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, ‡ 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
◆ 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
△ 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.
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 and stabilized by a fine balance between enthalpic and entropic contributions to non-covalent interactions. 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: 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. 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 ($\Delta H$) and entropy ($\Delta S$), 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 with harmonic torsional potentials that were parametrized using force constants from vibrational spectra of small molecules. These formulations were first applied to investigations of the structures of hemolysin, trypsin inhibitor, and hemoglobin and have now diversified into a large family of commonly used energy functions such as AMBER, DREIDING, OPLS, and CHARMM. Many of these energy functions also rely on new terms and parameterizations. For example, faster computers have enabled the derivation of parameters from ab initio quantum calculations. 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. Additionally, there are alternate models of electrostatics and solvation, such as the generalized Born approximation of the Poisson–Boltzmann equation and polarizable electrostatic terms that accommodate varying charge distributions.

The first version of the Rosetta energy function was developed for proteins by Simons et al. Initially, it used statistical potentials describing individual residue environments and frequent residue-pair interactions derived from the Protein Data Bank (PDB). 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. 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 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, a pairwise-additive implicit solvation model, a statistically derived electrostatics term, and a term for backbone-dependent rotamer preferences. Shortly thereafter, several terms were added, including an orientation-dependent hydrogen-bonding term, in agreement with electronic structure calculations. 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, hot-spot prediction, protein–protein docking, small-molecule docking, and specificity redesign as well as the first de novo designed protein backbone not found in nature and the first computationally designed new protein–protein interface.

The Rosetta energy function has changed dramatically since it was last described in complete detail by Rohl et al. in 2004. It has undergone significant advances ranging from improved models of hydrogen bonding and solvation to updated evaluation of backbone and rotamer conformations. 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 and the design of vaccines. Biomimetic peptides and self-assembling materials, and enzymes that perform new functions. 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 and for all macromolecules, including RNA, DNA, small-molecule ligands, noncanonical amino acids and backbones, and carbohydrates, 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 $\Delta E_{\text{total}}$, is computed as a linear combination of energy terms $E_i$ that are calculated as functions of geometric degrees of freedom ($\Theta$) and chemical identities ($\alpha\alpha$) and scaled by a weight on each term ($w_i$), as shown in eq 1:

$$\Delta E_{\text{total}} = \sum_i w_i E_i(\Theta, \alpha\alpha)$$

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
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, which calculates the interaction energy of atoms $i$ and $j$ in different residues given the sum of their atomic radii, $\sigma_{ij}$ (in Rosetta it has the same definition as $r_{ij}^{\text{min}}$ in CHARMM), the atom-pair distance, $d_{ij}$, and the geometric mean of their well depths, $\epsilon_{ij}$ (eq 2):

$$E_{\text{vdw}}(i, j) = \epsilon_{ij} \left( \frac{\sigma_{ij}}{d_{ij}} \right)^{12} - 2 \left( \frac{\sigma_{ij}}{d_{ij}} \right)^6$$

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 ($d_{ij} = \sigma_{ij}$) 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

---

**Table 1. Summary of Terms in REF15 for Proteins**

<table>
<thead>
<tr>
<th>term</th>
<th>description</th>
<th>weight</th>
<th>units</th>
<th>ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fa_atr</td>
<td>attractive energy between two atoms on different residues separated by a distance $d$</td>
<td>1.0</td>
<td>kcal/mol</td>
<td>5, 6</td>
</tr>
<tr>
<td>fa_rep</td>
<td>repulsive energy between two atoms on different residues separated by a distance $d$</td>
<td>0.55</td>
<td>kcal/mol</td>
<td>5, 6</td>
</tr>
<tr>
<td>fa_intra_rep</td>
<td>repulsive energy between two atoms on the same residue separated by a distance $d$</td>
<td>0.005</td>
<td>kcal/mol</td>
<td>5, 6</td>
</tr>
<tr>
<td>fa.sol</td>
<td>Gaussian exclusion implicit solution energy between protein atoms in different residues</td>
<td>1.0</td>
<td>kcal/mol</td>
<td>36</td>
</tr>
<tr>
<td>lk_ball_wtd</td>
<td>orientation-dependent solvation of polar atoms assuming ideal water geometry</td>
<td>1.0</td>
<td>kcal/mol</td>
<td>50, 71</td>
</tr>
<tr>
<td>fa_intra_sol</td>
<td>Gaussian exclusion implicit solution energy between protein atoms in the same residue</td>
<td>1.0</td>
<td>kcal/mol</td>
<td>36</td>
</tr>
<tr>
<td>fa_elec</td>
<td>energy of interaction between two nonbonded charged atoms separated by a distance $d$</td>
<td>1.0</td>
<td>kcal/mol</td>
<td>50</td>
</tr>
<tr>
<td>hbond_lr_bb</td>
<td>energy of short-range hydrogen bonds</td>
<td>1.0</td>
<td>kcal/mol</td>
<td>38, 49</td>
</tr>
<tr>
<td>hbond_sc</td>
<td>energy of side-chain–side-chain hydrogen bonds</td>
<td>1.0</td>
<td>kcal/mol</td>
<td>38, 49</td>
</tr>
<tr>
<td>dslf_fa13</td>
<td>energy of disulfide bridges</td>
<td>1.25</td>
<td>kcal/mol</td>
<td>49</td>
</tr>
<tr>
<td>rama_prepro</td>
<td>probability of backbone $\phi$, $\psi$ angles given the amino acid type</td>
<td>(0.45 kcal/mol)/$kT$</td>
<td>$kT$</td>
<td>50, 51</td>
</tr>
<tr>
<td>p_aa_pp</td>
<td>probability of amino acid identity given backbone $\phi$, $\psi$ angles</td>
<td>(0.4 kcal/mol)/$kT$</td>
<td>$kT$</td>
<td>51</td>
</tr>
<tr>
<td>fa_dun</td>
<td>probability that a chosen rotamer is native-like given backbone $\phi$, $\psi$ angles</td>
<td>(0.7 kcal/mol)/$kT$</td>
<td>$kT$</td>
<td>52</td>
</tr>
<tr>
<td>omega</td>
<td>backbone-dependent penalty for cis $\omega$ dihedrals that deviate from 0° and trans $\omega$ dihedrals</td>
<td>(0.6 kcal/mol)/AU</td>
<td>AU</td>
<td>72</td>
</tr>
<tr>
<td>pro_close</td>
<td>penalty for an open proline ring and proline $\omega$ bonding energy</td>
<td>(1.25 kcal/mol)/AU</td>
<td>AU</td>
<td>51</td>
</tr>
<tr>
<td>yhh_planarity</td>
<td>sinusoidal penalty for nonplanar tyrosine $\chi_2$ dihedral angle</td>
<td>(0.625 kcal/mol)/AU</td>
<td>AU</td>
<td>49</td>
</tr>
<tr>
<td>ref</td>
<td>reference energies for amino acid types</td>
<td>(1.0 kcal/mol)/AU</td>
<td>AU</td>
<td>1, 51</td>
</tr>
</tbody>
</table>

“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 $\epsilon_{\text{Nbb}} = 0.162$ kcal/mol and $\epsilon_{\text{Cbb}} = 0.063$ kcal/mol and atomic radii of 1.763 Å and 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 $\epsilon = 10$ and partial charges $q_{\text{Nbb}} = -0.604e$ and $q_{\text{Cbb}} = 0.090e$ (red) compared with Rosetta fa_elec (blue). The fa_elec model is shifted to reach zero at the cutoff distance of 6.0 Å.
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.23

The repulsive van der Waals energy, $E_{\text{a\_rep}}$, 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 $1/d_{ij}^{12}$ 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 $1/d_{ij}^{12}$ term with a softer linear term for $d_{ij} \leq 0.6 \sigma_{ij}$. The term is computed using the atom-type-specific parameters $m_{ij}$ and $b_{ij}$ which are fit to ensure derivative continuity at $d_{ij} = 0.6 \sigma_{ij}$. After the linear component, the function transitions smoothly to the 6–12 form until $d_{ij} = \sigma_{ij}$ where it reaches zero and remains zero (eq 3 and Figure 1A):

$$E_{\text{a\_rep}}(i,j) = \sum_{i,j} w_{ij}^{\text{conn}}$$

$$d_{ij} = 0.6 \sigma_{ij}$$

$$\begin{cases} 
    m_{ij} d_{ij} + b_{ij} & d_{ij} \leq 0.6 \sigma_{ij} \\
    \sigma_{ij} d_{ij}^{12} - 2 \left( \sigma_{ij} d_{ij}^{6} + 1 \right) & 0.6 \sigma_{ij} < d_{ij} \leq \sigma_{ij} \\\n    \sigma_{ij} < d_{ij} & \sigma_{ij} < d_{ij} \\
    0 & d_{ij} \geq \sigma_{ij}
\end{cases}$$

(3)

Rosetta also includes an intraresidue version of the repulsive component, $E_{\text{a\_intra\_rep}}$, with the same functional form as the $E_{\text{a\_rep}}$ term (eq 3). We include this term because the knowledge-based rotamer energy ($E_{\text{a\_dun}}$ see below) underestimates intraresidue collisions.

The attractive van der Waals energy, $E_{\text{a\_at}}$, has a value of $-\epsilon_{ij}$ for $d_{ij} \leq \sigma_{ij}$ and then transitions to the 6–12 potential as the distance increases (eq 4 and Figure 1B):

$$E_{\text{a\_at}} = \sum_{i,j} w_{ij}^{\text{conn}}$$

$$\begin{cases} 
    -\epsilon_{ij} & d_{ij} \leq \sigma_{ij} \\
    \epsilon_{ij} \left[ \sigma_{ij} d_{ij}^{12} - 2 \left( \sigma_{ij} d_{ij}^{6} + 1 \right) \right] & \sigma_{ij} < d_{ij} \leq 4.5 \text{ Å} \\
    f_{\text{poly}}(d_{ij}) & 4.5 \text{ Å} < d_{ij} \leq 6.0 \text{ Å} \\\n    0 & 6.0 \text{ Å} < d_{ij}
\end{cases}$$

(4)

For speed, we truncate the LJ term beyond $d_{ij} = 6.0$ Å, where the van der Waals forces are small. To avoid derivative discontinuities, we use a cubic polynomial function, $f_{\text{poly}}(d_{ij})$, for $d_{ij} > 4.5$ 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, $w_{ij}^{\text{conn}}$, to exclude the large repulsive energetic contributions that would otherwise be calculated for atoms separated by fewer than four chemical bonds (eq 5).

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

(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 ($n_{ij}^{\text{bonds}}$) 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 $w_{ij}^{\text{conn}}$ 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 $d_{ij}$ the dielectric constant, $\epsilon$, 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):

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

(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 $\epsilon$ in eq 6 with a sigmoidal function $\epsilon(d_{ij})$ that increases from $\epsilon_{\text{core}} = 6$ to $\epsilon_{\text{solvent}} = 80$ when $d_{ij}$ is between 0 and 4 Å (eqs 7 and 8):

$$\epsilon(d_{ij}) = g \left( \frac{d_{ij}}{4} \right) \epsilon_{\text{core}} + \left( 1 - g \left( \frac{d_{ij}}{4} \right) \right) \epsilon_{\text{solvent}}$$

(7)

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

(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 $E_{\text{ehic}}(d_{\text{max}})$ for $d_{ij} < 1.45$ 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 $d_{ij} = d_{\text{max}}$ (eq 9):

$$E_{\text{ehic}}(i,j, d_{ij}) = \begin{cases} 
    \frac{1}{d_{ij}} - \frac{1}{d_{\text{max}}} & d_{ij} \leq d_{\text{max}} \\
    0 & d_{\text{max}} < d_{ij}
\end{cases}$$

(9)

We use the cubic polynomials $f_{\text{elec,low}}(d_{ij})$ and $f_{\text{elec,high}}(d_{ij})$ to smooth between the traditional form and our adjustments while avoiding derivative discontinuities. The energy is also multiplied by the connectivity weight, $w_{ij}^{\text{conn}}$ (eq 5). The final
modified electrostatic potential is given by eq 10 and compared with the standard form in Figure 1C.

\[
E_{i\text{,sol}} = \sum_{i,j} w_{i,j}^{\text{conn}} E_{\text{elec}}^{\text{low}}(i, j, \sigma_{i,j}) \\
\quad + \sum_{i,j} \frac{\Delta G_{i,j}^{\text{free}}}{2\pi^{3/2} \lambda \sigma_i^3} \exp \left( -\frac{(d_{ij} - \sigma_{i,j})^2}{\lambda_i} \right)
\]

(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. Rosetta’s solvation model has two components: an isotropic solvation energy, called \(fa_{\text{sol}}\), which assumes that bulk water is uniformly distributed around the atoms (Figure 2A), and an anisotropic solvation energy, called \(lk_{\text{ball \_wt}}\), which accounts for specific waters near polar atoms that form the solvation shell (Figure 2B).

The isotropic (LK) model is based on the function \(f_{\text{desolv}}\) 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 \(\Delta G^{\text{vap}}\) term because we implement our own reference energy (discussed later). The energy of the atom-pair interaction varies with the separation distance, \(d_{ij}\), the experimentally determined vapor-to-water transfer free energy, \(\Delta G_{i,j}^{\text{vap}}\), the sum of the atomic radii, \(\sigma_{i,j}\), the correlation length, \(\lambda_i\), and the atomic volume of the desolvating atom, \(V_j\) (eq 11):

\[
f_{\text{desolv}} = -\frac{\Delta G_{i,j}^{\text{free}}}{2\pi^{3/2} \lambda \sigma_i^3} \exp \left( -\frac{(d_{ij} - \sigma_{i,j})^2}{\lambda_i} \right)
\]

At short distances, \(fa_{\text{rep}}\) prevents atoms from overlapping; however, many protocols briefly downweight or disable the \(fa_{\text{rep}}\) term. To avoid scenarios where \(f_{\text{desolv}}\) encourages atom-pair overlap in the absence of \(fa_{\text{rep}}\), we smoothly increase the value of the function to a constant at short distances where the van der Waals spheres overlap (\(d_{ij} = \sigma_{i,j}\)). 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 \(f_{\text{poly}}^{\text{low}}\) and \(f_{\text{poly}}^{\text{high}}\) 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:

\[
g_{\text{desolv}} = \begin{cases} 
  f_{\text{desolv}}(i, j, \sigma_{i,j}) & d_{ij} \leq \sigma_{i,j} - c_0 \\
  f_{\text{poly}}^{\text{low}}(i, j, d_{ij}) & \sigma_{i,j} - c_0 < d_{ij} \leq \sigma_{i,j} + c_1 \\
  f_{\text{desolv}}(i, j, d_{ij}) & \sigma_{i,j} + c_1 < d_{ij} \leq 4.5\text{ Å} \\
  f_{\text{poly}}^{\text{high}}(i, j, d_{ij}) & 4.5\text{ Å} < d_{ij} \leq 6.0\text{ Å} \\
  0 & 6.0\text{ Å} < d_{ij}
\end{cases}
\]

The total isotropic solvation energy, \(E_{i\text{,sol}}\), 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):

\[
E_{i\text{,sol}} = \sum_{i,j} w_{i,j}^{\text{conn}} [g_{\text{desolv}}(i, j) + g_{\text{desolv}}(j, i)]
\]
Rosetta also includes an intraresidue version of the isotropic solvation energy, \(fa_{\text{intra\_sol}}\), with the same functional form as the \(fa_{\text{sol}}\) term (eq 13).

A recent innovation (2016) is the addition of the energy term \(lk_{\text{ball\_wtd}}\) 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 \(i\), \(W_i = \{w_1, w_2, \ldots\}\). 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 \(\sigma_j\)) touches the van der Waals sphere of the water (with radius \(\sigma_w\)) 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 \(\nu_{ik}\), we compute an occlusion measure \(d_k^2\) to capture the gap between the hypothetical water and the desolvating atom \(j\), using the offset \(\Omega = 3.7\ \text{Å}^2\) to provide the ramp-up buffer (eq 14):

\[
d_k^2 = \| r_j - \nu_{ik} \|_2^2 - (\sigma_w + \sigma_j)^2 + \Omega
\]  

Next, we find the soft minimum of \(d_k^2\) over all water sites in \(W_i\) by computing the logarithmic average:

\[
d^2_{\text{min}}(i, j) = -\ln \left[ \frac{1}{|W_i|} \sum_{k \in W_i} \exp(-d_k^2) \right]
\]

Then \(d^2_{\text{min}}\) and \(\Omega\) are used to compute a damping function, \(f_{\text{frac}}\), 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):

\[
f_{\text{frac}}(i, j) = \begin{cases} 
1 & d^2_{\text{min}}(i, j) < 0 \\
1 - \left( \frac{d^2_{\text{min}}(i, j)}{\Omega} \right)^2 & 0 \leq d^2_{\text{min}}(i, j) < \Omega \\
0 & \Omega \leq d^2_{\text{min}}(i, j) 
\end{cases}
\]

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

\[
E_{lk_{\text{ball}}}(i, j) = w_{\text{aniso}} (\&_{\text{desolv}}(i, j)) f_{\text{frac}}(i, j)
\]

The amount of isotropic solvation energy subtracted is \(\&_{\text{desolv}}\) multiplied by \(w_{\text{iso}}\) where \(w_{\text{iso}}\) is an atom-type-specific weight, typically ∼0.3 (eq 18):
The total weight on the isotropic contribution through both the \( fa_\text{sol} \) and \( lk_\text{ball_wtd} \) terms is thus \( \sim 0.7 \). The isotropic and anisotropic components are then summed to yield a new desolvation function, \( h_{\text{desolv}} \) (eq 19):

\[
h_{\text{desolv}}(i, j) = E_{\text{lk_ball}}(i, j) + E_{\text{lk_ball}}(i, j)
\]

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

\[
E_{\text{lk_ball_wtd}} = \sum_{i \in P} w_{i,j}^\text{con} h_{\text{desolv}}(i, j) + \sum_{j \in P} w_{i,j}^\text{con} h_{\text{desolv}}(j, i)
\]

Figure 2 shows comparisons of \( fa_\text{sol}, lk_\text{ball} \) (eq 17), and the sum of \( fa_\text{sol} \) and \( lk_\text{ball_wtd} \) for the example of an asparagine NH2 desolvated from three different approach angles. As the approach angle varies, the sum of \( lk_\text{ball_wtd} \) and \( fa_\text{sol} \) creates a larger desolvation penalty when water sites are occluded and a smaller penalty otherwise, relative to \( fa_\text{sol} \) alone.

**Hydrogen Bonding.** Hydrogen bonds are partially covalent interactions that form when a nucleophilic heavy atom donates electron density to a polar hydrogen. At short ranges (<2.5 Å), they exhibit geometries that maximize orbital overlap. 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 \( fa_\text{elec} \) and a hydrogen-bonding model that evaluates energies on the basis of orientation preferences of hydrogen bonds found in high-resolution crystal structures. To derive this model, we curated intraprotein polar contacts from ~8000 high-resolution crystal structures (the Top8000 data set) 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 \( \theta_{AHD} \) formed by the acceptor, \( H \), and the donor heavy atom, \( D \); (3) the angle \( \theta_{BAH} \) formed by the acceptor’s parent atom (“base”), \( B \), the acceptor, and the donor; and (4) the torsion angle \( \phi_{BAH} \) formed by the donor \( H \), the acceptor, and two subsequent parent atoms \( B \) and \( B_2 \), the parent atom of \( A \) (or the first atom on the shortest path to the root atom (e.g., \( C_\beta \)). 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 (\( h_{\text{bond_lr_bb}} \)), short-range backbone hydrogen bonds (\( h_{\text{bond_sr_bb}} \)), hydrogen bonds between backbone and side-chain atoms (\( h_{\text{bond_bb_sc}} \)), and hydrogen bonds between side-chain atoms (\( h_{\text{bond_sc}} \)).

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-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 \( E_{\text{bond}} \) term depends on the orbital hybridization of the acceptor, \( \rho \), and the function \( f(x) \) (eq 22) is used for smoothing to avoid derivative discontinuities and ensure that edge-case hydrogen bonds are considered:

\[
E_{\text{bond}} = \sum_{H,A} w_H w_A (E_{\text{HA}} + E_{\text{AH}} + E_{\text{HBAH}} + E_{\text{BAHA}})
\]

\[
f(x) = \begin{cases} \frac{x}{0.025} - 2.5 & -0.025 \leq x \leq 0.1 \\ 0 & 0.1 \leq x \end{cases}
\]

**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 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 \( S \) atoms and computes the energies of these interactions using an orientation-dependent model called \( \text{dslf_fal13} \).

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 \( E_{\text{dSS}(d_{SS})} \). Second, the components due to the angles formed by \( C_\beta \) and \( C_\alpha \) with the S–S bond, \( \theta_{C_\alpha SS} \) and \( \theta_{C_\beta SS} \), respectively, are evaluated as \( E_{\text{C_\alpha SS}(d_{SS})} \).

Third, the components due to the dihedral angles formed by \( C_\gamma C_\alpha C_\beta \) and \( C_\alpha C_\beta C_\delta \) with the S–S bond, \( \phi_{C_\gamma C_\alpha C_\beta SS} \) and \( \phi_{C_\gamma C_\alpha C_\beta SS} \) respectively, are evaluated as \( E_{\text{C_\gamma C_\alpha C_\beta SS}(d_{SS})} \). Finally, the dihedral angle formed by \( C_\gamma C_\alpha C_\beta C_\delta \) and the S–S bond, \( \phi_{C_\gamma C_\alpha C_\beta C_\delta SS} \), is evaluated as \( E_{\text{C_\gamma C_\alpha C_\beta C_\delta SS}(\phi_{C_\gamma C_\alpha C_\beta SS})} \). The complete disulfide bonding energy evaluated for all S–S pairs is given by eq 23:
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. 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 \( \phi \) and \( \psi \) angles, we define an energy term called \( \text{rama\_prepro} \) based on Ramachandran maps for each amino acid using torsions from 3985 protein chains with a resolution of \( \leq 1.8 \) Å, \( R \) factor of \( \leq 0.22 \) and sequence identity of \( \leq 50\% \). Amino acids with low electron density (in the bottom 25th percentile of each residue type) were removed from the data set. The resulting \( \sim 581,000 \) residues were used in adaptive kernel density estimates of Ramachandran maps with a grid step of 10° for both \( \phi \) and \( \psi \). Residues preceding proline are also treated separately because they exhibit distinct \( \phi, \psi \) preferences due to steric interactions with the proline C\_δ. 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 (eq 24 and Figure 5):

\[
E_{\text{rama\_prepro}} = \sum_{i} \left\{ -\ln[P_{\text{reg}}(\phi, \psi|\text{aa})] \right\} \quad \text{C-terminus, } i + 1 \neq \text{Pro} \\
\left\{ -\ln[P_{\text{prepro}}(\phi, \psi|\text{aa})] \right\} \quad i + 1 = \text{Pro}
\]

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

\[
E_{\text{dalf},413} = \sum_{\text{SS}} E_{\text{dalf}}^{|dSS|} + E_{\text{dalf}}^{\phi\psi C}\phi_{\text{SSC}} + E_{\text{dalf}}^{C\psi\psi C}\phi_{\text{SSC}} + E_{\text{dalf}}^{C\phi\psi C}\phi_{\text{SSC}}
\]

Terms for Protein Backbone and Side-Chain Torsions.
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 \( \phi, \psi \) backbone conformation. This term, called \( p_{aa,pp} \), represents the propensity to observe an amino acid relative to the other 19 canonical amino acids. The knowledge-based propensity, \( P(aal\phi, \psi) \) (eq 25), was derived using the adaptive kernel density estimates for \( P(\phi, \psi|aa) \) and Bayes’ rule, and the equation for \( E_{p,aa,pp} \) is given in eq 26 (Figure SD):

\[
P(aal\phi, \psi) = \frac{P(\phi, \psi|aa)P(aa)}{\sum_{aa} P(\phi, \psi|aa)}
\]

\[
E_{p,aa,pp} = \sum_{r} -\ln \left[ \frac{P(aa, l\phi, \psi)}{P(aa)} \right]
\]

**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/bbddep2010/), which contains the frequencies, means, and standard deviations of individual \( \chi \) angles for each amino acid type. The probability has three components: (1) observing a specific rotamer given the backbone dihedral angles; (2) observing specific \( \chi \) angles given the rotamer; and (3) observing the terminal \( \chi \) angle distribution, which is either Gaussian-like or continuous when the terminal \( \chi \) angle is sp\(^2\)-hybridized (eq 27):

\[
P(\chi|l\phi, \psi, aa) = P(\text{rot}|l\phi, \psi, aa) \prod_{k<T} P(\chi_{k}|l\phi, \psi, \text{rot}, aa)
\]

\[
P(\chi_{k}|l\phi, \psi, \text{rot}, aa) = \exp \left[ -\frac{1}{2} \left( \frac{\chi_{k} - \mu\chi(l\phi, \psi|\text{rot}, aa)}{\sigma\chi(l\phi, \psi|\text{rot}, aa)} \right)^2 \right]
\]

where \( T \) is the number of rotamer \( \chi \) 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 \(-60^\circ\), \(\sim 180^\circ\), and \(-60^\circ\) with steep energy barriers between the wells). If the last \( \chi \) torsion is rotameric, the probability \( P(\chi|l\phi, \psi, \text{rot}, aa) \) is fixed at 1. On the other hand, a torsion is nonrotameric if its third atom is sp\(^3\)-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.

The probability of each rotamer, \( P(\text{rot}|l\phi, \psi, aa) \), 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 \( \mu\chi \) and standard deviation \( \sigma\chi \) for each \( \chi \) dihedral in the rotamer as functions of the backbone dihedrals, allowing us to compute a probability for the \( \chi \) values using eq 28:

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

The full form of \( E_{fa_dun} \) 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.

\[
E_{fa_dun} = \sum_{r} -\ln[P(\phi, \psi|aa)] + \sum_{k<T_r} \frac{1}{2} \frac{\chi_{k} - \mu\chi(\phi, \psi|\text{rot}, aa)}{\sigma\chi(\phi, \psi|\text{rot}, aa)}^2 - \ln[P(\chi_{T_r}|l\phi, \psi, \text{rot}, aa)]
\]

Figure 6. Energies for side-chain rotamer conformations. The Dunbrack rotamer energy, \( E_{fa_dun} \), is dependent on both the \( \phi \) and \( \psi \) backbone torsions and the \( \chi \) side-chain torsions. Here we demonstrate the variation of \( E_{fa_dun} \) when the backbone is fixed in an \( \alpha \)-helical conformation with \( \phi = -57^\circ \) and \( \psi = -47^\circ \) and the \( \chi \) values vary. \( \chi_1 \) is shown in blue, \( \chi_2 \) in red, and \( \chi_3 \) in green. (A) \( \chi \)-dependent Dunbrack energy of methionine with an sp\(^3\)-hybridized terminus. (B) \( \chi \)-dependent energy of glutamine with an sp\(^3\)-hybridized \( \chi_3 \) terminus. \( \chi_1, \chi_2 \), and \( \chi_3 \) of methionine and \( \chi_1 \) and \( \chi_2 \) of glutamine express rotameric behavior, while \( \chi_3 \) of the latter expresses broad nonrotameric behavior.

The energy from \( -\ln[P(\phi, \psi|aa)] \) is computed using bicubic-spline interpolation; \( P(\chi_{T_r}|l\phi, \psi, \text{rot}, aa) \) is computed using tricubic-spline interpolation. To save memory, \( \mu\chi(\phi, \psi|\text{rot}, aa) \) and \( \sigma\chi(\phi, \psi|\text{rot}, aa) \) are computed using bilinear interpolation, though this has the effect of producing derivative discontinuities at the \( (\phi, \psi) \) grid boundaries. These discontinuities, however, do not appear to produce noticeable artifacts.

**Terms for Special-Case Torsions.** Peptide bond dihedral angles, \( \omega_i \), remain mostly fixed in a cis or trans conformation and depend on the backbone \( \phi \) and \( \psi \) 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 behavior more accurately, Rosetta implements a backbone-dependent harmonic penalty...
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):

\[
E_{\text{omega}} = \sum_r \ln \left( \frac{1}{6\sqrt{2\pi}} \right) - \ln \left( \frac{1}{\sigma_r(\phi, \psi)}} \right) \sqrt{2\pi} \right) \\
+ \frac{[\omega_r - \mu_\omega(\phi_r, \psi_r)]^2}{2\sigma^2_r(\phi, \psi)}
\]

(30)

The means \(\mu_\omega\) and standard derivations \(\sigma_\omega\) are backbone \((\phi, \psi)\)-dependent, as given by kernel regressions of \(\omega\) on \(\phi\) and \(\psi\).

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. 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 \(O\) and \(N\) (\(\omega\)) given the observed mean \(\mu_\omega\) and standard deviation \(\sigma_\omega\), where \(r\) is the residue index. This term keeps the \(C_\beta\) atom in the peptide plane. Second, to ensure the correct geometry for the two hydrogens bound to \(C_\beta\), we build a virtual nitrogen atom, \(N_v\), off \(C_\beta\) whose coordinate is controlled by \(\chi_3\) (Figures 7A and 7C). The pro_close term seeks to align the virtual atom \(N_v\) directly on top of the real backbone nitrogen. The \(N-C_\beta-N_v\) bond angle and the \(N-C_\beta\) bond length are restrained to their ideal values.

\[
E_{\text{pro_close}} = \sum_{r \in \text{Pro}} \left\{ \frac{(\omega_r - \mu_\omega)^2}{\sigma^2_\omega} \right. \\
+ \frac{||N_v - N_{v,r}||^2}{\sigma^2_{N_vN_{v,r}}} \bigg|_{r \text{ is the N-terminus}} \\
\left. + \frac{||N_v - N_{v,r}||^2}{\sigma^2_{N_vN_{v,r}}} \right|_{r \text{ is the N-terminus}}
\]

(31)

Tyrosine also requires special treatment for its \(\chi_3\) angle because the hydroxyl hydrogen prefers to be in the plane of the aromatic ring. To enforce this preference, Rosetta implements a sinusoidal penalty to model the barrier to a \(\chi_3\) angle that deviates from planarity. This tyrosine hydroxyl penalty is called yhh_planarity (eq 32 and Figure 7C):

\[
E_{\text{yhh_planarity}} = \sum_r \frac{1}{2} \left[ \cos(\pi - 2\chi_{3,i}) + 1 \right]
\]

(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 as well as its ability to discriminate near-native conformations in the absence of experimental data. 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):

\[
E_{\text{cart_length}} = \frac{1}{2} \sum_{i=1}^n k_{\text{length}}(d_i - d_{i,0})^2
\]

(33)

\[
E_{\text{cart_angle}} = \frac{1}{2} \sum_{i=1}^n k_{\text{angle}}(\theta_i - \theta_{i,0})^2
\]

(34)

\[
E_{\text{cart_torsion}} = \frac{1}{2} \sum_{i=1}^l k_{\text{tension}} \left( \int_0^{2\pi} \left( \phi_i - \phi_{i,0} \right)^2 \rho_i \right)^2
\]

(35)

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

The function $f_{\text{pseud}}(x, y)$ wraps $x$ to the range $[0, y)$. To avoid double counting in the case of $E_{\text{cart, pseud}}$ the spring constant $k_{\text{pseud}}$ is zero when the torsion $\phi_1$ 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, $\Delta G^\text{ref}$, for each amino acid, $aa$, (eq 36):

$$E_{\text{ref}} = \sum_i \Delta G^\text{ref}_i(aa)$$

The $\Delta G^\text{ref}_i$ 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. 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 electrostatic energies partially capture torsional preferences and overlap can cause errors as a result of double counting of atomic or residue-specific contributions. 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 to optimize the agreement of Rosetta calculations with small-molecule thermodynamic data and high-resolution structural features. The energy function parameters have evolved over the years by optimization of the performance of multiple scientific benchmarks (Table 2). 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. and Conway and DiMaio, and O’Meara et al. 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. 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. The optimization data show a strong correlation between the experimental data and values predicted by Rosetta ($\Delta G$ upon mutation, $R = 0.994$; small-molecule $\Delta H$, 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.

### 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 ($\Delta G$) of mutation$^*$ and (2) modeling the structure of a protein—protein interface.

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

Rosetta calculates the $\Delta G$ of the T193V mutation to be $-4.95$ kcal/mol, and the experimentally measured value is $-1.11$ kcal/mol. 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 to identify important energetic contributions to...
the function print_atom_pair_energy_table the change in solvation energy (Figure 8C). made a large (>0.05 kcal/mol) and favorable contribution and I80 on HIV peptidase and residue F194 on the peptide with the resulting table, we found that a hydrophobic pocket around 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_residue_pair_energies to decompose the most favorable term, Δfa_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).

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 (∆fa_sol) is positive (unfavorable), indicating that polar residues are buried upon interface formation. Balancing the desolvation penalty, the changes in polar solvation energy (∆lk_ball_wtd) and electrostatics (∆fa_elec) are negative because of the formation of polar contacts at the interface. Finally, the changes in the three hydrogen-bonding energies (∆hbond_lr_bb, ∆hbond_bb_sc, and ∆hbond_sc) reflect the formation of backbone−backbone, backbone−side-chain, and side-chain−side-chain hydrogen bonds at the interface.

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

<table>
<thead>
<tr>
<th>I80 atoms</th>
<th>CB</th>
<th>CG1</th>
<th>CG2</th>
<th>CD1</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>CA</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.004</td>
</tr>
<tr>
<td>C</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.008</td>
</tr>
<tr>
<td>O</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.010</td>
</tr>
<tr>
<td>CB</td>
<td>0.000</td>
<td>0.054</td>
<td>0.000</td>
<td>-0.002</td>
</tr>
<tr>
<td>OG1 → CG1</td>
<td>0.008</td>
<td>-0.054</td>
<td>-0.316</td>
<td>-0.398</td>
</tr>
<tr>
<td>CG2 → CG2’</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.020</td>
</tr>
</tbody>
</table>

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.

**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). 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 Cβ atom less than 8.0 Å away from the Cβ 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 (∆fa_sol) is positive (unfavorable), indicating that polar residues are buried upon interface formation. Balancing the desolvation penalty, the changes in polar solvation energy (∆lk_ball_wtd) and electrostatics (∆fa_elec) are negative because of the formation of polar contacts at the interface. Finally, the changes in the three hydrogen-bonding energies (∆hbond_lr_bb, ∆hbond_bb_sc, and ∆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...
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 adapted the
RNA structures. and torsional potentials important for predicting and designing during side-chain rotamer packing. This model can predict the inclusion of both protonated and deprotonated side chains.

Watson developed a new polar solvation model that evaluates the hydrogen-bonding potential to capture planar hydrogen bonds between protein side chains and nucleic acid bases. The hydrogen-bonding potential to capture planar hydrogen bonds between protein side chains and nucleic acid bases.

Recently, 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).

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

<table>
<thead>
<tr>
<th>context</th>
<th>term</th>
<th>description</th>
<th>unit</th>
<th>ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>membrane</td>
<td>fa_mpso1v</td>
<td>solvation energy dependent on the protein orientation relative to the membrane</td>
<td>kcal/mol</td>
<td>118, 121</td>
</tr>
<tr>
<td></td>
<td>fa_mpso2v</td>
<td>one-body membrane environment energy dependent on the protein orientation relative to the membrane</td>
<td>kcal/mol</td>
<td>118, 121</td>
</tr>
<tr>
<td>pH</td>
<td>e_pH</td>
<td>likelihood of side-chain protonation given a user-specified pH kcal/mol</td>
<td>117</td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Energy Terms for Structure Prediction in Different Contexts
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**

*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**


**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 SGI 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**


Hydrogen Bonding Improves Structure Prediction with Rosetta.


# Supporting Information

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


<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major changes to the Rosetta energy function since 2000</td>
<td>2</td>
</tr>
<tr>
<td>Data for calibrating Rosetta energies to kcal/mol</td>
<td>3</td>
</tr>
<tr>
<td>Additional energy function details</td>
<td>4</td>
</tr>
<tr>
<td>Parameters for the Lennard-Jones and Lazaridis-Karplus energies</td>
<td>4</td>
</tr>
<tr>
<td>Analytical form of the hydrogen bonding potential</td>
<td>8</td>
</tr>
<tr>
<td>Analytical form of the disulfide bonding potential</td>
<td>9</td>
</tr>
<tr>
<td>Statistical potentials: Interpolation of energies rather than probabilities</td>
<td>10</td>
</tr>
<tr>
<td>Methods for energy-based analysis examples</td>
<td>12</td>
</tr>
<tr>
<td>Energy terms for biomolecules other than proteins</td>
<td>14</td>
</tr>
<tr>
<td>Compatibility with D-amino acids</td>
<td>14</td>
</tr>
<tr>
<td>Energy terms for non-canonical amino acids</td>
<td>15</td>
</tr>
<tr>
<td>Energy terms for carbohydrates</td>
<td>16</td>
</tr>
<tr>
<td>Energy terms for nucleic acids</td>
<td>18</td>
</tr>
<tr>
<td>References</td>
<td>21</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Energy Term</th>
<th>Adjustment</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lennard-Jones</td>
<td>Soften repulsive potential</td>
<td>Kuhlman et al. 2000⁷</td>
</tr>
<tr>
<td></td>
<td>Atomic radii matched to crystal structures</td>
<td>Kuhlman et al. 2003⁳</td>
</tr>
<tr>
<td></td>
<td>Shifted LJ Potential</td>
<td>Meiler &amp; Baker, 2006⁴</td>
</tr>
<tr>
<td></td>
<td>Extra soft repulsive potential</td>
<td>Scheffler 2006, Unpublished</td>
</tr>
<tr>
<td></td>
<td>Make derivatives continuous</td>
<td></td>
</tr>
<tr>
<td></td>
<td>New well-depth parameters</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Incorporation of hydrogens in the ( f_{a_atr} ) calculation</td>
<td></td>
</tr>
<tr>
<td>Solvation</td>
<td>Implementation of Lazaridis-Karplus Model</td>
<td>Kuhlman et al. 2000⁶</td>
</tr>
<tr>
<td></td>
<td>Make derivatives continuous</td>
<td>Sheffler 2006, Unpublished</td>
</tr>
<tr>
<td></td>
<td>Anisotropic Solvation Model</td>
<td>Yanover et al. 2011⁷</td>
</tr>
<tr>
<td></td>
<td>New atomic volume ( \lambda ) and ( \Delta G^{\text{free}} ) parameters</td>
<td>Park et al. 2016⁵</td>
</tr>
<tr>
<td>Electrostatics</td>
<td>Knowledge-based Pair term</td>
<td>Kuhlman et al. 2000⁶</td>
</tr>
<tr>
<td></td>
<td>Coulomb electrostatics for ligand interactions</td>
<td>Meiler &amp; Baker, 2006⁴</td>
</tr>
<tr>
<td></td>
<td>Coulomb electrostatics for nucleic acids</td>
<td>Yanover et al. 2011⁷</td>
</tr>
<tr>
<td></td>
<td>Coulomb electrostatics for proteins</td>
<td>O'Meara et al. 2015⁸</td>
</tr>
<tr>
<td></td>
<td>Sigmoid dielectric model</td>
<td>Park et al. 2016⁵</td>
</tr>
<tr>
<td>Hydrogen Bonding</td>
<td>Orientation-dependent hydrogen bond potential</td>
<td>Kortemme et al. 2003⁹</td>
</tr>
<tr>
<td></td>
<td>Favoring H-bonds in the ( \text{sp}^2 ) plane</td>
<td>O'Meara et al. 2015⁸</td>
</tr>
<tr>
<td></td>
<td>No H-bond environment dependence</td>
<td>Park et al. 2016⁵</td>
</tr>
<tr>
<td>Dunbrack Rotamers</td>
<td>Add 2002 backbone-dependent rotamer library</td>
<td>Dunbrack et al. 2002⁴</td>
</tr>
<tr>
<td></td>
<td>Replace 2002 version with the 2010 smoothed rotamer library</td>
<td>Shapovalov et al. 2011¹¹</td>
</tr>
<tr>
<td>Ramachandran &amp; ( p_{_aa_pp} )</td>
<td>Interpolation with bicubic splines</td>
<td>Leaver-Fay et al. 2013¹²</td>
</tr>
<tr>
<td></td>
<td>Correction for pre-proline backbone torsions</td>
<td>Park et al. 2016⁵</td>
</tr>
<tr>
<td>Side-chain specific</td>
<td>Penalty for Tyr hydroxyl hydrogen leaving aromatic plane</td>
<td>O'Meara et al. 2015⁸</td>
</tr>
<tr>
<td>Design reference</td>
<td>Refit reference energies with OptE</td>
<td>Leaver-Fay et al. 2011¹²</td>
</tr>
<tr>
<td>energy</td>
<td>Refit reference energies with DualOptE</td>
<td>O'Meara et al. 2015⁸</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Park et al. 2016⁵</td>
</tr>
</tbody>
</table>
Table S2: Rosetta revisions corresponding to major energy function updates

<table>
<thead>
<tr>
<th>Version</th>
<th>Rosetta Revision(^a)</th>
<th>Public Version(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score12</td>
<td>pre-#55611</td>
<td>Pre-Rosetta 3.5</td>
</tr>
<tr>
<td>Talaris2013</td>
<td>#55611</td>
<td>Rosetta 3.5</td>
</tr>
<tr>
<td>Talaris2014</td>
<td>#58602</td>
<td>v2016.13-dev58602</td>
</tr>
<tr>
<td>REF2015</td>
<td>#59248</td>
<td>v2017.05-dev59248</td>
</tr>
</tbody>
</table>

\(^a\) 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 ($\Delta H_{\text{vap}}$) 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](image)

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\textsuperscript{14,15} and Lazaridis-Karplus\textsuperscript{16} potentials. The updated parameters are in the following Rosetta database files:

**Table S3: Location of Rosetta atom type parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Database File</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radii, $\Delta G_{\text{free}}, \epsilon$</td>
<td>chemical/atom_type_sets/fa_standard/atom_properties.txt</td>
</tr>
<tr>
<td>Partial charges</td>
<td>chemical/residue_type_sets/residue_types/l-caa/*.params</td>
</tr>
<tr>
<td>lk_ball weights</td>
<td>chemical/atom_type_sets/fa_standard/extras/lk_ball_wtd.txt</td>
</tr>
<tr>
<td>hbond parameters</td>
<td>scoring/score_functions/hbonds/*</td>
</tr>
<tr>
<td>rama</td>
<td>scoring/score_functions/rama/*</td>
</tr>
<tr>
<td>p_aa_pp</td>
<td>scoring/score_functions/P_AA_pp/*</td>
</tr>
<tr>
<td>omega</td>
<td>scoring/score_functions/omega/*</td>
</tr>
<tr>
<td>fa_dun</td>
<td>rotamer/*</td>
</tr>
</tbody>
</table>

**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*

<table>
<thead>
<tr>
<th>Atom Type</th>
<th>Neria et al. (^{17}) Radius (Å)</th>
<th>Talaris2014 Radius (Å)</th>
<th>REF2015 (^{5}) Radius (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA_{bb}</td>
<td>2.3650</td>
<td>2.0000</td>
<td>2.0112</td>
</tr>
<tr>
<td>CH1</td>
<td>2.3650</td>
<td>2.0000</td>
<td>2.0112</td>
</tr>
<tr>
<td>CH2</td>
<td>2.2350</td>
<td>2.0000</td>
<td>2.0112</td>
</tr>
<tr>
<td>CH3</td>
<td>2.1650</td>
<td>2.0000</td>
<td>2.0112</td>
</tr>
<tr>
<td>CNH2</td>
<td>--</td>
<td>2.0000</td>
<td>1.9922</td>
</tr>
<tr>
<td>COO</td>
<td>2.1000</td>
<td>2.0000</td>
<td>1.9649</td>
</tr>
<tr>
<td>CO_{bb}</td>
<td>--</td>
<td>2.0000</td>
<td>1.9649</td>
</tr>
<tr>
<td>aroC</td>
<td>2.1000</td>
<td>2.0000</td>
<td>1.9859</td>
</tr>
<tr>
<td>NH2O</td>
<td>--</td>
<td>1.7500</td>
<td>1.7632</td>
</tr>
<tr>
<td>Narg</td>
<td>1.6000</td>
<td>1.7500</td>
<td>1.7632</td>
</tr>
<tr>
<td>Nbb</td>
<td>1.6000</td>
<td>1.7500</td>
<td>1.7632</td>
</tr>
<tr>
<td>Nhis</td>
<td>1.6000</td>
<td>1.7500</td>
<td>1.7632</td>
</tr>
<tr>
<td>Nlys</td>
<td>1.6000</td>
<td>1.7500</td>
<td>1.7632</td>
</tr>
<tr>
<td>Npro</td>
<td>1.6000</td>
<td>1.7500</td>
<td>1.7632</td>
</tr>
<tr>
<td>Ntrp</td>
<td>1.6000</td>
<td>1.7500</td>
<td>1.7632</td>
</tr>
<tr>
<td>OC_{bb}</td>
<td>1.6000</td>
<td>1.5500</td>
<td>1.5268</td>
</tr>
<tr>
<td>OH</td>
<td>1.6000</td>
<td>1.5500</td>
<td>1.5354</td>
</tr>
<tr>
<td>ONH2</td>
<td>--</td>
<td>1.5500</td>
<td>1.5760</td>
</tr>
<tr>
<td>OOC</td>
<td>--</td>
<td>1.5500</td>
<td>1.4492</td>
</tr>
<tr>
<td>S</td>
<td>0.0430</td>
<td>1.9000</td>
<td>2.0171</td>
</tr>
<tr>
<td>HN_{bb}</td>
<td>--</td>
<td>1.0000</td>
<td>0.8773</td>
</tr>
<tr>
<td>Hapo</td>
<td>--</td>
<td>1.2000</td>
<td>1.4634</td>
</tr>
<tr>
<td>Haro</td>
<td>--</td>
<td>1.2000</td>
<td>1.3778</td>
</tr>
<tr>
<td>Hpol</td>
<td>0.8000</td>
<td>1.0000</td>
<td>0.8773</td>
</tr>
</tbody>
</table>
Table S5: Well-depth parameters from the *Neria* et al. force field, *Talaris2014*, and *REF2015*

<table>
<thead>
<tr>
<th>Atom Type</th>
<th>Neria <em>et al.</em>$^{17}$ $\epsilon$ (kcal/mol)</th>
<th>Talaris2014 $\epsilon$ (kcal/mol)</th>
<th>REF2015$^{5}$ $\epsilon$ (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAbb</td>
<td>0.0486</td>
<td>0.0486</td>
<td>0.0626</td>
</tr>
<tr>
<td>CH1</td>
<td>0.0486</td>
<td>0.0486</td>
<td>0.0626</td>
</tr>
<tr>
<td>CH2</td>
<td>0.1142</td>
<td>0.1142</td>
<td>0.0626</td>
</tr>
<tr>
<td>CH3</td>
<td>0.1811</td>
<td>0.1811</td>
<td>0.0626</td>
</tr>
<tr>
<td>CNH2</td>
<td>--</td>
<td>0.1200</td>
<td>0.0626</td>
</tr>
<tr>
<td>COO</td>
<td>0.1200</td>
<td>0.1200</td>
<td>0.0946</td>
</tr>
<tr>
<td>CObb</td>
<td>--</td>
<td>0.1400</td>
<td>0.1418</td>
</tr>
<tr>
<td>aroC</td>
<td>0.1200</td>
<td>0.1200</td>
<td>0.1418</td>
</tr>
<tr>
<td>NH2O</td>
<td>--</td>
<td>0.2834</td>
<td>0.0688</td>
</tr>
<tr>
<td>Narg</td>
<td>0.2384</td>
<td>0.2834</td>
<td>0.1617</td>
</tr>
<tr>
<td>Nbb</td>
<td>0.2384</td>
<td>0.2834</td>
<td>0.1617</td>
</tr>
<tr>
<td>Nhis</td>
<td>0.2384</td>
<td>0.2834</td>
<td>0.1617</td>
</tr>
<tr>
<td>Nlys</td>
<td>0.2384</td>
<td>0.2834</td>
<td>0.1617</td>
</tr>
<tr>
<td>Npro</td>
<td>0.2384</td>
<td>0.2834</td>
<td>0.1617</td>
</tr>
<tr>
<td>Ntrp</td>
<td>0.2384</td>
<td>0.2834</td>
<td>0.1617</td>
</tr>
<tr>
<td>OCbb</td>
<td>--</td>
<td>0.1591</td>
<td>0.1617</td>
</tr>
<tr>
<td>OH</td>
<td>0.1591</td>
<td>0.1591</td>
<td>0.1617</td>
</tr>
<tr>
<td>ONH2</td>
<td>--</td>
<td>0.1591</td>
<td>0.1424</td>
</tr>
<tr>
<td>OOC</td>
<td>--</td>
<td>0.2100</td>
<td>0.1619</td>
</tr>
<tr>
<td>S</td>
<td>0.0430</td>
<td>0.1600</td>
<td>0.1829</td>
</tr>
<tr>
<td>SH1</td>
<td>--</td>
<td>--</td>
<td>0.0999</td>
</tr>
<tr>
<td>HNbb</td>
<td>--</td>
<td>0.0500</td>
<td>0.4560</td>
</tr>
<tr>
<td>HS</td>
<td>--</td>
<td>--</td>
<td>0.4560</td>
</tr>
<tr>
<td>Hapo</td>
<td>--</td>
<td>0.0500</td>
<td>0.0050</td>
</tr>
<tr>
<td>Haro</td>
<td>--</td>
<td>0.0500</td>
<td>0.0508</td>
</tr>
<tr>
<td>Hpol</td>
<td>--</td>
<td>0.0500</td>
<td>0.0218</td>
</tr>
<tr>
<td>Atom Type</td>
<td>Lazaris-Karplus&lt;sup&gt;16&lt;/sup&gt; $\Delta G_{\text{free}}$ (kcal/mol)</td>
<td>Talaris2014, $\Delta G_{\text{free}}$ (kcal/mol)</td>
<td>REF2015&lt;sup&gt;5&lt;/sup&gt; $\Delta G_{\text{free}}$ (kcal/mol)</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>CAbb</td>
<td>-0.2500</td>
<td>1.0000</td>
<td>2.5338</td>
</tr>
<tr>
<td>CH0</td>
<td>-0.2500</td>
<td>-0.2500</td>
<td>1.4093</td>
</tr>
<tr>
<td>CH1</td>
<td>-0.2500</td>
<td>-0.2500</td>
<td>-3.5384</td>
</tr>
<tr>
<td>CH2</td>
<td>0.5200</td>
<td>0.5200</td>
<td>-1.8547</td>
</tr>
<tr>
<td>CH3</td>
<td>1.5000</td>
<td>1.5000</td>
<td>7.2929</td>
</tr>
<tr>
<td>CNH2</td>
<td>--</td>
<td>0.0000</td>
<td>3.0770</td>
</tr>
<tr>
<td>COO</td>
<td>0.1200</td>
<td>-1.4000</td>
<td>-3.3326</td>
</tr>
<tr>
<td>CObb</td>
<td>--</td>
<td>1.0000</td>
<td>3.1042</td>
</tr>
<tr>
<td>aroC</td>
<td>0.8000</td>
<td>0.0800</td>
<td>1.7979</td>
</tr>
<tr>
<td>NH2O</td>
<td>--</td>
<td>-7.8000</td>
<td>-8.1016</td>
</tr>
<tr>
<td>Narg</td>
<td>-10.0000</td>
<td>-10.0000</td>
<td>-8.9684</td>
</tr>
<tr>
<td>Nbb</td>
<td>-7.8000</td>
<td>-5.0000</td>
<td>-9.9695</td>
</tr>
<tr>
<td>Nhis</td>
<td>-4.0000</td>
<td>-4.0000</td>
<td>-9.7396</td>
</tr>
<tr>
<td>Nlys</td>
<td>-20.000</td>
<td>-20.000</td>
<td>-20.865</td>
</tr>
<tr>
<td>Npro</td>
<td>-1.5500</td>
<td>-1.5500</td>
<td>-0.9846</td>
</tr>
<tr>
<td>Ntrp</td>
<td>-8.9000</td>
<td>-8.9000</td>
<td>-8.4131</td>
</tr>
<tr>
<td>OCbb</td>
<td>-10.0000</td>
<td>-5.0000</td>
<td>-8.0068</td>
</tr>
<tr>
<td>OH</td>
<td>-6.7000</td>
<td>-6.7000</td>
<td>-8.1335</td>
</tr>
<tr>
<td>ONH2</td>
<td>--</td>
<td>-5.8500</td>
<td>-6.5916</td>
</tr>
<tr>
<td>OOC</td>
<td>--</td>
<td>-10.0000</td>
<td>-9.2398</td>
</tr>
<tr>
<td>S</td>
<td>-4.1000</td>
<td>-4.1000</td>
<td>-1.7072</td>
</tr>
<tr>
<td>SH1</td>
<td>-2.7000</td>
<td>--</td>
<td>3.2916</td>
</tr>
</tbody>
</table>
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.

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

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

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

The third component, $E_{\text{hbond}}^{BzBAH}(\rho, \phi_{BzBAH}, \theta_{BAH})$, is dependent on the hybridization of the acceptor, $\rho$. 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.

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

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

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

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.

$$E_{\text{hbond}}^{BzBAH}(\rho, \phi_{BzBAH}, \theta_{BAH}) = \begin{cases} 
H(\phi_{BzBAH})F(\phi_{BzBAH}) + \left(1 - H(\phi_{BzBAH})G(\phi_{BzBAH})\right) \quad & \rho \sim \text{sp}^2 \\
\frac{1}{2} f_{\text{poly}}(\cos(\theta_{BAH})) + \frac{1}{4} \left(1 + \cos(\phi_{BzBAH})\right) \quad & \rho \sim \text{sp}^3 \\
f_{\text{poly}}(\cos(\theta_{BAH})) \quad & \rho \sim \text{ring} 
\end{cases} \quad \text{(S7)}$$
Figure S2: Analytic form of the hydrogen bonding BA\(\chi\) potential for sp\(2\) 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 BA\(\chi\) torsion, \(\phi\).

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 \(E_{\text{dslf}}^{\text{SS}}(d_{\text{ss}})\) is defined by Eq. S8 given the sulfur-sulfur distance, \(d\), mean distance \(\overline{d}_{\text{ss}}\), standard deviation \(\sigma_{\text{ss}}\), and fitting parameters \(\alpha_{\text{dslf}}^{\text{SS}}\), \(\epsilon_m\) and \(w_{\text{SS}}\).

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

Next, the energy of the angle formed by a \(C_\beta\) and two sulfur atoms \(E_{\text{dslf}}^{\text{CSS}}(\theta_{\text{CSS}})\) is defined by Eq. S9 given the angle \(\theta\) and von Mises parameters \(A_{\text{CSS}}, w_{\text{CSS}}, \kappa_{\text{CSS}}, \text{ and } \mu_{\text{CSS}}\).

\[
E_{\text{dslf}}^{\text{CSS}}(\theta_{\text{CSS}}) = w_{\text{CSS}} \left( -\ln(A_{\text{CSS}}) - \kappa_{\text{CSS}} \cos(\theta - \mu_{\text{CSS}}) \right)
\]  
(S9)

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

\[
E_{\text{dslf}}^{\text{CSSC}_\beta}(\phi_{\text{CSS}}) = w_{\text{CSSC}_\beta} \ln \left( \Sigma_{k=2} \exp \left( A_{k,\text{CSSC}_\beta} + \kappa_{k,\text{CSSC}_\beta} \cos \left( \phi - \mu_{k,\text{CSSC}_\beta} \right) \right) + \epsilon_m \right)
\]  
(S10)

Finally, the energy of the torsion formed by \(C_\alpha, C_\beta\) and the two adjacent sulfur atoms \(E_{\text{dslf}}^{\text{CSSC}_\alpha}(\theta_{\text{CSS}})\) is defined by Eq. S11 given the torsion angle \(\phi\) and the von Mises fitting parameters \(A_{k,\text{CSSC}_\alpha}, \kappa_{k,\text{CSSC}_\alpha}, \mu_{k,\text{CSSC}_\alpha}\) and \(\epsilon_m\).

\[
E_{\text{dslf}}^{\text{CSSC}_\alpha}(\theta_{\text{CSS}}) = -w_{\text{CSSC}_\alpha} \ln \left( \Sigma_{l=3} \exp \left( A_{k,\text{CSSC}_\alpha} + \kappa_{k,\text{CSSC}_\alpha} \cos \left( \theta - \mu_{k,\text{CSSC}_\alpha} \right) \right) + \epsilon_m \right)
\]  
(S11)
**Statistical potentials: Interpolation of energies rather than probabilities**

The Rosetta energy function uses probabilities from the Dunbrack backbone-dependent rotamer library\textsuperscript{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 $\phi, \psi$-dependent terms were stored on a $10^\circ \times 10^\circ$ 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 $\psi$. This method resulted in large inaccuracies because $P$ can vary by orders of magnitude over very short ranges of $\phi$ and $\psi$. 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 $\psi$ and $I_0(\kappa)$ is the modified Bessel function of order zero needed to normalize the distribution.

$$ P(x) = \frac{\exp(\kappa - \cos(x - \mu))}{2\pi I_0(\kappa)} \quad \text{(S13)} $$

First, **Figure S3A** shows the probability distribution, $P$ and its linear interpolation based on the $10^\circ \times 10^\circ$ 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\(^\text{19}\) 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.\(^\text{20}\)

```
```

The energies were averaged for each ensemble of five structures. The ΔΔG 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\(^\text{21}\) 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_file_name`.

**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:

```bash
relax.linuxgccrelease -s 1ztx_unbound.pdb -relax:restrain:consriants 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.\textsuperscript{22} using the *REF2015* energy function.

```bash
docking_prepack_protocol.linux -s 1ztx_relaxed.pdb -partners LH_G \
-dock_atmin -docking:sc_min

docking_protocol.linuxgccrelease -s 1ztx_unbound_prepacked.pdb -native \
1ztx_native.pdb -ignore_unrecognized_res -ex2ar -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

<table>
<thead>
<tr>
<th>Term</th>
<th>Can score</th>
</tr>
</thead>
<tbody>
<tr>
<td>fa_atr</td>
<td>All molecules</td>
</tr>
<tr>
<td>fa_rep</td>
<td>All molecules</td>
</tr>
<tr>
<td>fa_intra_rep</td>
<td>All molecules</td>
</tr>
<tr>
<td>fa_sol</td>
<td>All molecules</td>
</tr>
<tr>
<td>lk_ball</td>
<td>All molecules</td>
</tr>
<tr>
<td>fa_intra_sol</td>
<td>All molecules</td>
</tr>
<tr>
<td>fa_elec</td>
<td>All molecules</td>
</tr>
<tr>
<td>hbond_sr_bb</td>
<td>All molecules</td>
</tr>
<tr>
<td>hbond_lr_bb</td>
<td>All molecules</td>
</tr>
<tr>
<td>hbond_bb_sc</td>
<td>All molecules</td>
</tr>
<tr>
<td>hbond_sc</td>
<td>All molecules</td>
</tr>
<tr>
<td>dsfl_fa13</td>
<td>L-, D-, and mixed D/L disulfide bonds between cysteine or cysteine-like residues (e.g., homocysteine, penacillamine)</td>
</tr>
<tr>
<td>rama_prepro</td>
<td>Glycine, canonical L-amino acids, their D-counterparts, and similar alpha-amino acids that can use canonical rama tables.</td>
</tr>
<tr>
<td>p_aa_pp</td>
<td>Glycine, canonical L-amino acids, their D-counterparts, and similar alpha-amino acids that can use canonical rama tables.</td>
</tr>
<tr>
<td>omega</td>
<td>All α-amino acids, or β-amino acids.</td>
</tr>
<tr>
<td>fa_dun</td>
<td>All polymer building blocks.</td>
</tr>
<tr>
<td>pro_close</td>
<td>L- and D-proline.</td>
</tr>
<tr>
<td>yhh_planarity</td>
<td>L- and D-tyrosine.</td>
</tr>
<tr>
<td>ref</td>
<td>Glycine, canonical L-amino acids, and their D-counterparts.</td>
</tr>
</tbody>
</table>

Compatibility with D-amino acids

To make the energy terms compatible with D-amino acids, several modifications were made to the torsional terms. First, the \( \phi, \psi \) values were inverted in the \texttt{rama_prepro}, \texttt{omega}, and \texttt{p_aa_pp} terms to accommodate the chirality of the backbone. Accordingly, the derivatives were inverted to ensure that mirror-image structures energy-minimize identically. Second, the \texttt{fa_dun} score term was modified
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),\textsuperscript{24} hydrogen bond surrogates (HBS), oligo-peptoids,\textsuperscript{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, $\sigma_{ij}$, and the geometric mean of atomic well-depths, $\epsilon_{ij}$, are based on the CHARMM 24\textsuperscript{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 $K_\phi$ and $n$ come from CHARMM 24. A given set of 4 atoms types may have multiple $K_\phi$ and $n$ parameters that are summed in a Fourier series to more accurately describe the rotation about the central bond of the torsion.

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

Explicit Unfolded State Energy (EUSE) represents the unfolded energy of the protein and compensates for the difficulty 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 $U(AA_r, t)$ 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.\textsuperscript{27}
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 \(\alpha\)-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 \(R_{1B}(AA_r, t1b)\) 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 \(R_{2B}(T_i, t2b)\) 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.

\[
E_{TCRE} = (\sum_r \sum_{t1b} W_{t1b} R_{1B}(AA_r, t1b)) + (\sum_i \sum_{t2b} W_{t2b} R_{2B}(T_i, t2b))
\] (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 \texttt{sugar\_bb} term which rewards preferred glycosidic torsion angles.\(^{29}\) The \texttt{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).

\[
E_{\text{sugar\_bb}} = d + \sum_i a_i e^{-(x-b)^2/c_i}
\] (S17)

For \(\omega\) 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 \(\theta\), and the energy difference relative to the minimum \(b\) (Eq. S18). This function approximates the so-called Gauche effect.

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

The \texttt{sugar\_bb} score per residue is the sum of each function for each glycosidic torsion in the residue. \textbf{Table S8} lists the functional form for each torsion and linkage type. (The functions assume that \(D\)-sugars are in the \(^4\)C\(_1\) chair conformation and that \(L\)-sugars are in the \(^4\)C\(_1\) chair conformation.)
### Table S8: Functional form of the sugar backbone energy for each torsion and linkage type

<table>
<thead>
<tr>
<th>Angle</th>
<th>Ax./eq. designation</th>
<th>Stereoisomer</th>
<th>Exocyclic</th>
<th>Range</th>
<th>Functional form</th>
</tr>
</thead>
<tbody>
<tr>
<td>φ</td>
<td>axial (α)</td>
<td>D</td>
<td>—</td>
<td>–180°–180°</td>
<td>Gaussian</td>
</tr>
<tr>
<td></td>
<td>equatorial (β)</td>
<td>D</td>
<td>—</td>
<td>–180°–180°</td>
<td>Gaussian</td>
</tr>
<tr>
<td></td>
<td>axial (α)</td>
<td>L</td>
<td>—</td>
<td>–180°–180°</td>
<td>Gaussian, x=–φ</td>
</tr>
<tr>
<td></td>
<td>equatorial (β)</td>
<td>L</td>
<td>—</td>
<td>–180°–180°</td>
<td>Gaussian, x=–φ</td>
</tr>
<tr>
<td></td>
<td>ax. (parent at odd O)</td>
<td>D (parent)</td>
<td>no</td>
<td>0–360°</td>
<td>Gaussian</td>
</tr>
<tr>
<td></td>
<td>eq. (parent at odd O)</td>
<td>D (parent)</td>
<td>no</td>
<td>0–360°</td>
<td>Gaussian</td>
</tr>
<tr>
<td></td>
<td>ax. (parent at even O)</td>
<td>D (parent)</td>
<td>no</td>
<td>0–360°</td>
<td>Gaussian</td>
</tr>
<tr>
<td></td>
<td>eq. (parent at even O)</td>
<td>D (parent)</td>
<td>no</td>
<td>0–360°</td>
<td>Gaussian</td>
</tr>
<tr>
<td></td>
<td>ax. (parent at odd O)</td>
<td>L (parent)</td>
<td>no</td>
<td>0–360°</td>
<td>Gaussian, 360°–ψ</td>
</tr>
<tr>
<td></td>
<td>eq. (parent at odd O)</td>
<td>L (parent)</td>
<td>no</td>
<td>0–360°</td>
<td>Gaussian, 360°–ψ</td>
</tr>
<tr>
<td></td>
<td>ax. (parent at even O)</td>
<td>L (parent)</td>
<td>no</td>
<td>0–360°</td>
<td>Gaussian, 360°–ψ</td>
</tr>
<tr>
<td></td>
<td>eq. (parent at even O)</td>
<td>L (parent)</td>
<td>no</td>
<td>0–360°</td>
<td>Gaussian, 360°–ψ</td>
</tr>
<tr>
<td></td>
<td>axial (α)</td>
<td>D</td>
<td>yes</td>
<td>0–360°</td>
<td>Gaussian</td>
</tr>
<tr>
<td></td>
<td>equatorial (β)</td>
<td>D</td>
<td>yes</td>
<td>0–360°</td>
<td>Gaussian</td>
</tr>
<tr>
<td></td>
<td>axial (α)</td>
<td>L</td>
<td>yes</td>
<td>0–360°</td>
<td>Gaussian, 360°–ψ</td>
</tr>
<tr>
<td></td>
<td>equatorial (β)</td>
<td>L</td>
<td>yes</td>
<td>0–360°</td>
<td>Gaussian, 360°–ψ</td>
</tr>
<tr>
<td>ω</td>
<td>axial (parent O4)</td>
<td>D (parent)</td>
<td>yes</td>
<td>0–120°</td>
<td>parabolic</td>
</tr>
<tr>
<td></td>
<td>axial (parent O4)</td>
<td>D (parent)</td>
<td>yes</td>
<td>120°–240°</td>
<td>parabolic</td>
</tr>
<tr>
<td></td>
<td>axial (parent O4)</td>
<td>D (parent)</td>
<td>yes</td>
<td>240°–360°</td>
<td>parabolic</td>
</tr>
<tr>
<td></td>
<td>eq. (parent O4)</td>
<td>D (parent)</td>
<td>yes</td>
<td>0–120°</td>
<td>parabolic</td>
</tr>
<tr>
<td></td>
<td>eq. (parent O4)</td>
<td>D (parent)</td>
<td>yes</td>
<td>120°–240°</td>
<td>parabolic</td>
</tr>
<tr>
<td></td>
<td>eq. (parent O4)</td>
<td>D (parent)</td>
<td>yes</td>
<td>240°–360°</td>
<td>parabolic</td>
</tr>
<tr>
<td></td>
<td>axial (parent O4)</td>
<td>L (parent)</td>
<td>yes</td>
<td>0–360°</td>
<td>parabolic, 360°–ω</td>
</tr>
<tr>
<td></td>
<td>axial (parent O4)</td>
<td>L (parent)</td>
<td>yes</td>
<td>120°–240°</td>
<td>parabolic, 360°–ω</td>
</tr>
<tr>
<td></td>
<td>axial (parent O4)</td>
<td>L (parent)</td>
<td>yes</td>
<td>240°–360°</td>
<td>parabolic, 360°–ω</td>
</tr>
<tr>
<td></td>
<td>eq. (parent O4)</td>
<td>L (parent)</td>
<td>yes</td>
<td>0–120°</td>
<td>parabolic, 360°–ω</td>
</tr>
<tr>
<td></td>
<td>eq. (parent O4)</td>
<td>L (parent)</td>
<td>yes</td>
<td>120°–240°</td>
<td>parabolic, 360°–ω</td>
</tr>
<tr>
<td></td>
<td>eq. (parent O4)</td>
<td>L (parent)</td>
<td>yes</td>
<td>240°–360°</td>
<td>parabolic, 360°–ω</td>
</tr>
</tbody>
</table>
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. This term scales the $fa_{elec}$ potential as a function of the angle ($\kappa$) between the normal to the plane of the base ($z_i$) and the distance vector 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.

$$E_{stack\_elec} = \sum_{r_1 < r_2} \sum_{i,j} f(\kappa_i, \kappa_j) E_{fa_{elec}}$$ (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).

$$f(\kappa_i, \kappa_j) = \cos^2(\kappa_i) + \cos^2(\kappa_j)$$ (S20)

**Base stacking.** $\pi - \pi$ stacking interactions are not explicitly captured by $fa_{atr}$; thus, Rosetta includes an additional stacking bonus term, called $fa_{stack}$. 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 ($\kappa_i$) between the normal to the plane of the base ($z_i$) and the distance vector from atoms $i$ to $j$ ($d_{ij}$), 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.
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: $\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \nu_1, \nu_2, \chi, O2'$. 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).

$$E_{\text{rna_torsion}} = \sum_k -\ln (P(A_k)) \quad (S24)$$

To accommodate special cases, separate potentials were derived for each of the $\delta, \epsilon, \nu_1, \nu_2, \chi, O2'$ torsions depending on whether the sugar pucker is 2'-endo or 3'-endo. Additionally, a separate $\chi$ potential was derived for purines and pyrimidines. For $\zeta$, there are three separate potentials depending on whether the $\alpha$ 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.

$$E_{\text{geom_sol}} = \sum_{r_1<r_2} \sum_{i,j} E_{\text{hbond}}(r_i - v_j) \quad (S25)$$
Figure S5. Torsion potentials for RNA

RNA torsional potential for (A) $\alpha$, (B) $\beta$, (C) $\gamma$, (D) $\delta$, (E) $\epsilon$, (F) $\zeta$ when the $\alpha$ torsion of the following residue is gauche$^-$ (orange), trans (cyan), or gauche$^+$ (purple) (G) $\chi$ for purines (lighter red and blue) and pyrimidines (darker red and blue), (H) $\nu_1$, (I) $\nu_2$. Potentials when the sugar pucker is C2'-endo are shown in red and C3'-endo shown in blue.
References


