Dedicated software engineers maintain the code and maintain a platform for scientific benchmark testing. Physicists and chemists develop new energy terms that better model the physical rules found in nature. Structural biologists maintain a focus on created biological features and functions. We look forward to leveraging this powerful interdisciplinary scientific team as we head into the next decade of energy function advances.

CONCLUSION: A LIVING ENERGY FUNCTION

For the first time since 2004,[48] we have documented all of the mathematical and physical details of the Rosetta all-atom energy function, highlighting the latest upgrades to both the underlying science and the speed of calculations. In addition, we have illustrated how the energies can be used to analyze output models from Rosetta simulations. These advances have enabled Rosetta's achievements in biomolecular structure prediction and design over the past 15 years. Still, the energy function is far from complete and will continue to evolve long after this publication. Thus, we hope that this document will serve as an important resource for understanding the foundational physical and mathematical concepts in the energy function. Furthermore, we hope to encourage both current and future Rosetta developers and users to understand the strengths and shortcomings of the energy function as it applies to the scientific questions they are trying to answer.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b00125.

Description of changes to the Rosetta energy function since 2000; data describing the calibration of Rosetta energies in kcal/mol; additional details of energy terms and details on smoothing of statistical terms; energy terms for D-amino acids, noncanonical amino acids, carbohydrates, and nucleic acids; and methods describing example energy calculations (PDF)

Protocol capture within an interactive Python notebook demonstrating the usage of the `print_atom_pair_energy_table` function (ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: jgray@jhu.edu.

ORCID

Rebecca F. Alford: 0000-0003-0306-8476
Hahnbeom Park: 0000-0002-7129-1912
Michael S. Pacella: 0000-0001-8919-147X
Jeffrey J. Gray: 0000-0001-6380-2324

Author Contributions

R.F.A., J.R.J., A.L.-F., T.K., B.K., and J.J.G. wrote the manuscript. R.F.A., J.R.J., M.S.P., and J.J.G. generated analysis scripts and examples. A.L.-F., M.J.O., F.P.D., H.P., M.V.S., P.B., R.L.D., T.K., D.B., B.K., and J.J.G. wrote, verified, and/or contributed figures on protein energy terms. P.D.R., K.K., V.K.M., J.W.L., R.B., and R.D. wrote, verified, and/or contributed figures on non-protein energy terms.

Funding

R.F.A. was funded by a Hertz Foundation Fellowship and an NSF Graduate Research Fellowship. J.R.J. and J.J.G. were funded by NIH GM-078221. A.L.-F., J.J.G., and B.K. were funded by NIH GM-73141. M.J.O. was funded by NSF GM-114961. P.D.R. and R.B. were funded by the Simons Foundation. M.V.S. and R.L.D. were funded by NIH GM-084453 and NIH GM-111819. M.S.P. was funded by NSF BMAT 1507736. J.W.L. was funded by NIH F32-CA189246. K.K. was funded by an NSF Graduate Research Fellowship and an SGF Galiban Fellowship. D.B., H.P., and V.K.M. were funded by NIH GM-092802. T.K. was funded by NIH GM-110089 and GM-117189.

Notes

The authors declare the following competing financial interest(s): Drs. Gray, Baker, Bonneau, Kuhlman, Kortemme, and Bradley are unpaid members of the Executive Board of the Rosetta Commons. Under institutional participation agreements between the University of Washington, acting on behalf of the Rosetta Commons, and each institution participating in this article, each institution may be entitled to a portion of revenue received on licensing of software described here.

ACKNOWLEDGMENTS

Development of the Rosetta energy function would not be possible without the entire Rosetta Commons collaboration: a community of scientists, engineers, and software developers that have worked together for almost 20 years. We estimate that hundreds of scientists from the 50 institutions in the Rosetta Commons have made minor and major contributions to the advancement of the all-atom energy function. When writing this paper, it was impossible to compile a complete list of energy function contributors. Instead, our author list reflects a small subset of the historical contributors who provided text, figures, and expertise needed to write a comprehensive, complete, and accurate description of the current energy function formulation. We also want to give special recognition to the early pioneers of the Rosetta energy function who are not coauthors: Carol Rohl, Kim Simons, Charlie Strauss, Ingo Ruczinski, William Sheffler, Jens Meiler, Ora Schuler-Furman, James Havranek, and Ian Davis. We also acknowledge individuals that contributed to assembling the manuscript. We thank Sergey Lyskov for development of the benchmark server that enables continuous and transparent energy function testing and Morgan Nance, Henry Lessen, and Rocco Moretti for helpful comments on the manuscript.

REFERENCES

(1) Kuhlman, B.; Baker, D. Native Protein Sequences Are close to Optimal for Their Structures. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (19), 10383−10388.

(2) Richardson, J. S. The Anatomy and Taxonomy of Protein Structure. *Adv. Protein Chem.* **1981**, *34*, 167−339.

(3) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K. W.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y.-E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. Rosetta3: An Object-Oriented Software Suite for the

Simulation and Design of Macromolecules. *Methods Enzymol.* **2011**, *487*, 545−574.

(4) Anfinsen, C. B. Principles That Govern the Folding of Protein Chains. *Science* **1973**, *181* (4096), 223−230.

(5) Jones, J. E. On the Determination of Molecular Fields.—I. From the Variation of the Viscosity of a Gas with Temperature. *Proc. R. Soc. London, Ser. A* **1924**, *106*, 441−462.

(6) Jones, J. E. On the Determination of Molecular Fields.—II. From the Equation of State of a Gas. *Proc. R. Soc. London, Ser. A* **1924**, *106*, 463−477.

(7) Levitt, M.; Lifson, S. Refinement of Protein Conformations Using a Macromolecular Energy Minimization Procedure. *J. Mol. Biol.* **1969**, *46* (2), 269−279.

(8) Urey, H. C.; Bradley, C. A. The Vibrations of Pentatonic Tetrahedral Molecules. *Phys. Rev.* **1931**, *38* (11), 1969−1978.

(9) Westheimer, F. Calculation of the Magnitude of Steric Effects. In *Steric Effects in Organic Chemistry*; Newman, M. S., Ed.; Wiley: New York, 1956; pp 523−555.

(10) Lifson, S.; Warshel, A. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and N-Alkane Molecules. *J. Chem. Phys.* **1968**, *49* (11), 5116−5129.

(11) Warshel, A.; Lifson, S. Consistent Force Field Calculations. II. Crystal Structures, Sublimation Energies, Molecular and Lattice Vibrations, Molecular Conformations, and Enthalpies of Alkanes. *J. Chem. Phys.* **1970**, *53* (2), 582−594.

(12) Levitt, M. Energy Refinement of Hen Egg-White Lysozyme. *J. Mol. Biol.* **1974**, *82* (3), 393−420.

(13) Gelin, B. R.; Karplus, M. Sidechain Torsional Potentials and Motion of Amino Acids in Porteins: Bovine Pancreatic Trypsin Inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* **1975**, *72* (6), 2002−2006.

(14) Levinthal, C.; Wodak, S. J.; Kahn, P.; Dadivanian, A. K. Hemoglobin Interaction in Sickle Cell Fibers. I: Theoretical Approaches to the Molecular Contacts. *Proc. Natl. Acad. Sci. U. S. A.* **1975**, *72* (4), 1330−1334.

(15) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1996**, *118* (9), 2309−2309.

(16) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94* (26), 8897−8909.

(17) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657−1666.

(18) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4* (2), 187−217.

(19) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545−1614.

(20) Sun, H. COMPASS: An Ab Initio Force-Field Optimized for Condensed-Phase Applications Overview with Details on Alkane and Benzene Compounds. *J. Phys. Chem. B* **1998**, *102* (38), 7338−7364.

(21) Tanaka, S.; Scheraga, H. A. Model of Protein Folding: Inclusion of Short-, Medium-, and Long-Range Interactions. *Proc. Natl. Acad. Sci. U. S. A.* **1975**, *72* (10), 3802−3806.

(22) Tanaka, S.; Scheraga, H. A. Model of Protein Folding: Incorporation of a One-Dimensional Short-Range (Ising) Model into a Three-Dimensional Model. *Proc. Natl. Acad. Sci. U. S. A.* **1977**, *74* (4), 1320−1323.

(23) Miyazawa, S.; Jernigan, R. L. Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.* **1996**, *256* (3), 623−644.

(24) Wilmanns, M.; Eisenberg, D. Three-Dimensional Profiles from Residue-Pair Preferences: Identification of Sequences with Beta/alpha-Barrel Fold. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90* (4), 1379−1383.

(25) Jones, D. T.; Taylor, W. R.; Thornton, J. M. A New Approach to Protein Fold Recognition. *Nature* **1992**, *358* (6381), 86−89.

(26) Bowie, J. U.; Lüthy, R.; Eisenberg, D. A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science* **1991**, *253* (5016), 164−170.

(27) Sippl, M. J. Calculation of Conformational Ensembles from Potentials of Mean Force. An Approach to the Knowledge-Based Prediction of Local Structures in Globular Proteins. *J. Mol. Biol.* **1990**, *213* (4), 859−883.

(28) Skolnick, J.; Kolinski, A. Simulations of the Folding of a Globular Protein. *Science* **1990**, *250* (4984), 1121−1125.

(29) Bashford, D.; Case, D. A. Generalized Born Models of Macromolecular Solvation Effects. *Annu. Rev. Phys. Chem.* **2000**, *51* (1), 129−152.

(30) Warshel, A.; Kato, M.; Pisliakov, A. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **2007**, *3* (6), 2034−2045.

(31) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences Using Simulated Annealing and Bayesian Scoring Functions. *J. Mol. Biol.* **1997**, *268* (1), 209−225.

(32) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.

(33) Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B. A.; Bystroff, C.; Baker, D. Improved Recognition of Native-like Protein Structures Using a Combination of Sequence-Dependent and Sequence-Independent Features of Proteins. *Proteins: Struct., Funct., Genet.* **1999**, *34* (1), 82−95.

(34) Kuhlman, B.; Baker, D. Native Protein Sequences Are close to Optimal for Their Structures. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (19), 10383−10388.

(35) Neria, E.; Fischer, S.; Karplus, M. Simulation of Activation Free Energies in Molecular Systems. *J. Chem. Phys.* **1996**, *105* (5), 1902−1921.

(36) Lazaridis, T.; Karplus, M. Effective Energy Function for Proteins in Solution. *Proteins: Struct., Funct., Genet.* **1999**, *35* (2), 133−152.

(37) Dunbrack, R. L., Jr.; Cohen, F. E. Bayesian Statistical Analysis of Protein Side-Chain Rotamer Preferences. *Protein Sci.* **1997**, *6* (8), 1661−1681.

(38) Kortemme, T.; Morozov, A. V.; Baker, D. An Orientation-Dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes. *J. Mol. Biol.* **2003**, *326* (4), 1239−1259.

(39) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. Close Agreement between the Orientation Dependence of Hydrogen Bonds Observed in Protein Structures and Quantum Mechanical Calculations. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (18), 6946−6951.

(40) Bradley, P.; Misura, K. M. S.; Baker, D. Toward High-Resolution de Novo Structure Prediction for Small Proteins. *Science* **2005**, *309* (5742), 1868−1871.

(41) Kortemme, T.; Baker, D. A Simple Physical Model for Binding Energy Hot Spots in Protein-Protein Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (22), 14116−14121.

(42) Kortemme, T.; Kim, D. E.; Baker, D. Computational Alanine Scanning of Protein−Protein Interfaces. *Sci. STKE* **2004**, *2004* (219), pl2.

(43) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. Protein-Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations. *J. Mol. Biol.* **2003**, *331* (1), 281−299.

(44) Meiler, J.; Baker, D. ROSETTALIGAND: Protein-Small Molecule Docking with Full Side-Chain Flexibility. *Proteins: Struct., Funct., Genet.* **2006**, *65* (3), 538−548.

(45) Kortemme, T.; Joachimiak, L. A.; Bullock, A. N.; Schuler, A. D.; Stoddard, B. L.; Baker, D. Computational Redesign of Protein-Protein Interaction Specificity. *Nat. Struct. Mol. Biol.* **2004**, *11* (4), 371−379.

(46) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **2003**, *302* (5649), 1364−1368.

(47) Chevalier, B. S.; Kortemme, T.; Chadsey, M. S.; Baker, D.; Monnat, R. J.; Stoddard, B. L. Design, Activity, and Structure of a Highly Specific Artificial Endonuclease. *Mol. Cell* **2002**, *10* (4), 895−905.

(48) Rohl, C. A.; Strauss, C. E. M.; Misura, K. M. S.; Baker, D. Protein Structure Prediction Using Rosetta. *Methods Enzymol.* **2004**, *383*, 66−93.

(49) O'Meara, M. J.; Leaver-Fay, A.; Tyka, M. D.; Stein, A.; Houlihan, K.; DiMaio, F.; Bradley, P.; Kortemme, T.; Baker, D.; Snoeyink, J.; Kuhlman, B. Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **2015**, *11* (2), 609−622.

(50) Park, H.; Bradley, P.; Greisen, P., Jr.; Liu, Y.; Mulligan, V. K.; Kim, D. E.; Baker, D.; DiMaio, F. Simultaneous Optimization of Biomolecular Energy Function on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **2016**, *12* (12), 6201−6212.

(51) Leaver-Fay, A.; O'Meara, M. J.; Tyka, M.; Jacak, R.; Song, Y.; Kellogg, E. H.; Thompson, J.; Davis, I. W.; Pache, R. A.; Lyskov, S.; Gray, J. J.; Kortemme, T.; Richardson, J. S.; Havranek, J. J.; Snoeyink, J.; Baker, D.; Kuhlman, B. Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. *Methods Enzymol.* **2013**, *523*, 109−143.

(52) Shapovalov, M. V.; Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011**, *19* (6), 844−858.

(53) DiMaio, F.; Song, Y.; Li, X.; Brunner, M. J.; Xu, C.; Conticello, V.; Egelman, E.; Marlovits, T. C.; Cheng, Y.; Baker, D. Atomic-Accuracy Models from 4.5-Å Cryo-Electron Microscopy Data with Density-Guided Iterative Local Refinement. *Nat. Methods* **2015**, *12* (4), 361−365.

(54) Vortmeier, G.; DeLuca, S. H.; Els-Heindl, S.; Chollet, C.; Scheidt, H. A.; Beck-Sickinger, A. G.; Meiler, J.; Huster, D. Integrating Solid-State NMR and Computational Modeling to Investigate the Structure and Dynamics of Membrane-Associated Ghrelin. *PLoS One* **2015**, *10* (3), e0122444.

(55) Correia, B. E.; Bates, J. T.; Loomis, R. J.; Baneyx, G.; Carrico, C.; Jardine, J. G.; Rupert, P.; Correnti, C.; Kalyuzhniy, O.; Vittal, V.; Connell, M. J.; Stevens, E.; Schroeter, A.; Chen, M.; Macpherson, S.; Serra, A. M.; Adachi, Y.; Holmes, M. A.; Li, Y.; Klevit, R. E.; Graham, B. S.; Wyatt, R. T.; Baker, D.; Strong, R. K.; Crowe, J. E.; Johnson, P. R.; Schief, W. R. Proof of Principle for Epitope-Focused Vaccine Design. *Nature* **2014**, *507* (7491), 201−206.

(56) Masica, D. L.; Schrier, S. B.; Specht, E. A.; Gray, J. J. De Novo Design of Peptide−Calcite Biomineralization Systems. *J. Am. Chem. Soc.* **2010**, *132* (35), 12252−12262.

(57) King, N. P.; Bale, J. B.; Sheffler, W.; McNamara, D. E.; Gonen, S.; Gonen, T.; Yeates, T. O.; Baker, D. Accurate Design of Co-Assembling Multi-Component Protein Nanomaterials. *Nature* **2014**, *510* (7503), 103−108.

(58) Siegel, J. B.; Zanghellini, A.; Lovick, H. M.; Kiss, G.; Lambert, A. R.; St. Clair, J. L.; Gallaher, J. L.; Hilvert, D.; Gelb, M. H.; Stoddard, B. L.; Houk, K. N.; Michael, F. E.; Baker, D. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels−Alder Reaction. *Science* **2010**, *329* (5989), 309−313.

(59) Wolf, C.; Siegel, J. B.; Tinberg, C.; Camarca, A.; Gianfrani, C.; Paski, S.; Guan, R.; Montelione, G.; Baker, D.; Pultz, I. S. Engineering of Kuma030: A Gliadin Peptidase That Rapidly Degrades Immunogenic Gliadin Peptides in Gastric Conditions. *J. Am. Chem. Soc.* **2015**, *137* (40), 13106−13113.

(60) Kilambi, K. P.; Reddy, K.; Gray, J. J. Protein-Protein Docking with Dynamic Residue Protonation States. *PLoS Comput. Biol.* **2014**, *10* (12), e1004018.

(61) Alford, R. F.; Koehler Leman, J.; Weitzner, B. D.; Duran, A. M.; Tilley, D. C.; Elazar, A.; Gray, J. J. An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Comput. Biol.* **2015**, *11* (9), e1004398.

(62) Das, R.; Karanicolas, J.; Baker, D. Atomic Accuracy in Predicting and Designing Noncanonical RNA Structure. *Nat. Methods* **2010**, *7* (4), 291−294.

(63) Thyme, S. B.; Baker, D.; Bradley, P. Improved Modeling of Side-Chain−Base Interactions and Plasticity in Protein−DNA Interface Design. *J. Mol. Biol.* **2012**, *419* (3−4), 255−274.

(64) Joyce, A. P.; Zhang, C.; Bradley, P.; Havranek, J. J. Structure-Based Modeling of Protein: DNA Specificity. *Briefings Funct. Genomics* **2015**, *14* (1), 39−49.

(65) Lemmon, G.; Meiler, J. Rosetta Ligand Docking with Flexible XML Protocols. *Methods Mol. Biol.* **2012**, *819*, 143−155.

(66) Combs, S. A.; DeLuca, S. L.; DeLuca, S. H.; Lemmon, G. H.; Nannemann, D. P.; Nguyen, E. D.; Willis, J. R.; Sheehan, J. H.; Meiler, J. Small-Molecule Ligand Docking into Comparative Models with Rosetta. *Nat. Protoc.* **2013**, *8* (7), 1277−1298.

(67) Renfrew, P. D.; Choi, E. J.; Bonneau, R.; Kuhlman, B. Incorporation of Noncanonical Amino Acids into Rosetta and Use in Computational Protein-Peptide Interface Design. *PLoS One* **2012**, *7* (3), e32637.

(68) Drew, K.; Renfrew, P. D.; Craven, T. W.; Butterfoss, G. L.; Chou, F.-C.; Lyskov, S.; Bullock, B. N.; Watkins, A.; Labonte, J. W.; Pacella, M.; Kilambi, K. P.; Leaver-Fay, A.; Kuhlman, B.; Gray, J. J.; Bradley, P.; Kirshenbaum, K.; Arora, P. S.; Das, R.; Bonneau, R. Adding Diverse Noncanonical Backbones to Rosetta: Enabling Peptidomimetic Design. *PLoS One* **2013**, *8* (7), e67051.

(69) Bhardwaj, G.; Mulligan, V. K.; Bahl, C. D.; Gilmore, J. M.; Harvey, P. J.; Cheneval, O.; Buchko, G. W.; Pulavarti, S. V. S. R. K.; Kaas, Q.; Eletsky, A.; Huang, P.-S.; Johnsen, W. A.; Greisen, P. J.; Rocklin, G. J.; Song, Y.; Linsky, T. W.; Watkins, A.; Rettie, S. A.; Xu, X.; Carter, L. P.; Bonneau, R.; Olson, J. M.; Coutsias, E.; Correnti, C. E.; Szyperski, T.; Craik, D. J.; Baker, D. Accurate de Novo Design of Hyperstable Constrained Peptides. *Nature* **2016**, *538* (7625), 329−335.

(70) Labonte, J. W.; Adolf-Bryfogle, J.; Schief, W. R.; Gray, J. J. Residue-Centric Modeling and Design of Saccharide and Glycoconjugate Structures. *J. Comput. Chem.* **2017**, *38* (5), 276−287.

(71) Yanover, C.; Bradley, P. Extensive Protein and DNA Backbone Sampling Improves Structure-Based Specificity Prediction for C2H2 Zinc Fingers. *Nucleic Acids Res.* **2011**, *39* (11), 4564−4576.

(72) Berkholz, D. S.; Driggers, C. M.; Shapovalov, M. V.; Dunbrack, R. L., Jr.; Karplus, P. A. Nonplanar Peptide Bonds in Proteins Are Common and Conserved but Not Biased toward Active Sites. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (2), 449−453.

(73) Khatib, F.; Cooper, S.; Tyka, M. D.; Xu, K.; Makedon, I.; Popovic, Z.; Baker, D.; Players, F. Algorithm Discovery by Protein Folding Game Players. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (47), 18949−18953.

(74) Grigoryan, G.; Ochoa, A.; Keating, A. E. Computing van Der Waals Energies in the Context of the Rotamer Approximation. *Proteins: Struct., Funct., Genet.* **2007**, *68* (4), 863−878.

(75) Dahiyat, B. I.; Mayo, S. L. Probing the Role of Packing Specificity in Protein Design. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94* (19), 10172−10177.

(76) Warshel, A.; Russell, S. T. Calculations of Electrostatic Interactions in Biological Systems and in Solutions. *Q. Rev. Biophys.* **1984**, *17* (3), 283−422.

(77) Hubbard, R. E.; Kamran Haider, M. Hydrogen Bonds in Proteins: Role and Strength. In *Encyclopedia of Life Sciences*; John Wiley & Sons: Chichester, U.K., 2010.

(78) Li, X.-Z.; Walker, B.; Michaelides, A. Quantum Nature of the Hydrogen Bond. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (16), 6369−6373.

(79) Richardson, J. S.; Keedy, D. A.; Richardson, D. C. "The Plot" Thickens: More Data, More Dimensions, More Uses. In *Biomolecular Forms and Functions: A Celebration of 50 Years of the Ramachandran Map*; Bansal, M., Srinivasan, N., Eds.; World Scientific: Singapore, 2013; pp 46−61.

(80) Wang, C.; Bradley, P.; Baker, D. Protein−Protein Docking with Backbone Flexibility. *J. Mol. Biol.* **2007**, *373* (2), 503−519.

(81) Ho, B. K.; Thomas, A.; Brasseur, R. Revisiting the Ramachandran Plot: Hard-Sphere Repulsion, Electrostatics, and H-Bonding in the Alpha-Helix. *Protein Sci.* **2003**, *12* (11), 2508−2522.

(82) Wang, G.; Dunbrack, R. L. PISCES: A Protein Sequence Culling Server. *Bioinformatics* **2003**, *19* (12), 1589−1591.

(83) Ting, D.; Wang, G.; Shapovalov, M.; Mitra, R.; Jordan, M. I.; Dunbrack, R. L. Neighbor-Dependent Ramachandran Probability Distributions of Amino Acids Developed from a Hierarchical Dirichlet Process Model. *PLoS Comput. Biol.* **2010**, *6* (4), e1000763.

(84) Finkelstein, A. V.; Badretdinov, A. Y.; Gutin, A. M. Why Do Protein Architectures Have Boltzmann-like Statistics? *Proteins: Struct., Funct., Genet.* **1995**, *23* (2), 142−150.

(85) Shortle, D. Propensities, Probabilities, and the Boltzmann Hypothesis. *Protein Sci.* **2003**, *12* (6), 1298−1302.

(86) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The Penultimate Rotamer Library. *Proteins: Struct., Funct., Genet.* **2000**, *40* (3), 389−408.

(87) MacArthur, M. W.; Thornton, J. M. Influence of Proline Residues on Protein Conformation. *J. Mol. Biol.* **1991**, *218* (2), 397−412.

(88) McDonald, I. K.; Thornton, J. M. Satisfying Hydrogen Bonding Potential in Proteins. *J. Mol. Biol.* **1994**, *238* (5), 777−793.

(89) Conway, P.; Tyka, M. D.; DiMaio, F.; Konerding, D. E.; Baker, D. Relaxation of Backbone Bond Geometry Improves Protein Energy Landscape Modeling. *Protein Sci.* **2014**, *23* (1), 47−55.

(90) Hall, M. B. *Insight II*, version 12000; Accelrys, Inc.: San Diego, CA, 2005.

(91) Engh, R. A.; Huber, R. IUCr. Accurate Bond and Angle Parameters for X-Ray Protein Structure Refinement. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991**, *47* (4), 392−400.

(92) Renfrew, P. D.; Butterfoss, G. L.; Kuhlman, B. Using Quantum Mechanics to Improve Estimates of Amino Acid Side Chain Rotamer Energies. *Proteins: Struct., Funct., Genet.* **2008**, *71* (4), 1637−1646.

(93) Barton, R. R.; Ivey, J. S. Nelder−Mead Simplex Modifications for Simulation Optimization. *Manage. Sci.* **1996**, *42* (7), 954−973.

(94) Ó Conchúir, S.; Barlow, K. A.; Pache, R. A.; Ollikainen, N.; Kundert, K.; O'Meara, M. J.; Smith, C. A.; Kortemme, T. A. Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design. *PLoS One* **2015**, *10* (9), e0130433.

(95) Song, Y.; Tyka, M.; Leaver-Fay, A.; Thompson, J.; Baker, D. Structure-Guided Forcefield Optimization. *Proteins: Struct., Funct., Genet.* **2011**, *79* (6), 1898−1909.

(96) Conway, P.; DiMaio, F. Improving Hybrid Statistical and Physical Forcefields through Local Structure Enumeration. *Protein Sci.* **2016**, *25* (8), 1525−1534.

(97) Kellogg, E. H.; Leaver-Fay, A.; Baker, D. Role of Conformational Sampling in Computing Mutation-Induced Changes in Protein Structure and Stability. *Proteins: Struct., Funct., Genet.* **2011**, *79* (3), 830−838.

(98) Mandell, D. J.; Coutsias, E. A.; Kortemme, T. Sub-Angstrom Accuracy in Protein Loop Reconstruction by Robotics-Inspired Conformational Sampling. *Nat. Methods* **2009**, *6* (8), 551−552.

(99) Tyka, M. D.; Keedy, D. A.; André, I.; DiMaio, F.; Song, Y.; Richardson, D. C.; Richardson, J. S.; Baker, D. Alternate States of Proteins Revealed by Detailed Energy Landscape Mapping. *J. Mol. Biol.* **2011**, *405* (2), 607−618.

(100) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. Protein-Protein Docking Benchmark Version 4.0. *Proteins: Struct., Funct., Genet.* **2010**, *78* (15), 3111−3114.

(101) Chaudhury, S.; Berrondo, M.; Weitzner, B. D.; Muthu, P.; Bergman, H.; Gray, J. J. Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2. *PLoS One* **2011**, *6* (8), e22477.

(102) Raveh, B.; London, N.; Zimmerman, L.; Schueler-Furman, O. Rosetta FlexPepDock Ab-Initio: Simultaneous Folding, Docking and Refinement of Peptides onto Their Receptors. *PLoS One* **2011**, *6* (4), e18934.

(103) Song, Y.; DiMaio, F.; Wang, R. Y.-R.; Kim, D.; Miles, C.; Brunette, T.; Thompson, J.; Baker, D. High-Resolution Comparative Modeling with RosettaCM. *Structure* **2013**, *21* (10), 1735−1742.

(104) Altman, M. D.; Nalivaika, E. A.; Prabu-Jeyabalan, M.; Schiffer, C. A.; Tidor, B. Computational Design and Experimental Study of Tighter Binding Peptides to an Inactivated Mutant of HIV-1 Protease. *Proteins: Struct., Funct., Genet.* **2008**, *70* (3), 678−694.

(105) Chaudhury, S.; Lyskov, S.; Gray, J. J. PyRosetta: A Script-Based Interface for Implementing Molecular Modeling Algorithms Using Rosetta. *Bioinformatics* **2010**, *26* (5), 689−691.

(106) Nybakken, G. E.; Oliphant, T.; Johnson, S.; Burke, S.; Diamond, M. S.; Fremont, D. H. Structural Basis of West Nile Virus Neutralization by a Therapeutic Antibody. *Nature* **2005**, *437* (7059), 764−769.

(107) Havranek, J. J.; Duarte, C. M.; Baker, D. A Simple Physical Model for the Prediction and Design of Protein−DNA Interactions. *J. Mol. Biol.* **2004**, *344* (1), 59−70.

(108) Chen, Y.; Kortemme, T.; Robertson, T.; Baker, D.; Varani, G. A New Hydrogen-Bonding Potential for the Design of Protein-RNA Interactions Predicts Specific Contacts and Discriminates Decoys. *Nucleic Acids Res.* **2004**, *32* (17), 5147−5162.

(109) Renfrew, P. D.; Craven, T. W.; Butterfoss, G. L.; Kirshenbaum, K.; Bonneau, R. A Rotamer Library to Enable Modeling and Design of Peptoid Foldamers. *J. Am. Chem. Soc.* **2014**, *136* (24), 8772−8782.

(110) Nivedha, A. K.; Thieker, D. F.; Makeneni, S.; Hu, H.; Woods, R. J. Vina-Carb: Improving Glycosidic Angles during Carbohydrate Docking. *J. Chem. Theory Comput.* **2016**, *12* (2), 892−901.

(111) Nivedha, A. K.; Makeneni, S.; Foley, B. L.; Tessier, M. B.; Woods, R. J. Importance of Ligand Conformational Energies in Carbohydrate Docking: Sorting the Wheat from the Chaff. *J. Comput. Chem.* **2014**, *35* (7), 526−539.

(112) Das, R.; Baker, D. Automated de Novo Prediction of Native-like RNA Tertiary Structures. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (37), 14664−14669.

(113) Sripakdeevong, P.; Kladwang, W.; Das, R. An Enumerative Stepwise Ansatz Enables Atomic-Accuracy RNA Loop Modeling. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (51), 20573−20578.

(114) Chou, F.-C.; Kladwang, W.; Kappel, K.; Das, R. Blind Tests of RNA Nearest-Neighbor Energy Prediction. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (30), 8430−8435.

(115) Bazzoli, A.; Karanicolas, J. "Solvent Hydrogen-Bond Occlusion": A New Model of Polar Desolvation for Biomolecular Energetics. *J. Comput. Chem.* **2017**, *38*, 1321−1331.

(116) Combs, S. A. Identification and Scoring of Partial Covalent Interactions in Proteins and Protein Ligand Complexes. Ph.D. Dissertation, Vanderbilt University, Nashville, TN, 2013.

(117) Kilambi, K. P.; Gray, J. J. Rapid Calculation of Protein pKa Values Using Rosetta. *Biophys. J.* **2012**, *103* (3), 587−595.

(118) Barth, P.; Schonbrun, J.; Baker, D. Toward High-Resolution Prediction and Design of Transmembrane Helical Protein Structures. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (40), 15682−15687.

(119) Yarov-Yarovoy, V.; Schonbrun, J.; Baker, D. Multipass Membrane Protein Structure Prediction Using Rosetta. *Proteins: Struct., Funct., Genet.* **2006**, *62* (4), 1010−1025.

(120) Wang, Y.; Barth, P. Evolutionary-Guided de Novo Structure Prediction of Self-Associated Transmembrane Helical Proteins with near-Atomic Accuracy. *Nat. Commun.* **2015**, *6*, 7196.

(121) Lazaridis, T. Effective Energy Function for Proteins in Lipid Membranes. *Proteins: Struct., Funct., Genet.* **2003**, *52* (2), 176−192.

Supporting Information

“The Rosetta all-atom energy function for macromolecular modeling and design”

Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O’Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Phillip Bradley, Roland L. Dunbrack Jr., Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, Jeffrey J. Gray

Major changes to the Rosetta energy function since 2000

The all-atom Rosetta energy function for proteins has undergone significant upgrades since the original implementation in 2000. These changes range from improved atomic parameters and models of hydrogen bonding to smoothing routines that eliminate errors during minimization. An overview of these advances is listed in **Table S1**.

Table S1: Major changes to the Rosetta Energy Function since 2000

Energy Term	Adjustment	Ref.
Lennard-Jones	Soften repulsive potential Atomic radii matched to crystal structures Shifted LJ Potential Extra soft repulsive potential Make derivatives continuous New well-depth parameters Incorporation of hydrogens in the `fa_atr` calculation	Kuhlman *et al.* 2000[1] Kuhlman *et al.* 2003[2] Tsai *et al.* 2003[3] Meiler & Baker, 2006[4] Scheffler 2006, Unpublished Park et al. 2016[5]
Solvation	Implementation of Lazaridis-Karplus Model Make derivatives continuous Anisotropic Solvation Model New atomic volume λ and ΔG^{free} parameters	Kuhlman *et al.* 2000[6] Sheffler 2006, Unpublished Yanover *et al.* 2011[7] Park *et al.* 2016[5]
Electrostatics	Knowledge-based Pair term Coulomb electrostatics for ligand interactions Coulomb electrostatics for nucleic acids Coulomb electrostatics for proteins Sigmoid dielectric model Avoidance of dipole splitting for local interactions New partial charges	Kuhlman *et al.* 2000[6] Meiler & Baker, 2006[4] Yanover *et al.* 2011[7] O'Meara *et al.* 2015[8] Park *et al.* 2016[5]
Hydrogen Bonding	Orientation-dependent hydrogen bond potential Favoring H-bonds in the sp^2 plane No H-bond environment dependence Weights on hydrogen bond donors and acceptors	Kortemme *et al.* 2003[9] O'Meara *et al.* 2015[8] Park *et al.* 2016[5]
Dunbrack Rotamers	Add 2002 backbone-dependent rotamer library Replace 2002 version with the 2010 smoothed rotamer library	Dunbrack *et al.* 2002[10] Shapovalov *et al.* 2011[11] Leaver-Fay *et al.* 2013[12]
Ramachandran & p_aa_pp	Interpolation with bicubic splines Correction for pre-proline backbone torsions	Leaver-Fay *et al.* 2013[12] Park *et al.* 2016[5]
Side-chain specific	Penalty for Tyr hydroxyl hydrogen leaving aromatic plane	O'Meara *et al.* 2015[8]
Design reference energy	Refit reference energies with OptE Refit reference energies with DualOptE	Leaver-Fay *et al.* 2011[12] O'Meara *et al.* 2015[8] Park *et al.* 2016[5]

Table S2: Rosetta revisions corresponding to major energy function updates

Version	Rosetta Revision[a]	Public Version[b]
Score12	pre-#55611	Pre-Rosetta 3.5
Talaris2013	#55611	Rosetta 3.5
Talaris2014	#58602	v2016.13-dev58602
REF2015	#59248	v2017.05-dev59248

[a] Internal code revision number available to member institutions of the Rosetta Commons.
[b] Download the public Rosetta release from http://www.rosettacommons.org.

Data for calibrating Rosetta energies to kcal/mol

The parameters and weights in *REF2015* were recently fit[5] such that Rosetta simulations reproduce high-resolution protein structural data and thermodynamic data for small molecules from Jorgensen *et al.*[13] Thus, the Rosetta energy is now expressed in kcal/mol. In support, **Figure S1** compares experimental data and Rosetta predictions of density, heat of vaporization (ΔH_{vap}) and heat capacity ($C_{p(l)}$) for seventeen molecules: ethane, propane, isobutene, cyclohexane, benzene, toluene, phenol, methanol, ethanol, 2-propanol, tert-Butyl alcohol, methane thiol, ethane thiol, dimethyl sulfide, acetamide, *N*-methylamide, *N*-methylformamide, dimethyl ether, ethanol and propanone.

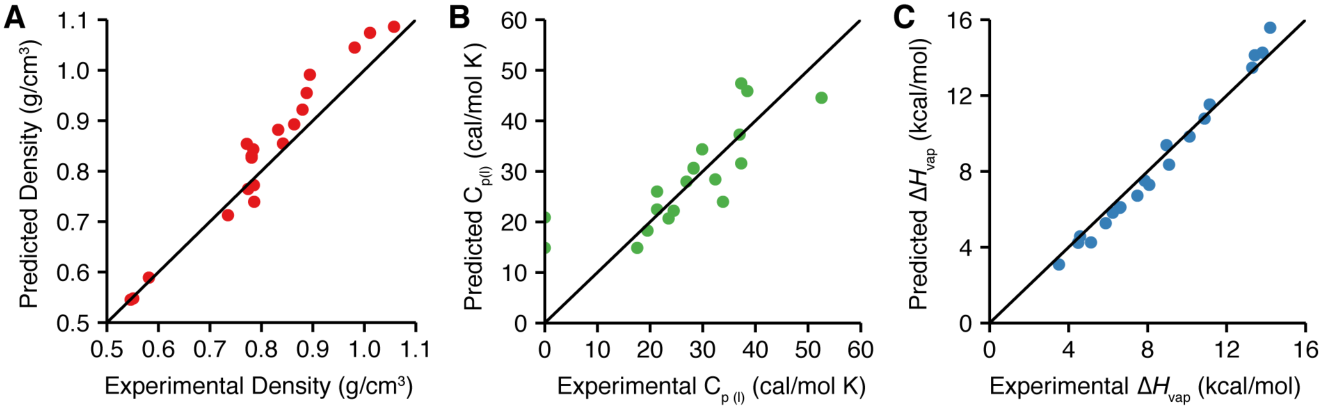


Figure S1: Comparison of Rosetta simulations with experimental thermodynamic data
Comparison between Rosetta predictions and experimental thermodynamic measurements for seventeen small molecules (A) Density (B) Heat Capacity and (C) Heat of vaporization.

Additional energy function details

Parameters for the Lennard-Jones and Lazaridis-Karplus energies

New experiments and numerical methods to optimize the energy function have led to updated atomic-parameters used by the Lennard-Jones[14,15] and Lazaridis-Karplus[16] potentials. The updated parameters are in the following Rosetta database files:

Table S3: Location of Rosetta atom type parameters

Parameter	**Database File**
Radii, $\Delta G^{\text{free}}, \epsilon$	`chemical/atom_type_sets/fa_standard/atom_properties.txt`
Partial charges	`chemical/residue_type_sets/residue_types/l-caa/*.params`
`lk_ball` weights	`chemical/atom_type_sets/fa_standard/extras/lk_ball_wtd.txt`
`hbond` parameters	`scoring/score_functions/hbonds/*`
`rama`	`scoring/score_functions/rama/*`
`p_aa_pp`	`scoring/score_functions/P_AA_pp/*`
`omega`	`scoring/score_functions/omega/*`
`fa_dun`	`rotamer/*`

Tables S4-S6 present a comparison between selected atomic parameters between the original source publication and the values in Rosetta energy functions, *Talaris2014* and *REF2015*.

Table S4: Atomic radii values from the *Neria* et al. force field, *Talaris2014*, and *REF2015*

Atom Type	Neria *et al.*[17] Radius (Å)	*Talaris2014* Radius (Å)	*REF2015*[5] Radius (Å)
CAbb	2.3650	2.0000	2.0112
CH1	2.3650	2.0000	2.0112
CH2	2.2350	2.0000	2.0112
CH3	2.1650	2.0000	2.0112
CNH2	--	2.0000	1.9922
COO	2.1000	2.0000	1.9649
CObb	--	2.0000	1.9649
aroC	2.1000	2.0000	1.9859
NH2O	--	1.7500	1.7632
Narg	1.6000	1.7500	1.7632
Nbb	1.6000	1.7500	1.7632
Nhis	1.6000	1.7500	1.7632
Nlys	1.6000	1.7500	1.7632
Npro	1.6000	1.7500	1.7632
Ntrp	1.6000	1.7500	1.7632
OCbb	1.6000	1.5500	1.5268
OH	1.6000	1.5500	1.5354
ONH2	--	1.5500	1.5760
OOC	--	1.5500	1.4492
S	0.0430	1.9000	2.0171
HNbb	--	1.0000	0.8773
Hapo	--	1.2000	1.4634
Haro	--	1.2000	1.3778
Hpol	0.8000	1.0000	0.8773

Table S5: Well-depth parameters from the *Neria* et al. force field, *Talaris2014*, and *REF2015*

Atom Type	Neria *et al.*[17] ϵ (kcal/mol)	*Talaris2014* ϵ (kcal/mol)	*REF2015*[5] ϵ (kcal/mol)
CAbb	0.0486	0.0486	0.0626
CH1	0.0486	0.0486	0.0626
CH2	0.1142	0.1142	0.0626
CH3	0.1811	0.1811	0.0626
CNH2	--	0.1200	0.0626
COO	0.1200	0.1200	0.0946
CObb	--	0.1400	0.1418
aroC	0.1200	0.1200	0.1418
NH2O	--	0.2834	0.0688
Narg	0.2384	0.2834	0.1617
Nbb	0.2384	0.2834	0.1617
Nhis	0.2384	0.2834	0.1617
Nlys	0.2384	0.2834	0.1617
Npro	0.2384	0.2834	0.1617
Ntrp	0.2384	0.2834	0.1617
OCbb		0.1591	0.1617
OH	0.1591	0.1591	0.1617
ONH2	--	0.1591	0.1424
OOC	--	0.2100	0.1619
S	0.0430	0.1600	0.1829
SH1	--	--	0.0999
HNbb	--	0.0500	0.4560
HS	--	--	0.4560
Hapo	--	0.0500	0.0050
Haro	--	0.0500	0.0508
Hpol	--	0.0500	0.0218

Table S6: ΔG^{free} parameters from Lazaridis & Karplus, *Talaris2014* and *REF2015*

Atom Type	Lazarids-Karplus[16] ΔG^{free} (kcal/mol)	*Talaris2014*, ΔG^{free} (kcal/mol)	*REF2015*[5] ΔG^{free} (kcal/mol)
CAbb	-0.2500	1.0000	2.5338
CH0	-0.2500	-0.2500	1.4093
CH1	-0.2500	-0.2500	-3.5384
CH2	0.5200	0.5200	-1.8547
CH3	1.5000	1.5000	7.2929
CNH2	--	0.0000	3.0770
COO	0.1200	-1.4000	-3.3326
CObb	--	1.0000	3.1042
aroC	0.8000	0.0800	1.7979
NH2O	--	-7.8000	-8.1016
Narg	-10.0000	-10.0000	-8.9684
Nbb	-7.8000	-5.0000	-9.9695
Nhis	-4.0000	-4.0000	-9.7396
Nlys	-20.000	-20.000	-20.865
Npro	-1.5500	-1.5500	-0.9846
Ntrp	-8.9000	-8.9000	-8.4131
OCbb	-10.0000	-5.0000	-8.0068
OH	-6.7000	-6.7000	-8.1335
ONH2	--	-5.8500	-6.5916
OOC	--	-10.0000	-9.2398
S	-4.1000	-4.1000	-1.7072
SH1	-2.7000	--	3.2916

Analytical form of the hydrogen bonding potential

To avoid expensive table lookups, the hydrogen bonding potential (**Eq. 21-22** in the main text) is given by component energies with simple analytical forms. For completeness, we detail these analytical forms below. The first two components, $E_{\text{hbond}}^{HA}(d_{HA})$ and $E_{\text{hbond}}^{HAD}(\theta_{HAD})$ are polynomial functions, $f_{\text{poly}}(P,x)$ where the polynomial P depends on the atom type of the acceptor and donor, and the order n varies from 6 to 10 (**Eq. S1**). The forms of $E_{\text{hbond}}^{HA}(d_{HA})$ and $E_{\text{hbond}}^{HAD}(\theta_{HAD})$ are given by **Eq. S2-3**.

$$f_{\text{poly}}(P,x) = C_0 + C_1 x + C_2 x^2 + \cdots + C_{n-1} x^{n-1} + C_n x^n \qquad \text{(S1)}$$

$$E_{\text{hbond}}^{HA}(d_{HA}) = F_{HA} \cdot f_{\text{poly}}(P, d_{HA}) \qquad \text{(S2)}$$

$$E_{\text{hbond}}^{HAD}(\theta_{HAD}) = F_{HAD} \cdot f_{\text{poly}}(P(\theta_{HAD})) \qquad \text{(S3)}$$

The third component, $E_{\text{hbond}}^{B_2BAH}(\rho, \phi_{B_2BAH}, \theta_{BAH})$, is dependent on the hybridization of the acceptor, ρ. For sp^2 hybridized acceptors, the potential is given as a combination of cosine and polynomial functions (**Eq. S4-5**) controlled by a cosine switch function (**Eq. S6**). The functional forms are also shown in **Fig. S2**.

$$F(\phi) = \begin{cases} \frac{d}{2}\cos(3(\pi - \phi)) + \frac{d-1}{2} & \frac{2\pi}{3} < \phi \\ \frac{m}{2}\cos\left(\frac{1}{l}\left(\pi - \frac{2\pi}{3}\phi\right)\right) + \frac{m-1}{2} & \frac{2\pi}{3} - l \le \phi \le \frac{2\pi}{3} \\ m - \frac{1}{2} & \phi < \frac{2\pi}{3} - l \end{cases} \qquad \text{(S4)}$$

$$G(\phi) = \begin{cases} d - \frac{1}{2} & \frac{2\pi}{3} < \phi \\ \frac{m-d}{2}\cos\left(\pi - \frac{1}{l}\left(\pi - \frac{2\pi}{3}\,\phi\right)\right) + \frac{m+d+1}{2} & \frac{2\pi}{3} - l \le \phi \le \frac{2\pi}{3} \\ m - \frac{1}{2} & \phi < \frac{2\pi}{3} - l \end{cases} \qquad \text{(S5)}$$

$$H(\phi) = \frac{\cos(2\phi)+1}{2} \qquad \text{(S6)}$$

For sp^3 hybridized acceptors, the potential is modeled as a composition of sine and cosine functions. If the acceptor is attached to a ring, the potential is modeled with a simple cosine function. The overall energy is given in **Eq. S7**.

$$E_{\text{hbond}}^{B_2BAH}(\rho, \phi_{B_2BAH}, \theta_{BAH}) = \begin{cases} H(\phi_{B_2BAH})F(\phi_{B_2BAH}) + \left(1 - H(\phi_{B_2BAH})G(\phi_{B_2BAH})\right) & \rho \sim \text{sp}^2 \\ f_{\text{poly}}(\cos(\theta_{BAH})) + \frac{1}{4}\left(1 + \cos(\phi_{B_2BAH})\right) & \rho \sim \text{sp}^3 \\ f_{\text{poly}}(\cos(\theta_{BAH})) & \rho \sim \text{ring} \end{cases} \qquad \text{(S7)}$$

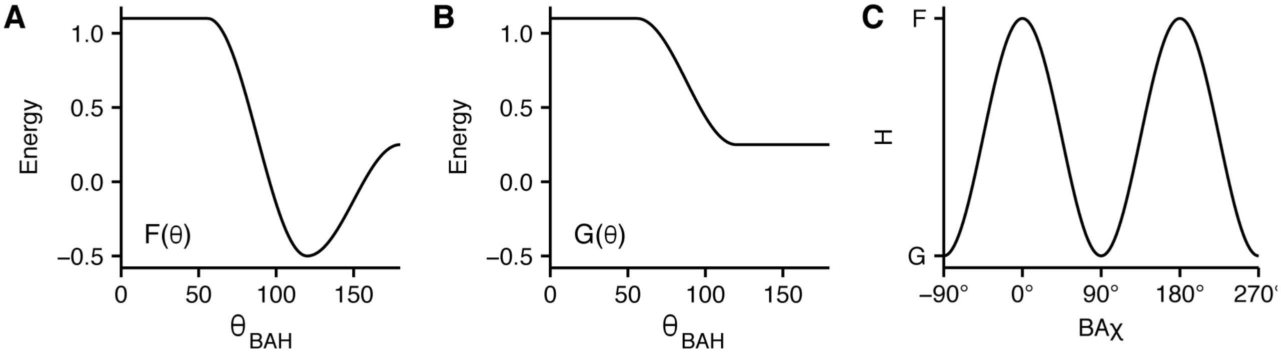


Figure S2: Analytic form of the hydrogen bonding BAχ potential for *sp2* hybridized acceptors
(A) Plot of the function $F(\theta)$ that models the energy of the BAH angle for an in-plane acceptor (B) Plot of the function $G(\theta)$ that models the energy of the BAH angle for an out-of-plane acceptor. (C) Switch function $H(\theta)$ that controls contributions from $F(\theta)$ and $G(\theta)$ at a specified value of the BAχ torsion, ϕ.

Analytical form of the disulfide bonding potential

Like the hydrogen bonding potential, the component disulfide bonding energies are defined by analytical forms. As defined by **Eq. 23** in the main text, the disulfide is computed given four component energies. First, the sulfur-sulfur distance energy $E_{\text{dslf}}^{SS}(d_{SS})$ is defined by **Eq. S8** given the sulfur-sulfur distance, d, mean distance $\overline{d_{ss}}$, standard deviation σ_{ss}, and fitting parameters α_d^{dslf}, ϵ_m and w_{SS}.

$$E_{\text{dslf}}^{SS}(d_{SS}) = w_{SS}\left(\left(\frac{d-\overline{d_{ss}}}{\sigma_{ss}}\right)^2 + \ln\left[\operatorname{erf}\left(\alpha_d^{\text{dslf}}\left(\frac{d-\overline{d_{ss}}}{\sigma_{\text{dslf}}^{SS}}\right)\right) + \epsilon_m\right]\right) \qquad \text{(S8)}$$

Next, the energy of the angle formed by a C_β and two sulfur atoms $E_{\text{dslf}}^{CSS}(\theta_{CSS})$ is defined by **Eq. S9** given the angle θ and von Mises parameters A_{CSS}, w_{CSS}, κ_{CSS}, and μ_{CSS}.

$$E_{\text{dslf}}^{C_\beta SS}(\theta_{CSS}) = w_{CSS}\left(-\ln(A_{CSS}) - \kappa_{CSS}\cos(\theta - \mu_{CSS})\right) \qquad \text{(S9)}$$

The energy of the torsion formed by $C_{\beta 1}$, $C_{\beta 2}$ and the two sulfur atoms $E_{\text{dslf}}^{C_\beta SSC_\beta}(\phi_{CSSC})$ is defined by **Eq. S10** given the torsion angle ϕ and the von Mises parameters $A_{k,C_\beta SSC_\beta}$, $\kappa_{k,C_\beta SSC_\beta}$, $\mu_{k,C_\beta SSC_\beta}$ and ϵ_{m}.

$$E_{\text{dslf}}^{C_\beta SSC_\beta}(\phi_{CSSC}) = w_{C_\beta SSC_\beta}\ln\left(\sum_{k\leq 2}\exp\left(A_{k,C_\beta SSC_\beta} + \kappa_{k,C_\beta SSC_\beta}\cos\left(\phi - \mu_{k,C_\beta SSC_\beta}\right)\right) + \ \epsilon_{\text{m}}\right) \quad \text{(S10)}$$

Finally, the energy of the torsion formed by C_α, C_β and the two adjacent sulfur atoms $E_{\text{dslf}}^{C_\alpha C_\beta S,S}(\theta_{CCSS})$ is defined by **Eq. S11** given the torsion angle ϕ and the von Mises fitting parameters $A_{k,C_\alpha C_\beta SS}$, $\kappa_{k,C_\alpha C_\beta SS}$, $\mu_{k,C_\alpha C_\beta SS}$ and ϵ_{m}.

$$E_{\text{dslf}}^{C_\alpha C_\beta S,S}(\theta_{CCSS}) = -w_{C_\alpha C_\beta SS}\ln\left(\sum_{i\leq 3}\exp\left(A_{k,C_\alpha C_\beta SS} + \kappa_{k,C_\alpha C_\beta SS}\cos\left(\theta - \mu_{k,C_\alpha C_\beta SS}\right)\right) + \ \epsilon_{\text{m}}\right) \quad \text{(S11)}$$

Statistical potentials: Interpolation of energies rather than probabilities

The Rosetta energy function uses probabilities from the Dunbrack backbone-dependent rotamer library[18] to derive torsional energies E using the inverted Boltzmann relation the probability P (**Eq. S12**):

$$E = -kT \ln P \quad \text{(S12)}$$

Prior to 2012, the probabilities for the ϕ, ψ-dependent terms were stored on a 10° x 10° grid used for energy calculations. These probabilities were calculated using bilinear interpolation and then converted to energies using **Eq. S12** and the derivatives were calculated by linearly interpolating $1/P$ and dP/dx to compute $d(-\log P)/dx = -(1/P)\, dP/dx$ with $x = \phi$ or ψ. This method resulted in large inaccuracies because P can vary by orders of magnitude over very short ranges of ϕ and ψ. In addition, the linearly interpolated derivatives are constant between grid points, so that gradient-based minimization results in moving structures to the nearest grid point where the derivative changes sign. Therefore, it is more accurate to provide P and $E = -\ln P$ at each grid point and then interpolate the energies using bicubic interpolation.

Here we demonstrate why interpolating energies is better than interpolating the probabilities. **Figure S3** compares the different interpolation strategies for a toy problem: a one-dimensional probability distribution with a discrete rotamer modeled with the following von Mises function (**Eq. S13**). Here, the location constant $\mu = 180^{\circ}$, the concentration constant $\kappa = 20$, $x = \phi$ or ψ and $I_0(\kappa)$ is the modified Bessel function of order zero needed to normalize the distribution.

$$P(x) = \frac{\exp(\kappa - \cos(x - \mu))}{2\pi I_o(\kappa)} \qquad \text{(S13)}$$

First, **Figure S3A** shows the probability distribution, P and its linear interpolation based on the 10° x 10° grid. Here, the difference between the curves demonstrate the effect of approximating P by linear interpolation. This effect would be more severe for steeper functions such as the Ramachandran probability density function. **Figure S3B** compares the $E = -\ln P$ calculation with two approaches to interpolating the function: interpolate P and then compute the energies versus compute the energies at the grid points and then interpolate. The second scenario clearly mitigates several errors which can be further improved using cubic rather than linear interpolation. Like the first panel, the benefits of cubic interpolation are clearer with steeper functions.

Figures S3C and **S3D** demonstrate that the effects of interpolating energies are more pronounced for the derivatives of P and E respectively. Previously, Rosetta computed the derivative of P as $dP/dx = [P(x+10) - P(x)]/10$ (**Fig. S3C**). The linear interpolation of this derivative includes noticeable artifacts. **Figure S3D** shows the four energy derivative curves: (1) the exact analytical expression $dE/dx = -(1/P)\, dP/dx$ where P is interpolated and dP/dx is the step function shown in **Fig. S3C** (2) $dE/dx = -(1/P)\, dP/dx$ where P and dP/dx are interpolated from the grid values, (3) $dE/dx = -(1/P)\, dP/dx$ where both P and dP/dx are interpolated from the grid values and (4) calculation of E and dE/dx on the grid followed by interpolation of dE/dx in between the grid points. The linear interpolation of dE/dx provides the closest match to the analytical expression. The current Rosetta energy function interpolates energies rather than probabilities: both P and E are stored in database files, dE/dx is calculated on the grid points, and then P, E and dE/dx are computed by bicubic spline interpolation.

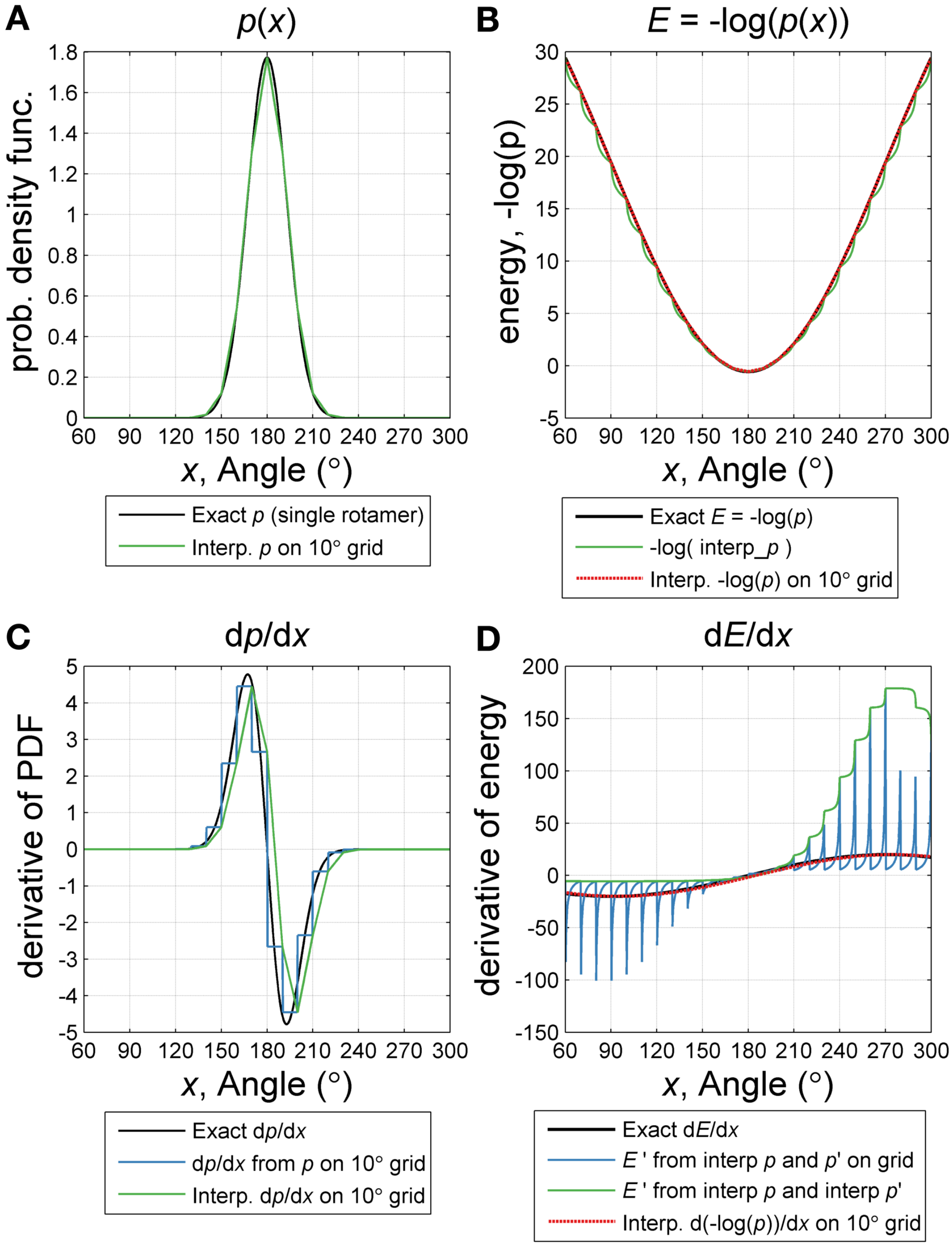


Figure S3. Approximating the energy and energy derivatives for torsional potentials

Comparison between the old and new approach of approximating the energy and energy derivatives using a toy example in one dimension. (A) Exact analytical expression of the probability distribution P(X) (black) compared to approximation of the grid (green). (B) Exact energy expression, -log p(x) (black) compared to interpolated probabilities (green) and interpolation on the grid (red). (C) Probability first-order derivatives: analytical expression (black), derivative approximation with no interpolation (blue), and derivative with linear interpolation (green). (D) Energy derivatives: exact (black), calculation as a step function (blue), calculation by linear interpolation (green), calculation from grid values (red).

Methods for energy-based analysis examples

ΔΔG of Mutation. The coordinate file for 1kgj was downloaded from the Protein Data Bank[19] and cleaned to remove any non-canonical amino acids. The PDB was refined with fast relax constrained to native coordinates using Cartesian-space refinement and the *REF2015* energy function using the following command line:

```
relax.linuxgccrelease –s 1kgj.pdb –use_input_sc \
–constrain_relax_to_start_coords –ignore_unrecognized_res –nstruct 1000 \
–relax:coord_constrain_to_sidechains –relax:ram_constraints false \
–relax:Cartesian –relax:min_type lbfgs_armijo_nonmonotone
```

After refinement, the lowest scoring model was used to generate five structures of the native conformation and five structures of the T193V mutated conformation using a Cartesian version of Rosetta's ddg protocol.[20]

```
cartesian_ddg.linuxgccrelease –s 1kgj_refined_lowest.pdb –ddg:mut_file \
$MUT_FILE –ddg:iterations 5 –optimization:default-max_cycles 200 –bbnbr 1 \
–relax:min_type lbfgs_armijo_nonmonotone –fa_max_dis 9.0
```

The energies were averaged for each ensemble of five structures. The Δ*ΔG* was then calculated as the difference between the average energy of the mutated ensemble and the average energy of the native ensemble.

To determine which specific interactions underlie the observed differences in solvation, we first needed to identify which residue-pair interactions contribute most to the change in solvation energy. Because the mutation is taking place at residue 193, we can safely restrict our search to residue-pair interactions involving residue 193. Now, we use the PyRosetta[21] tool `print_residue_pair_energies()` to obtain a list of all residue pair interactions involving residue 193. Inspecting the output in `native_residue_pair_interactions.csv` and `mutant_residue_pair_interactions.csv` we can find a list of significant pair energy changes between residue 193 and other surrounding residues.

PyRosetta tools can also be used to analyze atom-pair interactions that contribute most strongly to the critical residue-pair interactions. The scoring machinery in Rosetta treats a residue (protein amino acid, sugar monosaccharide, nucleic acid base) as the simplest unit for calculating pairwise energies. All two body energy terms must define `residue_pair_energy()` to calculate the pairwise energy between two residues. For energies such as hydrogen bonding this is necessary because scoring an individual hydrogen bond using the distance and orientation dependent potential described in the main text requires knowledge of not only the donor hydrogen and the acceptor atoms but also the acceptor and donor base atoms to calculate an energy. However, for other terms in the Rosetta score function (such as Lennard Jones attraction/repulsion, implicit solvation, and electrostatics) the `residue_pair_energy()` method simply sums up all of the pairwise interactions between all atoms in each of the residues. These atom pair energies are not normally reported by the scoring function, however in some situations they can assist in pinpointing which specific atom pair interactions are influencing the residue pair energy most strongly.

The PyRosetta toolkit provides two tools for analyzing specific atom pair energies. First, the `etable_atom_pair_energies()` method takes two residues (`res1`, `res2`) and atom indices specifying one atom on each residue (`atom_index_1`, `atom_index_2`) and calculates atom pair energies for Lennard Jones attractive/repulsive, implicit solvation, and electrostatics using a specified score function (`sfxn`).

The second tool, `print_atom_pair_energy_table(),` is designed to output energies for all pairwise atom pair interactions between two specified residues. For ease of viewing this tool outputs the pairwise energies as a table formatted in a .csv file. The tool takes a `score_type` and score function (`sfxn`) as inputs in addition to two specified residues (`res1`, `res2`) and a specified `output_filename`.

Docking. The coordinate file for 1ztx was downloaded from the Protein Data Bank and cleaned to remove any non-canonical amino acids. The structure was first refined to remove significant clashes in the structure using the following command line:

```
relax.linuxgccrelease -s 1ztx_unbound.pdb -relax:ramp_constriants false \
-relax:constrain_relax_to_start_coords -ex1 -ex2 -use_input_sc -flip_HNQ \
-no_optH false
```

Next, the structure was prepacked and then docked using the procedure described in Chaudhury *et al.*[22] using the *REF2015* energy function.

```
docking_prepack_protocollinux -s 1ztx_relaxed.pdb -partners LH_G \
-dock_rtmin -docking:sc_min

docking_protocol.linuxgccrelease —s 1ztx_unbound_prepacked.pdb —native \
1ztx_native.pdb -ignore_unrecognized_res -ex1 -ex2aro -dock_pert 3 8 \
-partners LH_G -nstruct 1000
```

Finally, the interface scores were extracted from the output score file for analysis.

Energy terms for biomolecules other than proteins

An active research area is the development of energy functions compatible for biomolecules other than proteins containing the 20 canonical amino acids. So far, this has involved two approaches: (1) generalizing terms to score non-l amino acids and (2) developing new terms to accommodate other biomolecules. Below, we provide details of the main non-protein energy functions currently being developed in Rosetta.

Generalizing the Existing Energy Terms

The physically-derived terms in the Rosetta energy function capture forces that are general to all biomolecules. Therefore, these terms were generalized to be compatible will D-amino acids, nucleic acids, carbohydrates, and other biomolecules.

Table S7: Summary of energy term compatibility with other biomolecules

Term	Can score
`fa_atr`	All molecules
`fa_rep`	All molecules
`fa_intra_rep`	All molecules
`fa_sol`	All molecules
`lk_ball`	All molecules
`fa_inra_sol`	All molecules
`fa_elec`	All molecules
`hbond_sr_bb`	All molecules
`hbond_lr_bb`	All molecules
`hbond_bb_sc`	All molecules
`hbond_sc`	All molecules
`dslf_fa13`	L-, D-, and mixed D/L disulfide bonds between cysteine or cysteine-like residues (e.g., homocysteine, penacillamine)
`rama_prepro`	Glycine, canonical L-amino acids, their D-counterparts, and similar alpha-amino acids that can use canonical rama tables.
`p_aa_pp`	Glycine, canonical L-amino acids, their D-counterparts, and similar alpha-amino acids that can use canonical rama tables.
`omega`	All α-amino acids, or β-amino acids.
`fa_dun`	All polymer building blocks.
`pro_close`	L- and D-proline.
`yhh_planarity`	L- and D-tyrosine.
`ref`	Glycine, canonical L-amino acids, and their D-counterparts.

Compatibility with D-amino acids

To make the energy terms compatible with D-amino acids, several modifications were made to the torsional terms.[23] First, the ϕ, ψ values were inverted in the `rama_prepro`, `omega`, and `p_aa_p` terms to accommodate the chirality of the backbone. Accordingly, the derivatives were inverted to ensure that mirror-image structures energy-minimize identically. Second, the `fa_dun` score term was modified to invert main chain and side-chain torsional values. Special amino acid-specific score terms, such as `pro_close` and `yhh_planarity`, were updated to recognize D-proline and D-tyrosine, respectively. The `dslf_fa13` term was symmetrized to ensure that mirror-image conformations of mixed D/L disulfides score identically. Finally, the `ref` term was altered to ensure that D-amino acids have a reference energy penalty or bonus identical to that of their L-counterparts. All other score terms were compatible with arbitrary molecules without modification.

Energy terms for non-canonical amino acids

Toward the goal of designing proteins with non-canonical amino acids, Renfrew *et al.* implemented an energy function with terms derived from molecular mechanics. This energy function, called `mm_std`, removes the terms that depend on residue identity (i.e. `rama_prepro`, `p_aa_pp`, `omega`, and `fa_dun`) and replaces them with terms that capture the internal and torsional energy preferences: `mm_lj_intra_rep`, `mm_lj_intra_atr`, and `mm_twist`. The `ref` term is replaced by either a term that explicitly models the unfolded state, (`unfolded`), or a pair of terms that capture the change in energy experienced by an atom of a specific type going from an unfolded to folded environment (`split_unfolded_1b` and `split_unfolded_2b`). These terms were developed toward the goal of designing proteins containing non-canonical alpha-amino acid residues. It has also been used to model oligo-oxypiperizines (OOPs),[24] hydrogen bond surrogates (HBS), oligo-peptoids,[25] and hybrid molecules.

Intra-residue van der Waals interactions are calculated between atom pairs from the same residue using a Lennard-Jones 6-12 potential. Like `fa_rep` and `fa_atr`, the potential is divided between attractive (`mm_lj_intra_atr`) and repulsive (`mm_lj_intra_rep`) components that can be weighted separately. The two terms have the same functional form as the inter-reside terms (Eq. 3 and 4 in the main text) but with the following differences. The summed atomic radii, σ_{ij}, and the geometric mean of atomic well-depths, ϵ_{ij}, are based on the CHARMM 24[26] parameters. The terms are applied to all atom pairs in a residue with a bond separation of 3 or more. Some atom-type pairs have different parameters when separated by 3 bonds (and involved in a proper torsion) and when separated by 4 of more bonds, but no connectivity weight is applied. Both attractive and repulsive energies are calculated for hydrogens. The attractive potential is not smoothed and consequently is evaluated to 8 Å.

The **torsional term**, called `mm_twist` (**Eq. S14**), is a molecular mechanics torsion term. It is evaluated for all atom quads involved in proper torsions. To match the intra-residue van der Waals term the parameters for K_θ and n come from CHARMM 24. A given set of 4 atoms types may have multiple K_θ and n parameters that are summed in a Fourier series to more accurately describe the rotation about the central bond of the torsion.

$$E_{\text{twist}} = \sum_i K_\theta (1 - \cos(n\theta)) \qquad \text{(S14)}$$

Explicit Unfolded State Energy (EUSE) represents the unfolded energy of the protein and compensates for the difficultly in packing large side chains (**Eq. S15**). The `ref` term is fit during the weight optimization protocol which is only trained on protein data and therefore incompatible with non-protein residues. The EUSE is the sum over each residue and each term in the energy function where $U(\text{AA}_r, t)$ is the unfolded reference value of residue type, AA_r, of residue, r, and energy term t. The unfolded reference values are the Boltzmann weighted average energies of the central residue of 5-mer fragments of high quality protein structures. The central residue of each fragment was mutated to the residue of choice, repacked and scored and the Boltzmann weighted average for each energy term, t, for each residue type is stored. For peptoids, only XXGPX fragments were used to mimic an oligo-peptoid environment.[27]

$$E_{\text{unfolded}} = \sum_r \sum_t W_t U(AA_r, t) \quad \text{(S15)}$$

Two-Component Reference Energy (TCRE) is a reference energy that compensates for some of the shortcomings of the EUSE; primarily the dependence of the EUSE on short peptide fragments which limits the types of oligomer chemistry to those that contain an α-amino acid backbone (e.g. OOPs, HBS, peptoids; **Eq. S16**). The one-body component is the sum over each residue and each one-body energy term in the energy function where $R_{1\text{B}}(\text{AA}_r, t1\text{b})$ is the one-body reference value of the residue type, AA_r, of residue, r, and one-body scoring term $t1b$. The one-body reference values are the unweighted $t1b$ energies for each energy term, taken from lowest energy conformation of that residue type in the context of a didpeptide model system. The two-body component is the sum over each atom and each two-body energy term in the energy function where $R_{2\text{B}}(T_i, t2\text{b})$ is the two-body reference value for atom type, T_i, of atom, i, and two-body energy term $t2b$. The two-body reference values are the median $t2b$ energy of an atom of type T_i in the context of a folded protein.

$$E_{\text{TCRE}} = \overbrace{\left(\sum_r \sum_{t1b} W_{t1\text{b}} R_{1\text{B}}(\text{AA}_r, t1\text{b})\right)}^{\text{one-body}} + \overbrace{\left(\sum_i \sum_{t2b} W_{t2\text{b}} R_{2\text{B}}(T_i, t2\text{b})\right)}^{\text{two-body}} \quad \text{(S16)}$$

Reference values were determined using structures from the Top8000 database.[28] The effect is to produce a single reference value for a residue type just like the `ref` and `unfolded` terms. The term is a measure of the difference between the base energy of inherent to a peptide sequence and the average interaction that sequence would make with itself when folded. Currently W_{t1b} and W_{t1b} are set to the weight of that term in the energy function but could be modified.

Energy terms for carbohydrates

To model realistic carbohydrate geometries, Rosetta implements the `sugar_bb` term which rewards preferred glycosidic torsion angles.[29] The `sugar_bb` term is a mixture of functions specific to glycosidic torsions and linkage types. For most torsion/linkage types, Rosetta uses the CHarbohydrate-Intrinsic (CHI) energy functions developed from quantum mechanical calculations with isomers of *O*-linked tetrahydropyran oligomers.[30,31] The data were fit to Gaussian functions and matched with statistical data. Together, they are used to compute the energy, given as a function of some torsion angle *x* in degrees, magnitude of the Gaussian distribution *a*, midpoint of the distribution *b*, the intercept of the distribution *d*, and a constant *c* which is twice the square width of the distribution (**Eq. S17**).

$$E_{\text{sugar\_bb}} = d + \sum_i a_i e^{-(x-b_i)^2/c_i} \quad \text{(S17)}$$

For ω torsions, the energy is instead modeled using a series of parabolic functions with coefficients fit to statistical data and centered around the ideal staggered and Gauche conformations. This energy is defined as a function of the torsion angle *x* (in degrees), a constant to define the parabola width, *k*, the vertex of the parabola θ, and the energy difference relative to the minimum *b* (**Eq. S18**). This function approximates the so-called Gauche effect.

$$E_{\text{sugar\_bb}} = k(x - \theta)^2 + b \quad \text{(S18)}$$

The `sugar_bb` score per residue is the sum of each function for each glycosidic torsion in the residue. **Table S8** lists the functional form for each torsion and linkage type. (The functions assume that D-sugars are in the $^4C_1$ chair conformation and that L-sugars are in the $^1C^4$ chair conformation.)

Table S8: Functional form of the sugar backbone energy for each torsion and linkage type

Angle	Ax./eq. designation	Stereoisomer	Exocyclic	Range	Functional form
ϕ	axial (α)	D	—	–180°–180°	Gaussian
	eqiuatorial (β)	D	—	–180°–180°	Gaussian
	axial (α)	L	—	–180°–180°	Gaussian, *x*=–ϕ
	equatorial (β)	L	—	–180°–180°	Gaussian, *x*=–ϕ
ψ	ax. (parent at odd O)	D (parent)	no	0–360°	Gaussian
	eq. (parent at odd O)	D (parent)	no	0–360°	Gaussian
	ax. (parent at even O)	D (parent)	no	0–360°	Gaussian
	eq. (parent at even O)	D (parent)	no	0–360°	Gaussian
	ax. (parent at odd O)	L (parent)	no	0–360°	Gaussian, 360°–ψ
	eq. (parent at odd O)	L (parent)	no	0–360°	Gaussian, 360°–ψ
	ax. (parent at even O)	L (parent)	no	0–360°	Gaussian, 360°–ψ
	eq. (parent at even O)	L (parent)	no	0–360°	Gaussian, 360°–ψ
	axial (α)	D	yes	0–360°	Gaussian
	equatorial (β)	D	yes	0–360°	Gaussian
	axial (α)	L	yes	0–360°	Gaussian, 360°–ψ
	equatorial (β)	L	yes	0–360°	Gaussian, 360°–ψ
ω	axial (parent O4)	D (parent)	yes	0–120°	parabolic
	axial (parent O4)	D (parent)	yes	120°–240°	parabolic
	axial (parent O4)	D (parent)	yes	240°–360°	parabolic
	eq. (parent O4)	D (parent)	yes	0–120°	parabolic
	eq. (parent O4)	D (parent)	yes	120°–240°	parabolic
	eq. (parent O4)	D (parent)	yes	240°–360°	parabolic
	axial (parent O4)	L (parent)	yes	0–360°	parabolic, 360°–ω
	axial (parent O4)	L (parent)	yes	120°–240°	parabolic, 360°–ω
	axial (parent O4)	L (parent)	yes	240°–360°	parabolic, 360°–ω
	eq. (parent O4)	L (parent)	yes	0–120°	parabolic, 360°–ω
	eq. (parent O4)	L (parent)	yes	120°–240°	parabolic, 360°–ω
	eq. (parent O4)	L (parent)	yes	240°–360°	parabolic, 360°–ω

Energy terms for nucleic acids

The Rosetta energy function captures van der Waals and electrostatic forces general to all biomolecules. However, these terms do not capture rules specific to the geometry and pairing of nucleic acid bases. Therefore, Das and coworkers have implemented terms to explicitly capture these rules.

Electrostatics. The standard Rosetta electrostatic potential (`fa_elec`) disfavors Watson-Crick base pairs due to repulsion between the fixed positive charges on the hydrogen atoms in close proximity in G-C and A-U pairs. To alleviate this problem, Rosetta uses two modified terms to evaluate electrostatics involving RNA bases. First, electrostatic interactions between phosphate atoms are evaluated using the standard `fa_elec` potential (**Eq. 10** in the main text), via a term called `fa_elec_rna_phos_phos`. Second, electrostatic interactions between RNA bases are captured using the `stack_elec` term.[32] This term scales the `fa_elec` potential as a function of the angle (κ_i) between the normal to the plane of the base (z_i) and the vector $d_{i,j}$ between base heavy atoms i and j in residues r_1 and r_2, respectively (Figure S4). The equation for `stack_elec` is given by **Eq. S19**.

$$E_{\text{stack\_elec}} = \textstyle\sum_{r_1<r_2} \sum_{i,j} f(\kappa_i, \kappa_j) E_{\text{fa\_elec}} \qquad \text{(S19)}$$

The scaling function $f(\kappa_i, \kappa_j)$ suppresses the electrostatic energy to zero when the bases are coplanar and maintains the full value of the energy when the bases are stacked (**Eq. S20; Fig. S4B**).

$$f(\kappa_i, \kappa_j) = cos^2(\kappa_i) + cos^2(\kappa_j) \qquad \text{(S20)}$$

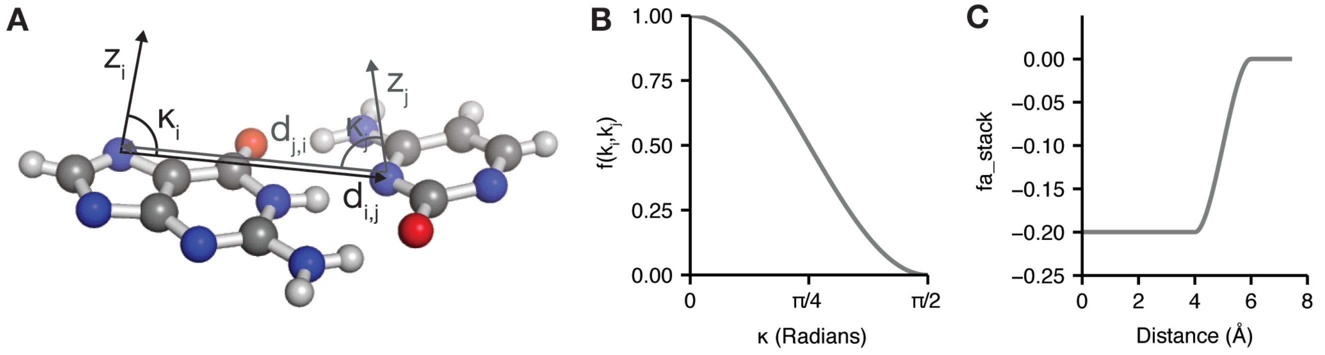


Figure S4. **Electrostatic and stacking energies for RNA.**
(A) `fa_stack` and `stack_elec` are scaled as a function of the angle, κ_i, between the normal to the base, z_i, and the distance vector between atoms i and j. (B) The scaling function takes the form $f(\kappa_i) = \cos^2(\kappa_i)$, such that the weight is equal to 1.0 when the bases are stacked and 0 when they are coplanar. (C) The `fa_stack` energy for stacked bases (when $f(\kappa_i) = 1.0$).

Base stacking. $\pi - \pi$ stacking interactions are not explicitly captured by `fa_atr`; thus, Rosetta includes an additional stacking bonus term, called `fa_stack`.[33] The `fa_stack` term applies a constant bonus for base atoms less than 4 Å from each other to reward neighboring stacked bases. Like the `stack_elec` term, `fa_stack` also depends on the angle (κ_i) between the normal to the plane of the base (z_i) and the distance vector from atoms i to j ($d_{i,j}$), such that stacked, but not coplanar bases receive this bonus (**Eq. S19: Fig. 4C**). The potential is smoothed to zero between 4 Å and 6 Å using a smoothing function given in **Eq. S21-S23**.

$$E_{\text{fa\_stack}} = \sum_{r_1<r_2} \sum_{i,j} f\left(\kappa_i, \kappa_j\right) g\left(\left|d_{i,j}\right|\right) \qquad \text{(S21)}$$

$$g\left(\left|d_{i,j}\right|\right) = \begin{cases} -0.2, & \left|d_{i,j}\right| \leq 4.0 \\ -0.2h(\left|d_{i,j}\right|) & 4.0 < \left|d_{i,j}\right| < 6.0 \\ \;0.0, & \left|d_{i,j}\right| \geq 6.0 \end{cases} \qquad \text{(S22)}$$

$$h\left(\left|d_{i,j}\right|\right) = \; -0.2\left[2\left(\frac{\left|d_{i,j}\right|-4}{2}\right)^3 - 3\left(\frac{\left|d_{i,j}\right|-4}{2}\right)^2 + 1\right] \qquad \text{(S23)}$$

RNA torsions. Like carbohydrates and non-canonical amino acids, nucleic acids require a separate term to evaluate specific torsional energies. For RNA, the `rna_torsion` term evaluates the energies for the nucleic acid backbone and side chain torsions: α, β, γ, δ, ε, ζ, v_1, v_2, χ, O2'. The torsional energies are computed as a function of the frequency of some general torsion A found in RNA structures in the PDB (**Eq. S24, Fig. S5**).

$$E_{\text{rna\_torsion}} = \sum_k -\ln\left(P(A_k)\right) \qquad \text{(S24)}$$

To accommodate special cases, separate potentials were derived for each of the δ, ε, v_1, v_2, χ, O2' torsions depending on whether the sugar pucker is 2'-endo or 3'-endo. Additionally, a separate χ potential was derived for purines and pyrimidines. For ζ, there are three separate potentials depending on whether the α torsion of the following residue is gauche$^-$, trans, or gauche$^+$. Additionally, a set of four harmonic restraints, together comprising `rna_sugar_close`, are applied to ensure that the RNA sugar ring remains closed: a bond distance restraint between atoms O4' and C1', and three angle restraints for the O4'-C1'-C2', C4'-O4'-C1', and O4'-C1'-first base atom angles.

Solvation. The full atom RNA potential contains an orientation-dependent desolvation penalty for polar atoms (`geom_sol`). The penalty is equal to the sum of the values of the orientation-dependent Rosetta hydrogen bonding energies for virtual water molecules placed at the positions of each occluding atom. The form of this term is given by **Eq. S25**.

$$E_{\text{geom\_sol}} = \sum_{r_1<r_2} \sum_{i,j} E_{\text{hbond}}(r_i - v_j) \qquad \text{(S25)}$$

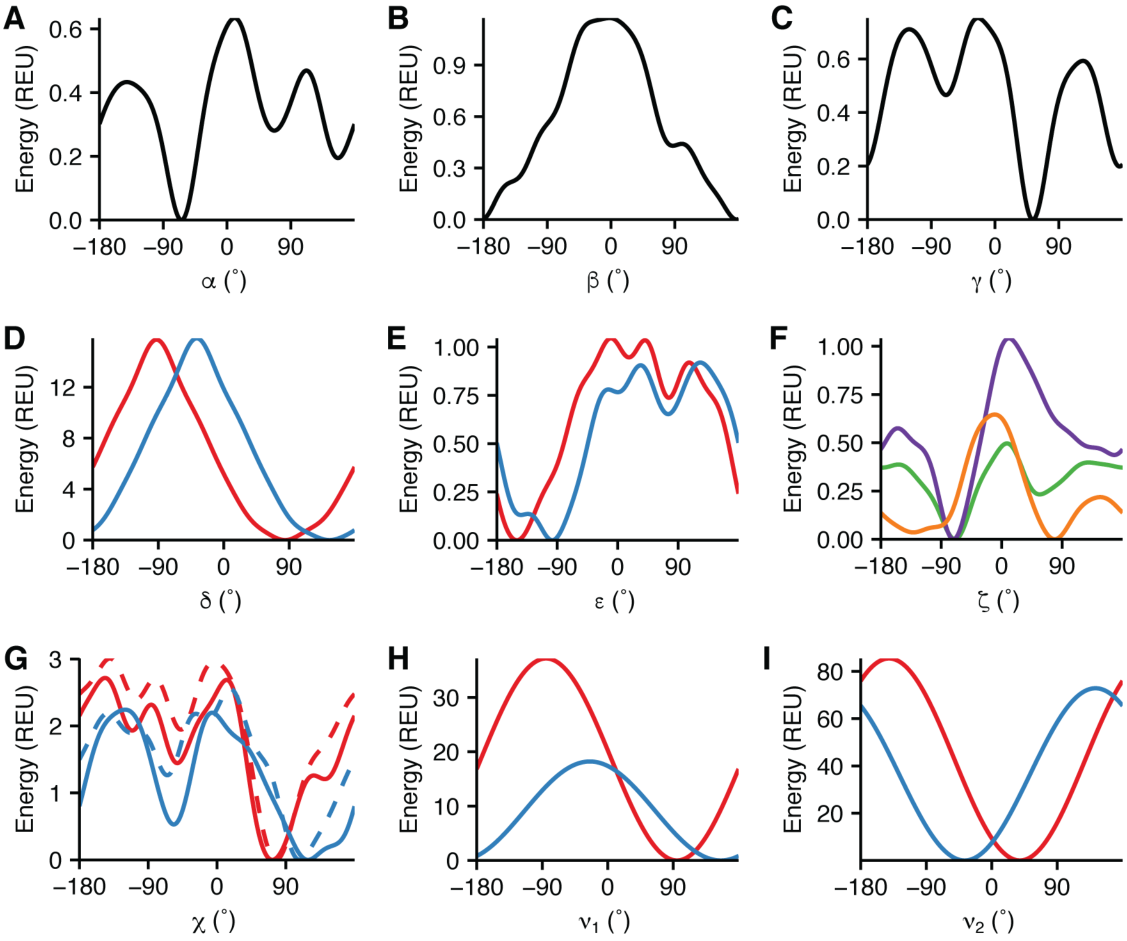


Figure S5. Torsion potentials for RNA

RNA torsional potential for (A) α, (B) β, (C) γ, (D) δ, (E) ε, (F) ζ when the α torsion of the following residue is gauche^{-} (orange), trans (cyan), or gauche^{+} (purple) (G) χ for purines (lighter red and blue) and pyrimidines (darker red and blue), (H) v_1, (I) v_2. Potentials when the sugar pucker is C2'-endo are shown in red and C3'-endo shown in blue.

References

(1) Kuhlman, B.; Baker, D. Native Protein Sequences Are close to Optimal for Their Structures. *Proc Natl Acad Sci U S A* **2000**, *97* (19), 10383–10388.

(2) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science (80-.).* **2003**, *302* (5649), 1364–1368.

(3) Tsai, J.; Bonneau, R.; Morozov, A. V.; Kuhlman, B.; Rohl, C. A.; Baker, D. An Improved Protein Decoy Set for Testing Energy Functions for Protein Structure Prediction. *Proteins Struct. Funct. Bioinforma.* **2003**, *53* (1), 76–87.

(4) Meiler, J.; Baker, D. ROSETTALIGAND: Protein-Small Molecule Docking with Full Side-Chain Flexibility. *Proteins Struct. Funct. Bioinforma.* **2006**, *65* (3), 538–548.

(5) Park, H.; Bradley, P.; Greisen, P.; Liu, Y.; Kim, D. E.; Baker, D.; DiMaio, F. Simultaneous Optimization of Biomolecular Energy Function on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **2016**, *12* (12), 6201–6212.

(6) Kuhlman, B.; Baker, D. Native Protein Sequences Are close to Optimal for Their Structures. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (19), 10383–10388.

(7) Yanover, C.; Bradley, P. Extensive Protein and DNA Backbone Sampling Improves Structure-Based Specificity Prediction for C2H2 Zinc Fingers. *Nucleic Acids Res.* **2011**, *39* (11), 4564–4576.

(8) O'Meara, M. J.; Leaver-Fay, A.; Tyka, M. D.; Stein, A.; Houlihan, K.; DiMaio, F.; Bradley, P.; Kortemme, T.; Baker, D.; Snoeyink, J.; Kuhlman, B. Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **2015**, *11* (2), 609–622.

(9) Kortemme, T.; Morozov, A. V; Baker, D. An Orientation-Dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes. *J. Mol. Biol.* **2003**, *326* (4), 1239–1259.

(10) Dunbrack, R. L.; Cohen, F. E.; Cohen, F. E. Bayesian Statistical Analysis of Protein Side-Chain Rotamer Preferences. *Protein Sci.* **1997**, *6* (8), 1661–1681.

(11) Shapovalov, M. V; Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011**, *19* (6), 844–858.

(12) Leaver-Fay, A.; O'Meara, M. J.; Tyka, M.; Jacak, R.; Song, Y.; Kellogg, E. H.; Thompson, J.; Davis, I. W.; Pache, R. A.; Lyskov, S.; Gray, J. J.; Kortemme, T.; Richardson, J. S.; Havranek, J. J.; Snoeyink, J.; Baker, D.; Kuhlman, B. Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. *Methods Enzymol.* **2013**, *523*, 109–143.

(13) William L. Jorgensen, *; David S. Maxwell, and; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. **1996**.

(14) Lennard-Jones, J. On the Determination of Molecular Fields II: From the Variation of Viscosity of a Gas with Temperature. *R. Soc. London, Ser. A, Contain. Pap. a Math. Phys. Character* **1924**, *106*, 464–477.

(15) Lennard-Jones, J. On the Determination of Molecular Fields I: From the Variation of Viscosity of a Gas with Temperature. *R. Soc. London, Ser. A, Contain. Pap. a Math. Phys. Character* **1924**, *106*, 441–462.

(16) Lazaridis, T.; Karplus, M. Effective Energy Function for Proteins in Solution. *Proteins* **1999**, *35* (2), 133–152.

(17) Neria, E.; Fischer, S.; Karplus, M. Simulation of Activation Free Energies in Molecular Systems. *J. Chem. Phys.* **1996**, *105* (5), 1902–1921.

(18) Shapovalov, M. V; Dunbrack Jr., R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011**, *19*

(6), 844–858.

(19) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(20) Kellogg, E. H.; Leaver-Fay, A.; Baker, D. Role of Conformational Sampling in Computing Mutation-Induced Changes in Protein Structure and Stability. *Proteins* **2011**, *79* (3), 830–838.

(21) Chaudhury, S.; Lyskov, S.; Gray, J. J. PyRosetta: A Script-Based Interface for Implementing Molecular Modeling Algorithms Using Rosetta. *Bioinformatics* **2010**, *26* (5), 689–691.

(22) Chaudhury, S.; Berrondo, M.; Weitzner, B. D.; Muthu, P.; Bergman, H.; Gray, J. J. Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2. *PLoS One* **2011**, *6* (8), e22477.

(23) Bhardwaj, G.; Mulligan, V. K.; Bahl, C. D.; Gilmore, J. M.; Harvey, P. J.; Cheneval, O.; Buchko, G. W.; Pulavarti, S. V. S. R. K.; Kaas, Q.; Eletsky, A.; Huang, P.-S.; Johnsen, W. A.; Greisen, P. J.; Rocklin, G. J.; Song, Y.; Linsky, T. W.; Watkins, A.; Rettie, S. A.; Xu, X.; Carter, L. P.; Bonneau, R.; Olson, J. M.; Coutsias, E.; Correnti, C. E.; Szyperski, T.; Craik, D. J.; Baker, D. Accurate de Novo Design of Hyperstable Constrained Peptides. *Nature* **2016**, *538* (7625), 329–335.

(24) Lao, B. B.; Drew, K.; Guarracino, D. A.; Brewer, T. F.; Heindel, D. W.; Bonneau, R.; Arora, P. S. Rational Design of Topographical Helix Mimics as Potent Inhibitors of Protein–Protein Interactions. *J. Am. Chem. Soc.* **2014**, *136* (22), 7877–7888.

(25) Drew, K.; Renfrew, P. D.; Craven, T. W.; Butterfoss, G. L.; Chou, F.-C.; Lyskov, S.; Bullock, B. N.; Watkins, A.; Labonte, J. W.; Pacella, M.; Kilambi, K. P.; Leaver-Fay, A.; Kuhlman, B.; Gray, J. J.; Bradley, P.; Kirshenbaum, K.; Arora, P. S.; Das, R.; Bonneau, R. Adding Diverse Noncanonical Backbones to Rosetta: Enabling Peptidomimetic Design. *PLoS One* **2013**, *8* (7), e67051.

(26) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.

(27) Renfrew, P. D.; Craven, T. W.; Butterfoss, G. L.; Kirshenbaum, K.; Bonneau, R. A Rotamer Library to Enable Modeling and Design of Peptoid Foldamers. *J. Am. Chem. Soc.* **2014**, *136* (24), 8772–8782.

(28) Richardson, J. S.; Keedy, D. A.; Richardson, D. C. In Biomolecular Forms and Functions: A Celebration of 50 Years of the Ramachandran Map. *World Sci. Publ. Co. Pte. Ltd Singapore* **2013**, 46–61.

(29) Labonte, J. W.; Aldof-Bryfogle, J.; Schief, W. R.; Gray, J. J. Residue-Centric Modeling and Design of Saccharide and Glycoconjugate Structures. *J. Comput. Chem.* **2017**, *38* (5), 276–287.

(30) Nivedha, A. K.; Thieker, D. F.; Makeneni, S.; Hu, H.; Woods, R. J. Vina-Carb: Improving Glycosidic Angles during Carbohydrate Docking. *J. Chem. Theory Comput.* **2016**, *12* (2), 892–901.

(31) Nivedha, A. K.; Makeneni, S.; Foley, B. L.; Tessier, M. B.; Woods, R. J. Importance of Ligand Conformational Energies in Carbohydrate Docking: Sorting the Wheat from the Chaff. *J. Comput. Chem.* **2014**, *35* (7), 526–539.

(32) Chou, F.-C.; Kladwang, W.; Kappel, K.; Das, R. Blind Tests of RNA Nearest-Neighbor Energy Prediction. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (30), 8430–8435.

(33) Sripakdeevong, P.; Kladwang, W.; Das, R. An Enumerative Stepwise Ansatz Enables Atomic-Accuracy RNA Loop Modeling. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (51), 20573–20578.