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

Rebecca F. Alford,[1] Andrew Leaver-Fay,[2] Jeliazko R. Jeliazkov,[3] Matthew J. O'Meara,[4] Frank P. DiMaio,[5] Hahnbeom Park,[6] Maxim V. Shapovalov,[7] P. Douglas Renfrew,[8,9] Vikram K. Mulligan,[6] Kalli Kappel,[10] Jason W. Labonte,[1] Michael S. Pacella,[11] Richard Bonneau,[8,9] Philip Bradley,[12] Roland L. Dunbrack Jr.,[7] Rhiju Das,[13] David Baker,[6] Brian Kuhlman,[2] Tanja Kortemme,[14] Jeffrey J. Gray[1,2§]

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

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

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

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

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

[6] Department of Biochemistry, University of Washington, Molecular Engineering and Sciences, 4000 15th Ave NE, Seattle, Washington 98195, United States

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

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

[9] Center for Computational Biology, Flatiron Institute, Simons Foundation, 162 5th Avenue, New York, New York 10010, United States

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

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

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

[13] Department of Biochemistry, Stanford University, B400 Beckman Center, 279 Campus Drive, Stanford, California 94305, United States

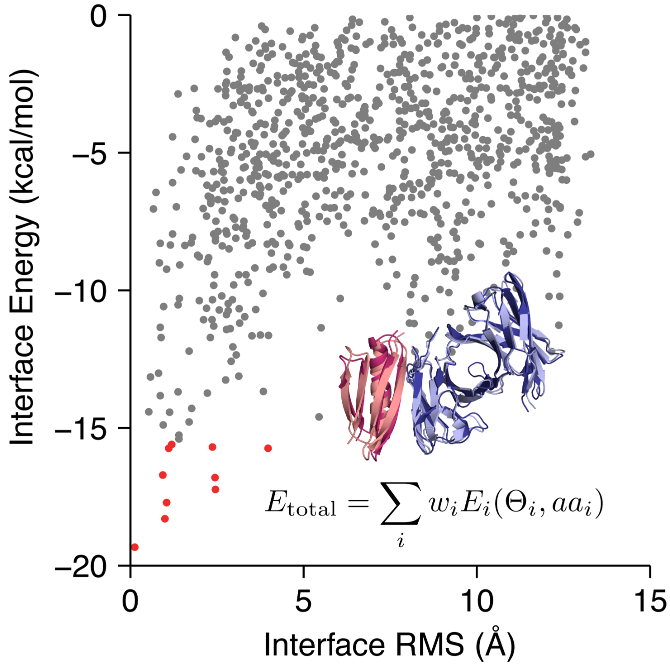
[14] Department of Bioengineering and Therapeutic Sciences, University of California at San Francisco, San Francisco, California 94158, United States

[§] Corresponding author

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 parameterized 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, *beta_nov15*. 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 capabilities from soluble proteins to also include membrane proteins, peptides containing non-canonical amino acids, carbohydrates, nucleic acids, and other macromolecules.

Table of Contents Graphic



Keywords

Rosetta, molecular modeling, energy function, force field, structure prediction, design

Introduction

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

Energy functions are based on Anfinsen's hypothesis that native-like protein conformations represent unique, low-energy, thermodynamically stable conformations.[4] These folded states reside in minima on the energy landscape, and they have a net favorable change in Gibbs free energy, which is the sum of contributions from both enthalpy (ΔH) and entropy (ΔS) 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 a linear combination of terms that compute energies as a function of various degrees of freedom.

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

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

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

The Rosetta energy function has changed dramatically since it was last described in complete detail by Rohl *et al*.[47] in 2004. It has undergone significant advances ranging from improved models of hydrogen bonding[48] and solvation,[49] to updated evaluation of backbone[50] and rotamer conformations.[51] Along the way, these developments have enabled Rosetta to address new biomolecular modeling problems including refinement of low-resolution X-ray structures,[52] development of protein binders,[53] and the design of vaccines,[54] biomineralization peptides,[55] self-assembling materials,[56] and enzymes that perform new functions.[57,58] Instead of arbitrary units, the energy function is now also calibrated to compute energies in kcal/mol. The details of these 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[59,60] and for all biomolecules including RNA,[61] DNA,[62,63] small-molecule ligands,[64] non-canonical amino acids and backbones,[65–67] and carbohydrates,[68] 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 concepts and calculations underlying the current Rosetta all-atom energy function called *beta_nov15*. Our discussion aims to expose the physical and mathematical details of the energy function required for rigorous understanding. In addition, we explain how to apply the computed energies to analyze structural models produced by Rosetta simulations. We hope this paper will provide critically needed documentation of the energy methods as well as an educational resource to help students and scientists interpret the results of these simulations.

Computing the total Rosetta energy

The Rosetta energy function approximates the energy of a biomolecule conformation. This quantity, called ΔE_{total}, is computed from a linear combination of energy terms E_i which are calculated as a function of geometric degrees of freedom, Θ, chemical identities, aa, and scaled by a weight on each term, w, as shown in **Eq. 1**.

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

Here, we explain the Rosetta energy function term by term. First, we describe energies of interactions between non-bonded atom-pairs important for atomic packing, electrostatics, and solvation. Second, we explain empirical potentials used to model hydrogen- and disulfide-bonds. Next, we explain statistical potentials used to describe backbone and side-chain torsional preferences in proteins. After, we explain a set of terms that accommodate features not explicitly captured yet important for native structural feature recapitulation. Finally, we discuss how the energy terms are combined into a single function used to approximate the energy of biomolecules. For reference, items in the `fixed width font` are names of energy terms in the Rosetta code. The energy terms are summarized in **Table 1**.

Table 1: Summary of Energy terms in the *beta_nov15* energy function

Term	Description	Weight	Units	Ref.
`fa_atr`	Attractive energy between two atoms on different residues separated by distance, *d*	1.0	kcal/mol	[5,6]
`fa_rep`	Repulsive energy between two atoms on different residues separated by distance, *d*	0.55	kcal/mol	[5,6]
`fa_intra_rep`	Repulsive energy between two atoms on the same residue, separated by distance, *d*	0.005	kcal/mol	[5,6]
`fa_sol`	Gaussian exclusion implicit solvation energy between protein atoms in different residues	1.0	kcal/mol	[36]
`lk_ball_wtd`	Orientation-dependent solvation of polar atoms assuming ideal water geometry	1.0	kcal/mol	[49,69]
`fa_intra_sol`	Gaussian exclusion implicit solvation energy between protein atoms in the same residue	1.0	kcal/mol	[36]
`fa_elec`	Energy of interaction between two non-bonded charged atoms separated by distance, *d*	1.0	kcal/mol	[49]
`hbond`	Energy of hydrogen bonds	1.0	kcal/mol	[38,48]
`dslf_fa13`	Energy of disulfide bridges	1.25	kcal/mol	[48]
`rama_prepro`	Probability of backbone ϕ, ψ angles given amino acid type	0.45 kcal/mol/kT	kT	[49,50]
`p_aa_pp`	Probability of amino acid identity given backbone ϕ, ψ angles	0.4 kcal/mol/kT	kT	[50]
`fa_dun`	Probability that a chosen rotamer is native-like given backbone ϕ, ψ angles	0.7 kcal/mol/kT	kT	[51]
`omega`	Backbone-dependent penalty for cis ω dihedrals that deviate from 0° and trans ω dihedrals that deviate from 180°	0.6 kcal/mol/AU	Arbitrary Units (AU)	[70]
`pro_close`	Penalty for an open proline ring and proline ω bonding energy	1.25 kcal/mol/AU	Arbitrary Units	[50]
`yhh_planarity`	Sinusoidal penalty for non-planar tyrosine χ_3 dihedral angle	0.625 kcal/mol/AU	Arbitrary Units	[48]
`ref`	Reference energies for amino acid types	1.0 kcal/mol/AU	Arbitrary Units	[1,50]

Terms for atom-pair interactions

van der Waals interactions are short-range attractive and repulsive forces that vary with atom-pair distance. Whereas attractive forces result from the cross-correlated motions of electrons in neighboring non-bonded atoms, repulsive forces occur because electrons cannot occupy the same orbitals by the Pauli exclusion principle. To model van der Waals interactions, Rosetta uses the Lennard-Jones (LJ) 6-12 potential[5,6] which calculates the interaction energy of atoms i and j in different residues given their summed atomic radii $\sigma_{i,j}$,[a] atom-pair distance, $d_{i,j}$, and the geometric mean of well depths, $\epsilon_{i,j}$ (**Eq. 2**). The atomic radii and well depths are derived from small molecule liquid phase data optimized in context of the energy model.[49]

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

Rosetta splits the LJ potential at the function's minimum ($d_{i,j} = \sigma_{i,j}$) into two components that can be weighted separately: attractive (`fa_atr`) and repulsive (`fa_rep`). By decomposing the function this way, we can alter component weights without changing the minimum-energy distance or introducing any derivative discontinuities. Many conformational sampling protocols in Rosetta take advantage of this splitting by slowly increasing the weight of the repulsive component to traverse rugged energy landscapes and to prevent structures from unfolding during sampling.[71]

The repulsive van der Waals energy, `fa_rep,` varies as a function of atom-pair distance. At short distances, atomic overlap results in strong forces that lead to large changes in the energy. The steep $1/d_{i,j}^{12}$ term can cause poor performance in minimization routines and overall structure prediction and design calculations.[72,73] To alleviate this problem, we weaken the repulsive component by replacing the $1/d_{i,j}^{12}$ term with a softer linear term when $d \leq 0.6\sigma_{i,j}$. The term is computed using the atom-type specific parameters $m_{i,j}$ and $b_{i,j}$ which are fit to ensure derivative continuity at $d = 0.6\sigma_{i,j}$ After the linear component, the function transitions smoothly to the 6-12 form until $d_{i,j} = \sigma$, where it reaches zero and remains zero (**Eq. 3**; **Fig. 1A**).

$$E_{\text{rep}}(i,j) = \sum\nolimits_{i,j} w_{i,j}^{\text{conn}} \begin{cases} m_{i,j}\, d_{i,j} + b_{i,j} & d_{i,j} \leq 0.6\sigma_{i,j} \\ \varepsilon_{i,j}\left[\left(\frac{\sigma_{i,j}}{d_{i,j}}\right)^{12} - 2\left(\frac{\sigma_{i,j}}{d_{i,j}}\right)^{6} + 1\right] & 0.6\sigma_{i,j} < d_{i,j} \leq \sigma_{i,j} \\ 0 & \sigma_{i,j} < d_{i,j} \end{cases} \qquad (3)$$

Rosetta also includes an intra-residue version of the repulsive component, `fa_intra_rep`, with the same functional form as the `fa_rep` term (**Eq. 3**). We include this term because the knowledge-based rotamer energy (`fa_dun`, below) under-estimates intra-residue collisions.

The attractive van der Waals energy, `fa_atr` has a value of $-\epsilon_{ij}$ when $d_{i,j} = 0$ and then transitions to the 6-12 potential as the distance increases (**Eq. 4**; **Fig. 1B**). For speed, we truncate the LJ term beyond 6.0 Å where the van der Waals forces are small. To avoid derivative discontinuities, we use a cubic

[a] In Rosetta, $\sigma_{i,j}$ has the same definition as the $r_{i,j}^{\text{min}}$ variable in CHARMM.

polynomial function, $f_{\mathrm{poly}}(d_{i,j})$ after 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 inflections points in the energy landscape during conformational sampling with gradient-based minimization (Sheffler 2006, Unpublished).

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

All three terms are multiplied by a connectivity weight $w_{i,j}^{\mathrm{conn}}$ to exclude the large repulsive energetic contributions that would otherwise be calculated for atoms separated by fewer than four chemical bonds (**Eq. 5**). This weight is common to molecular force fields that assume covalent bonds are not formed or broken during a simulation. Rosetta uses four chemical bonds as the "crossover" separation when $w_{i,j}^{\mathrm{conn}}$ transitions from zero to one (rather than the three chemical bonds used by traditional force fields) to limit the effects of double-counting due to knowledge-based torsional potentials.

$$w_{i,j}^{\mathrm{conn}} = \begin{cases} 0 & n_{i,j}^{\mathrm{bonds}} \le 3 \\ 0.2 & n_{i,j}^{\mathrm{bonds}} = 4 \\ 1 & n_{i,j}^{\mathrm{bonds}} \ge 5 \end{cases} \quad (5)$$

The comparison between **Eq. 2** and the modified LJ potential (**Eq. 3-4**) is shown in **Fig. 1A** and **Fig. 1B**.

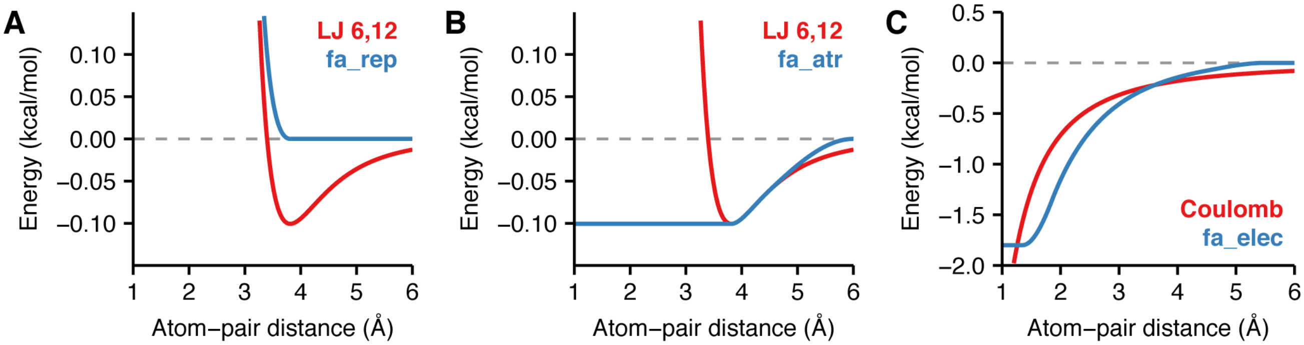


Figure 1: Van der Waals and electrostatics energies
Comparison between pairwise energies of non-bonded atoms computed by Rosetta and the form computed by traditional molecular mechanics force fields. Here, the interaction between the backbone nitrogen and carbon are used as an example. (A) Lennard-Jones van der Waals energy with well-depths $\epsilon_{\mathrm{Nbb}} = 0.162$ and $\epsilon_{\mathrm{Cbb}} = 0.063$ and atomic radii $r_{\mathrm{Nbb}} = 1.763$ and $r_{\mathrm{Cbb}} = 2.011$ (red) and Rosetta `fa_rep` (blue). (B) Lennard-Jones van der Waals energy (red) and Rosetta `fa_rep` (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 electrostatics energy with a dielectric constant $\epsilon = 10$, and partial charges $q_{\mathrm{Nbb}} = -0.604$ and $q_{\mathrm{Cbb}} = 0.090$ (red) compared with Rosetta `fa_elec` (blue). The `fa_elec` model is shifted to reach zero at the cutoff distance 6.0 Å.

Electrostatics. Non-bonded 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**).[49] Coulomb's law is a pairwise term commonly expressed in terms of the distance between atoms i and j ($d_{i,j}$), dielectric constant ϵ, partial atomic charges for each atom q_i and q_j, and Coulomb's constant, $C_0 = 322$ Å kcal/mol e^{-2} (with e being the elementary charge) (**Eq. 6**).

$$E_{\text{Coulomb}}(i,j) = \frac{C_o q_i q_j}{\epsilon} \frac{1}{d_{i,j}^2} \qquad (6)$$

To approximate electrostatic interactions in biomolecules, we modify the potential to account for the difference in dielectric constant between the protein core and solvent-exposed surface.[74] Specifically, we substitute the constant ϵ in **Eq. 6** with a sigmoidal function $\epsilon(d_{i,j})$ that increases from $\epsilon_{\text{core}} = 6$ to $\epsilon_{\text{solvent}} = 80$ when the atom-pair distance is between 0 Å and 4 Å (**Eq. 7-8**):

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

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

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

$$E_{\text{elec}}(i,j,d_{i,j}) = \frac{C_o q_i q_j}{\epsilon(d_{i,j})}\begin{cases} \frac{1}{d_{i,j}^2} - \frac{1}{d_{\max}^2} & d \le d_{\max} \\ 0 & d_{\max} < d \end{cases} \qquad (9)$$

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

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

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

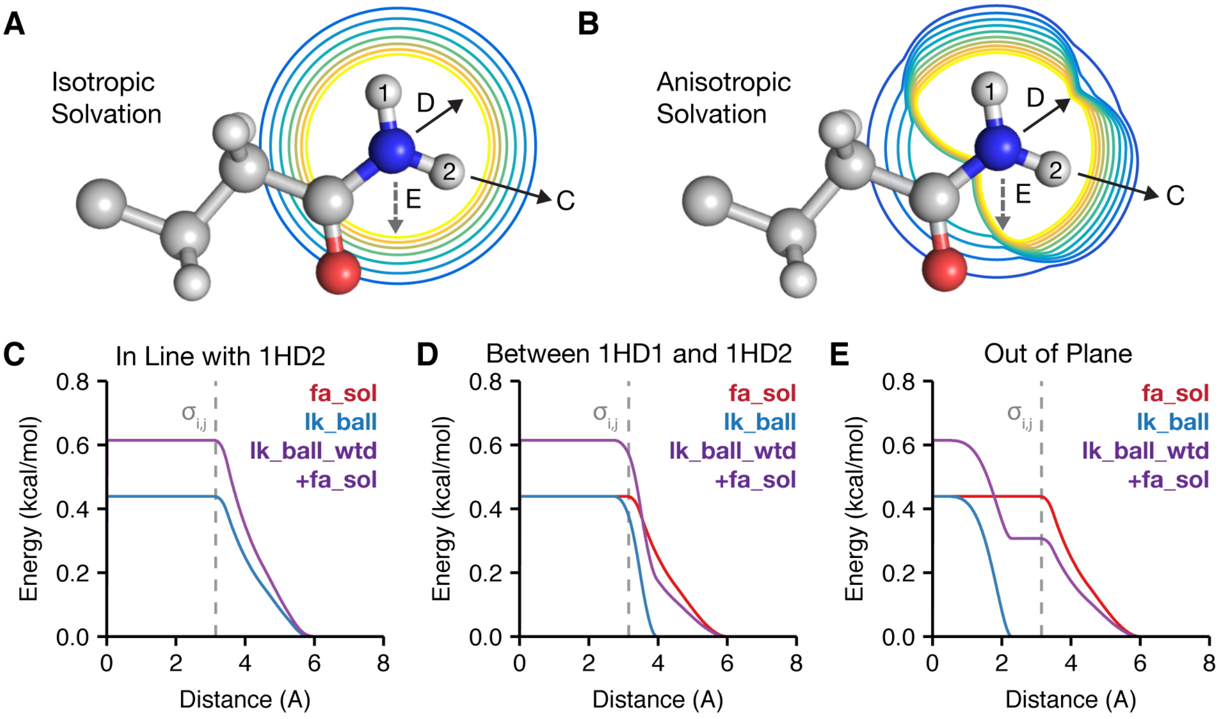


Figure 2: A two component Lazaridis-Karplus solvation model
Rosetta uses two energy terms to evaluate the desolvation of protein side chains: an isotropic (`fa_sol`) and anisotropic (`lk_ball_wtd`) term. (A) and (B) demonstrate the difference between isotropic and anisotropic solvation of the NH2 group by CH3 on the asparagine side chain. The contours vary from low energy (blue) to high energy (yellow). The arrows represent the approach vectors for the pair potentials shown in C-E. In the bottom panel, we compare `fa_sol`, `lk_ball` and `lk_ball_wtd` energies for the solvation of the NH2 group on asparagine for three different approach angles: (C) in line with the 1HD2 atom, (D) along the bisector of the angle between 1HD1 and 1HD2 and (E) vertically down from above the plane of the hydrogens (out of plane).

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

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

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

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

The total isotropic solvation energy (**Eq. 13**), `fa_sol`, is computed as a sum including atom *j* desolvating atom *i* and vice-versa and scaled by the previously-defined connectivity weight.

$$E_{\text{fa\_sol}} = \sum_{i,j} w_{i,j}^{\text{conn}} \left(g_{\text{desolv}}(i,j) + g_{\text{desolv}}(j,i)\right) \qquad (13)$$

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

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

To compute the anisotropic energy, we first calculate the set of ideal water sites around atom *i*, $\mathcal{W}_i = \{\boldsymbol{v}_{i1}, \boldsymbol{v}_{i2}, \ldots\}$. This set contains 1 to 3 water sites, depending on the atom type of atom *i*. Each site is 2.65 Å from atom i and has an optimal hydrogen-bond 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* (radius σ_j) touches the van der Waals sphere of the water at site *k* (radius σ_w), and then the term smoothly increases over a zone of partial overlap of approximately 0.5 Å. Thus, for each water site, k, with coordinates $\boldsymbol{v}_{i,k}$, we compute an occlusion measure d_k^2 to capture the gap between the hypothetical water and the desolvating atom *j* (**Eq. 14**), using the offset Ω = 3.7 Å^2 to provide the ramp-up buffer.

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

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

$$d^2_{\min}(i,j) = -\ln\big(\textstyle\sum_{k \in \mathcal{W}_i} \exp(-d^2_k)\big) \qquad (15)$$

Then, $d^2_{\min}$ and Ω are used to compute a damping function f_{lkfrac} (**Eq. 16**) that varies from zero when the desolvating atom is at least a van der Waals distance from any preferred water site to one when the desolvating atom overlaps a water site by more than ~ 0.5 Å.

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

We calculate the anisotropic energy of desolvating a polar atom $E_{\text{lk\_ball}}$ by scaling the desolvation function g_{desolv} by the damping function f_{lkfrac} and an atom-type specific weight w_{aniso} that is typically ~0.7 (**Eq. 17**). The amount of isotropic solvation energy subtracted is g_{desolv} multiplied by w_{iso}, where w_{iso} is an atom-type specific weight typically ~0.3 (**Eq. 18**; the total weight on the isotropic contribution through both `fa_sol` and `lk_ball_wtd` terms is thus ~0.7). The isotropic and anisotropic components are then summed to yield a new desolvation function, h_{desolv} (**Eq. 19**).

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

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

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

Like `fa_sol`, the energy of desolvating atom i by atom j and then j by i are summed to yield the overall `lk_ball_wtd` energy (**Eq. 20**) but only counting the desolvation of polar, hydrogen-bonding heavy atoms (O,N) defined as the set $\mathcal{P}$. **Fig. 2** shows a comparison between `fa_sol`, the `lk_ball` term (**Eq. 17**), and the sum of `fa_sol` and `lk_ball_wtd` for the example of an asparagine NH2 desolvated from three different approach angles. As the approach angle varies, the sum of `lk_ball_wtd` and `fa_sol` creates a larger desolvation penalty when waters sites are occluded and a smaller penalty otherwise, relative to the `fa_sol` term alone.

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

Hydrogen bonding. Hydrogen bonds are partially covalent interactions that form when a nucleophilic heavy atom donates electron density to a polar hydrogen.[75] At short ranges (< 2.5 Å), they exhibit geometries that maximize orbital overlap.[76] 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 energy of hydrogen bonds using `fa_elec` and a term called `hbond` that evaluates energies based on the orientation preferences of hydrogen bonds found in high-resolution crystal structures.[38,48] To derive this model, we curated intra-protein polar contacts from ~8,000 high resolution crystal structures (Top8000 dataset[77]) 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 (**Fig. 3A**): (1) the distance between the donor and acceptor, d_{HA} (2) the angle formed by the donor, acceptor, and donor-heavy atom, θ_{AHD} (3) the angle formed by the acceptor's parent atom ("base") B, acceptor, and the donor, θ_{BAH} and (4) the torsion, ϕ_{B_2BAH}, formed by the donor, acceptor, and two subsequent parent atoms B and B_2. (**Fig. 3A**). B, the parent atom of A, is the first atom on the shortest path to the root atom (e.g. C$_\alpha$). 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)

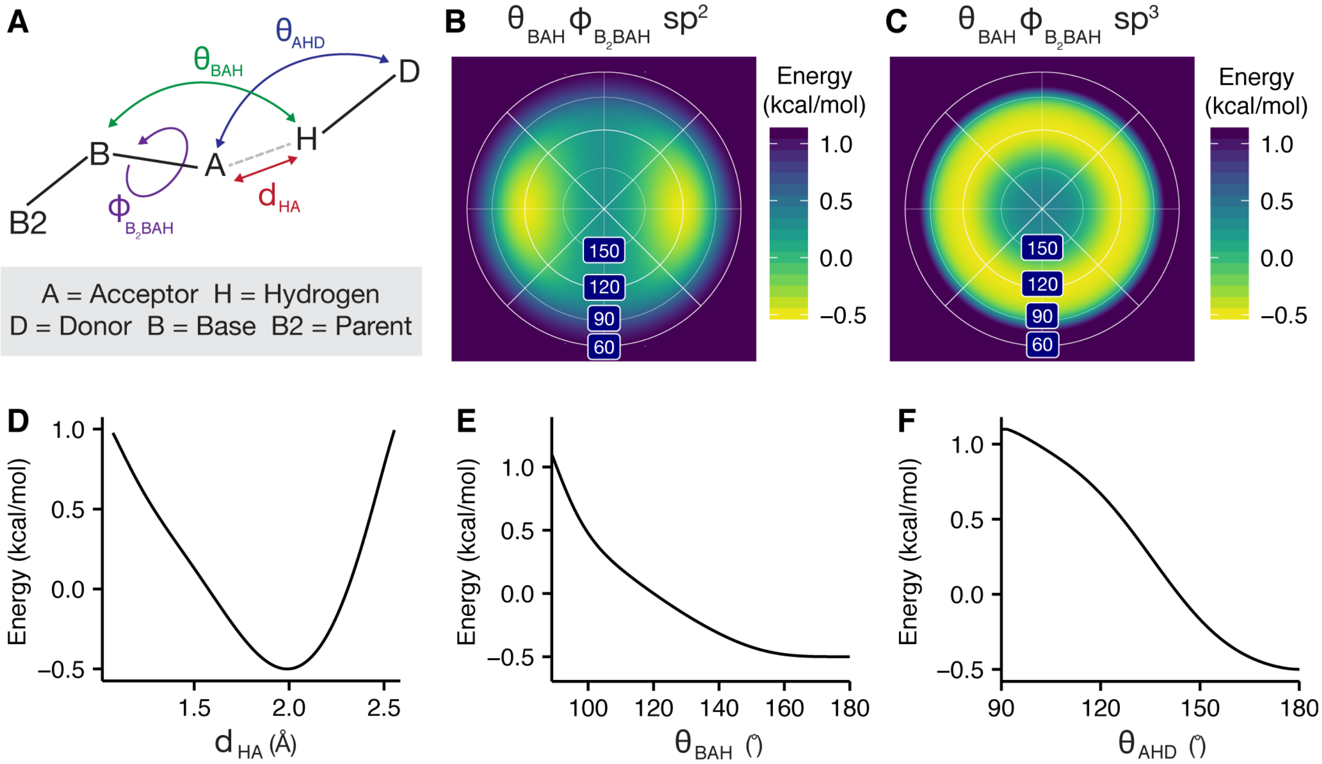


Figure 3: Orientation-dependent hydrogen bonding model
(A) Degrees of freedom evaluated by the hydrogen bonding term: acceptor—donor distance, d_{HA}, angle between the base, acceptor and hydrogen θ_{BAH}, angle between the acceptor, hydrogen, and donor, θ_{AHD}, and dihedral angle corresponding to rotation around the base—acceptor bond, ϕ_{B_2BAH}. (B) Lambert-azimuthal projection of the $E_{\text{hbond}}^{B_2BAH}$ energy landscape for an sp^2 hybridized acceptor.[48] (C) $E_{\text{hbond}}^{B_2BAH}$ energy landscape for an sp^3 hybridized acceptor. Example energies for the histidine imidazole ring acceptor hydrogen bonding with a protein backbone amide: (D) energy vs. the acceptor—donor distance, E_{hbond}^{HA} (E) energy vs. the acceptor-hydrogen-donor angle, E_{hbond}^{AHD} (F) energy vs. the base-acceptor—hydrogen angle, E_{hbond}^{BAH}.

To avoid over-counting, side-chain to backbone hydrogen bonds are excluded if the backbone group is already involved in a hydrogen bond. For speed, the component terms have simple analytic functional forms (**Fig. 3B-F**; Supporting Information **Eq. S1-7**). 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** where the $E_{\text{hbond}}^{B_2BAH}$ term depends on the orbital hybridization of the acceptor, ρ. Finally, the function is also smoothed with $f(x)$ (**Eq. 22**) to avoid derivative discontinuities and ensure that edge-case hydrogen bonds are considered.

$$E_{\text{hbond}} = \sum_{H,A} w_H w_A f\left(E_{\text{hbond}}^{HA}(d_{HA}) + E_{\text{hbond}}^{AHD}(\theta_{AHD}) + E_{\text{hbond}}^{BAH}(\theta_{BAH}) + E_{\text{hbond}}^{B_2BAH}(\rho, \phi_{B_2BAH}, \theta_{BAH})\right) \tag{21}$$

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

Disulfide bonding. Disulfide bonds are covalent interactions that link sulfur atoms in cysteine residues. Typically, in Rosetta, we rely on a tree-based kinematic system[78,79] to keep bond lengths and angles fixed so that we may sample conformation space 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 (below), the extra bonds cannot be represented with a tree (since a tree graph is acyclic), and thus must be treated explicitly. Thus, disulfide bonds are a special case of inter-residue covalent contact 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 Sγ atoms. Then, Rosetta computes the energy of these interactions using an orientation-dependent model called `dslf_fa13`.[48] The model was derived by curating intra-protein 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 1, 2, and 3, von Mises functions (Supporting Information **Eq. S8-11**).

The overall disulfide energy is computed as a function of six degrees of freedom (**Fig. 4**) that map to four component energies. First, the geometry of the sulfur-sulfur distance d_{SS} is evaluated by $E_{\text{dslf}}^{SS}(d)$. Second, the angle formed by either $C_{\beta 1}$ or $C_{\beta 2}$ with S-S bond is evaluated by $E_{\text{dslf}}^{CSS}(\theta)$. Third, the dihedral formed by either $C_{\alpha 1}C_{\beta 1}$ or $C_{\alpha 2}C_{\beta 2}$ with the S-S bond is evaluated by $E_{\text{dslf}}^{C_\alpha C_\beta SS}(\phi)$. Finally, the dihedral formed by $C_{\beta 1}, C_{\beta 2}$ and the S-S bond is evaluated by $E_{\text{dslf}}^{C_\beta SSC_\beta}(\phi)$. The complete disulfide bonding energy evaluated for all S-S pairs is given by **Eq. 23**.

$$E_{\text{dslf\_fa13}} = \sum_{S_1,S_2} E_{\text{dslf}}^{SS}(d_{SS}) + E_{\text{dslf}}^{CSS}\left(\theta_{C_{\beta 1}SS}\right) + E_{\text{dslf}}^{CSS}\left(\theta_{C_{\beta 2}SS}\right) + E_{\text{dslf}}^{C_\alpha C_\beta SS}\left(\phi_{C_{\alpha 1}C_{\beta 1}SS}\right) + E_{\text{dslf}}^{C_\alpha C_\beta SS}\left(\phi_{C_{\alpha 2}C_{\beta 2}SS}\right) + E_{\text{dslf}}^{C_\beta SSC_\beta}\left(\phi_{C_{\beta 1}SSC_{\beta 2}}\right) \tag{23}$$

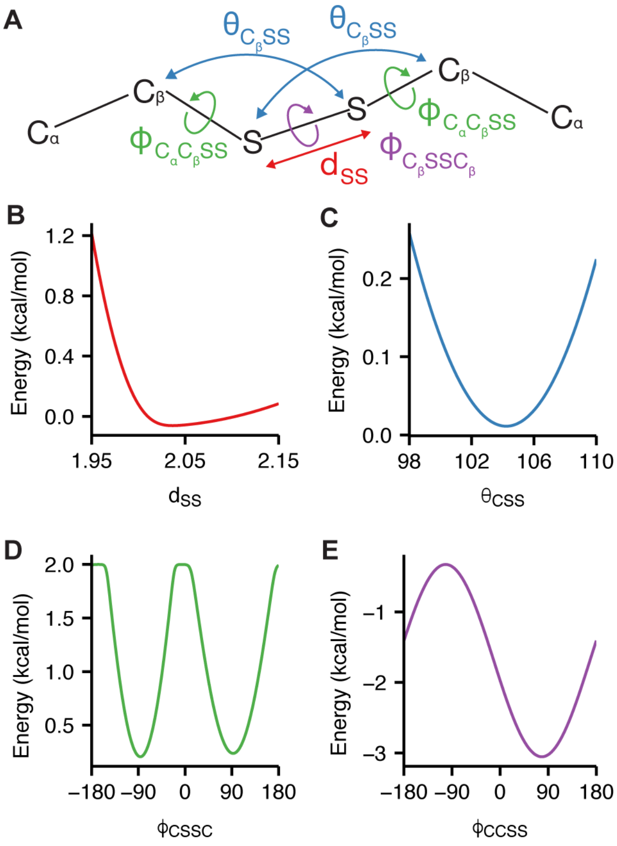


Figure 4: Orientation-dependent disulfide bonding model
(A) Degrees of freedom evaluated by the disulfide bonding energy: sulfur—sulfur distance, d_{ss}, angle between the β-carbon and two sulfur atoms, θ_{CSS}, dihedral corresponding to rotation about the α-Carbon and sulfur bond $\phi_{C_\alpha C_\beta SS}$, and dihedral corresponding to rotation about the S—S bond ϕ_{SS}. (B) E_{dslf}^{SS} (C) E_{dslf}^{CSS} (D) $E_{\text{dslf}}^{C_\alpha C_\beta SS}$ (E) $E_{\text{dslf}}^{C_\beta SSC_\beta}$.

Terms for Protein Backbone and Side Chain Torsions

Rosetta evaluates backbone and side-chain conformations in torsion space to greatly reduce the search domain and increase 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.[80] 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. To evaluate backbone ϕ and ψ angles, we defined an energy term called `rama_prepro` based on Ramachandran maps for each amino acid, using torsions from 3,985 protein chains with a resolution ≤ 1.8 Å, R-factor ≤ 0.22 and sequence identity $\leq 50\%$.[81] Amino acids with low electron density (in the bottom 25th percentile of each residue type) were removed from the data set. The resulting ~581,000 residues were used in adaptive kernel density estimates[51] of Ramachandran maps with a grid step of 10° for both ϕ and ψ. Residues preceding proline are also treated separately because they exhibit distinct ϕ, ψ preferences due to steric interactions with the proline's C_δ.[82] The energy, called `rama_prepro,` is then computed by converting the probabilities to energies at the grid points via the inverted Boltzmann relation[83] (**Eq. 24**; **Fig 5**). The energies are then evaluated using bicubic interpolation. The Supporting

Information includes a detailed discussion of why interpolation is performed on the backbone torsional *energies* rather than the *probabilities* (**Fig. S3**, **Eqs. S12-13**).

$$E_{\text{rama\_pre\_pro}} = \sum_i \begin{cases} -\ln\left[P_{\text{reg}}(\phi_i,\psi_i|\text{aa}_i)\right] & \text{C-terminus or } i\text{+1 is not a proline} \\ -\ln\left[P_{\text{prepro}}(\phi_i,\psi_i|\text{aa}_i)\right] & i\text{+1 is a proline} \end{cases} \qquad (24)$$

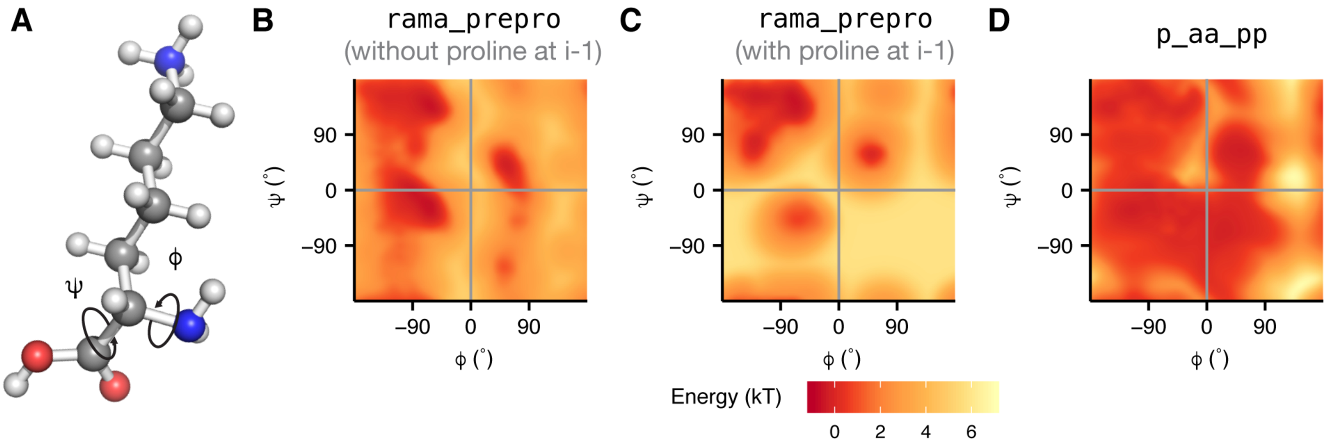


Figure 5: Backbone torsion energies
The backbone-dependent torsion energies are demonstrated for the lysine residue. (A) The ϕ angle is defined by the backbone atoms $C_{i-1} - N - C_\alpha - C$ and the ψ angle is defined by $N - C_\alpha - C - N_{i+1}$. (B) `rama_prepro` energy of lysine without a proline at i+1. (C) `rama_prepro` energy of lysine with a proline at i+1. (D) `p_aa_pp` energy of lysine.

Backbone design term. Rosetta also computes the likelihood of placing a specific amino acid side chain given an existing ϕ,ψ backbone conformation. This term, called `p_aa_pp` represents the propensity of observing an amino acid relative to the other 19 canonical amino acids.[84] The knowledge-based propensity, $P(\text{aa}|\phi,\psi)$ (**Eq. 25**) was derived using the adaptive kernel density estimates for $P(\phi,\psi|\text{aa})$ and Bayes' rule. The equation for `p_aa_pp` is given in **Eq. 26** (**Fig. 5D**).

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

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

Side-chain conformations. Protein side chains mostly occupy discrete conformations (rotamers) separated by large energy barriers. To evaluate rotamer conformations, Rosetta derives probabilities from the 2010 backbone-dependent rotamer library (dunbrack.fccc.edu/bbdep2010), which contains the frequencies, means, and standard deviations of individual χ angles for each χ angle k of each rotamer of each amino acid type.[51] The probability has three components: (1) observing a specific rotamer given the backbone dihedral angles (2) observing specific χ angles given the rotamer and (3) observing the terminal χ angle distribution, which is either Gaussian-like or continuous when the terminal χ angle is sp^2 hybridized (**Eq. 27**). Here, T represents the number of rotameric χ angles + 1.

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

The 2010 rotamer library distinguishes between rotameric and non-rotameric torsions. A torsion is rotameric when the third of the four atoms defining the torsion is sp^3 hybridized (i.e. preferring ~60°, ~180° and ~-60°, with steep energy barriers between the wells), If the last χ torsion is rotameric, probability $p(\chi_T|\phi,\psi,\text{rot,aa})$ is fixed at one. On the other hand, a torsion is non-rotameric its third atom is sp^2 hybridized: the library describes its probability distribution continuously, instead. The category of semi-rotameric amino acids with both rotameric and non-rotameric dihedrals encompasses eight amino acids: Asp, Asn, Gln, Glu, His, Phe, Tyr, and Trp.[85]

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

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

This formulation is reminiscent of the Gaussian distribution, except that it is missing the normalization coefficient of $\left(2\pi\sigma_{\chi_k}(\phi,\psi|\text{rot,aa})\right)^{-1/2}$. After taking the log of this probability, the term resembles Hooke's law where the spring constant is given by $\sigma_{\chi_k}^{-2}(\phi,\psi|\text{rot,aa})$.

The full form of `fa_dun` is given by **Eq. 29** as a sum over all residues *r*. The difference between the rotameric- and semi-rotameric models is also shown in **Fig. 6**.

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

The energy from $-\ln\left(P(\text{rot}_r|\phi_r,\psi_r,\text{aa}_r)\right)$ is computed using bicubic-spline interpolation; $P\left(\chi_{T_r,r}|\phi_r,\psi_r,\text{rot}_r,\text{aa}_r\right)$ is computed using tricubic-spline interpolation. To save memory, $\mu_{\chi_k}(\phi_r,\psi_r|\text{rot}_r,\text{aa}_r)$, and $\sigma_{\chi_k}(\phi_r,\psi_r|\text{rot}_r,\text{aa}_r)$ are computed using bilinear interpolation, though this has the effect of producing derivative discontinuities at the (ϕ,ψ) grid boundaries. These discontinuities, however, do not appear to produce noticeable artifacts.[50]

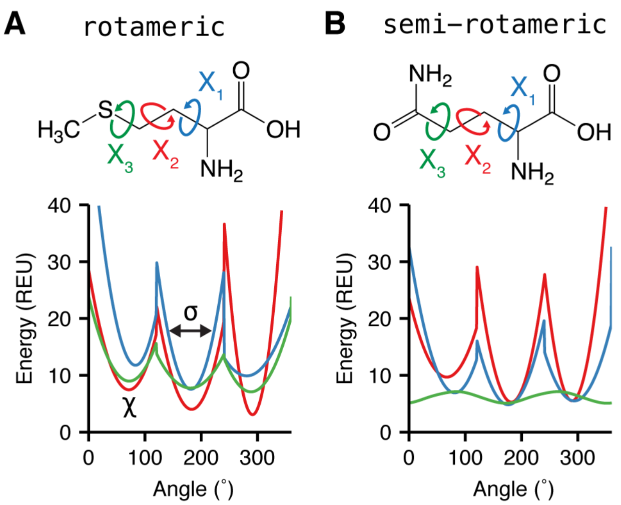


Figure 6: Energies for side-chain rotamer conformations
The Dunbrack rotamer energy, `fa_dun`, is dependent on both the ϕ and ψ backbone torsions and the χ side-chain torsions. Here, we demonstrate the variation of `fa_dun` when the backbone is fixed in an α-helical conformation with ϕ = -57˚ and ψ =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

Terms for special case torsions

Peptide bond dihedral angles, ω, remain mostly fixed in a *cis*- or *trans*- conformation and depend on the backbone ϕ and ψ angles. Since the electron pair on the backbone nitrogen donates electron density to the electrophilic carbonyl carbon, the peptide bond has partial double bond character. To model this barrier to rotation, Rosetta implements a backbone-dependent harmonic penalty centered near 0° for cis and 180° for trans (**Fig. 7A**). This energy, called `omega`, is evaluated on all peptide bonds in the biomolecule (**Eq. 30**). The means and standard derivations of ω, $\mu_\omega$and σ_ω, respectively, are backbone (ϕ,ψ) dependent, as given by kernel regressions of ω on ϕ and ψ.[70]

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

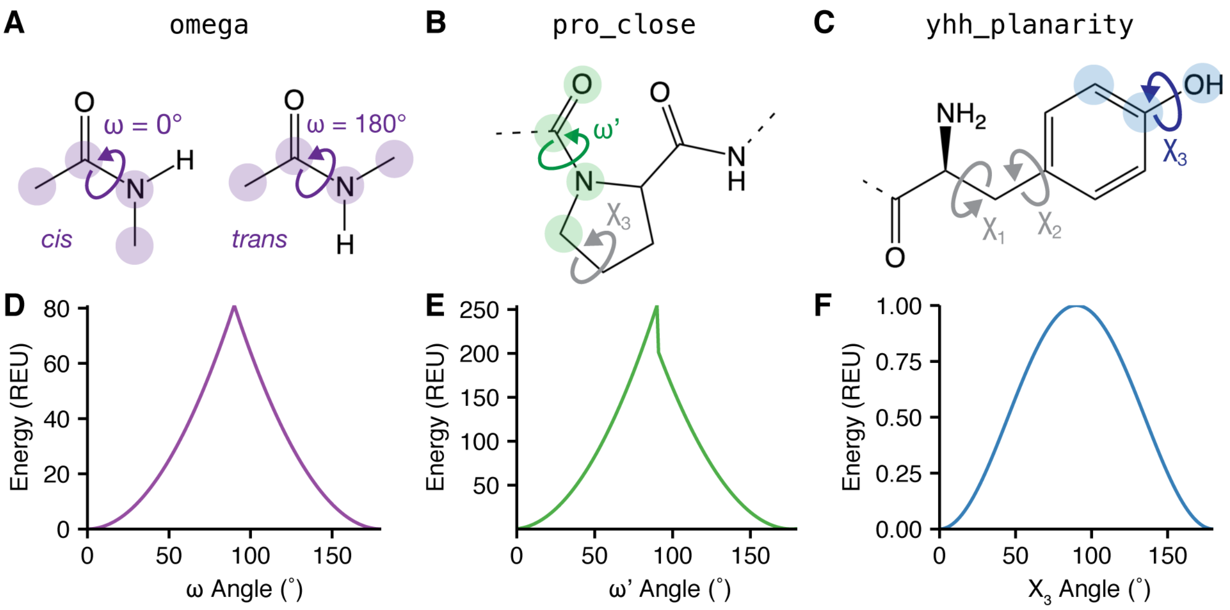


Figure 7: Special case torsion energies

Rosetta implements three additional energy terms to model torsional degrees of freedom with acute preferences. (A) Omega torsion corresponding to rotation about C-N (B) Proline secondary omega torsion corresponding to rotation about C-N related to the C-δ in the ring. (C) Tyrosine terminal χ torsion. (D) Omega energy (E) Proline closure energy (F) Tyrosine planarity energy.

Most Rosetta protocols only search over simple torsions within chains and rigid-body degrees of freedom between chains. However, **proline's side chain** requires special treatment because its ring cannot be represented by a kinematic tree.[86] Therefore, Rosetta implements a proline closure term, called `pro_close` (**Fig. 7B**). There are two components to this energy, shown in **Eq. 31**. First, there is a torsional potential that operates on the dihedral formed by $O_{r\text{-}1}$–$C_{r\text{-}1}$–N_r–$C_{\delta,r}$, called ω'_r given the observed mean $\mu_{\omega'}$ and standard deviation $\sigma_{\omega'}$, where i is the residue index. This term keeps the C_δ atom in the peptide plane. Second, to ensure correct geometry for the two hydrogens bound to C_δ, we build a virtual atom, N_v, off C_δ whose coordinate is controlled by χ_3 (**Fig. 7B**). The `pro_close` term seeks to align the virtual N_v atom, directly on top of the real backbone nitrogen. The N–C_δ–C_γ bond angle and the N–C_δ bond length are restrained to their ideal values.

$$E_{\text{pro\_close}} = \sum_{r \in \text{Pro}} \begin{cases} \frac{(\omega'_r - \mu_{\omega'})^2}{\sigma^2_{\omega'}} + \frac{\|\mathbf{N}_r - \mathbf{N}_{v,r}\|^2}{\sigma^2_{\mathrm{N,N}_v}} & r \text{ is not N-terminus} \\ \frac{\|\mathbf{N}_r - \mathbf{N}_{v,r}\|^2}{\sigma^2_{\mathrm{N,N}_v}} & r \text{ is N-terminus} \end{cases} \qquad (31)$$

Tyrosine also requires special treatment for its $\boldsymbol{\chi_3}$ **angle** because the hydroxyl hydrogen prefers to be in the plane of the aromatic ring.[87] To enforce this preference, Rosetta implements a sinusoidal penalty to model the barrier to a χ_3 angle that deviates from planarity. This tyrosine hydroxyl penalty is called `yhh_planarity` (**Eq. 32**; **Fig. 7C**).

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

Terms for modeling non-ideal 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 cryo-electron microscopy,[52] as well as its ability to discriminate near-native conformations in the absence of experimental data.[88] These data suggest that capturing non-ideal bond lengths and angles can be important for accurate modeling of minimum-energy protein conformations. To accommodate, Rosetta now allows these "non-ideal" angles and lengths to be included as additional degrees of freedom in refinement and includes a Cartesian-minimization mode where atom coordinates are explicit degrees of freedom in optimization.

To evaluate the energetics of non-ideal 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 (**Eq. 32-34**). Here, d_i is a bonded-atom-pair distance with $d_{i,0}$ as its ideal distance, θ_i is a bond angle with $\theta_{i,0}$ as its ideal angle, and ϕ_i is a bond torsion or improper torsion with $\phi_{i,0}$ as its ideal value and ρ_i as its periodicity. The ideal bond lengths and angles[89,90] were selected based on their ability to rebuild side chains observed in crystal structures (Kevin Karplus & James J. Havranek, unpublished); they were subsequently modified empirically.[50] The spring constants for the angle and length terms are from CHARMM32.[19] Finally, all planar groups and the C_β "pseudo-torsion" are constrained using empirically derived values and spring constants:

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

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

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

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

Terms for Protein Design

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

Rosetta calculates the reference energy as a sum of individual constant unfolded state reference energies, ΔG_i^{ref}, for each amino acid, aa$_i$ (**Eq. 36**).[1]

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

The $\Delta G_i^{\mathrm{ref}}$ values are empirically optimized by searching for values that maximize native sequence recovery (discussed below) during design simulations on a large set of high-resolution crystal structures.[49,50] 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 the terms using a weighted linear sum to approximate free energies (**Table 1**). Historically, we adjust the weights and parameters to balance the energetic contribution from each term. This balance is important because the van der Waals, solvation, and electrostatics energies partially capture torsional preferences and overlap can cause errors as a result of double counting atomic or residue specific contributions.[91] More recently, we fix physics-based terms with weights of 1.0 and perturb other weights and atomic-level parameters using a Nelder-Mead[92] scheme to optimize agreement of Rosetta calculations with small-molecule thermodynamic data and high-resolutions structural features.[49] The energy function parameters have evolved over the years by optimizing the performance of multiple scientific benchmarks (**Table 2**).[49,50,93] These benchmarks were chosen to test 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*.,[94] Conway *et al*.[95] and O'Meara *et al*.[46] have fit intra-term 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.*[96] **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 terms of Rosetta energy were not in kcal/mol, 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 interpretation. The current energy function (*beta_nov15*) was parameterized on small molecule thermodynamic data and high-resolution protein structures in units of kcal/mol.[49] The optimization data show a strong correlation (R = 0.994) between the experimental data and values predicted by Rosetta ($\Delta\Delta G$ upon mutation, small molecule ΔH_{vap}; **Fig. S2**); therefore, as is standard practice for molecular force fields such as OPLS, CHARMM, and AMBER, we now express energies in kcal/mol.

Table 2: Common energy function benchmarking methods

Test	Description	Ref.
Sequence Recovery	Percentage of the native sequence recovered after backbone redesign	[1,50]
Rotamer Recovery	Percentage of native rotamers recovered after full repacking	[50]
ΔΔG Prediction	Prediction of free energy changes upon mutation	[97]
Loop Modeling	Prediction of loop conformations	[98]
High-resolution refinement	Discrimination of native-like decoys upon refinement of *ab initio* protein models	[99]
Docking	Prediction of protein-protein interfaces	[100,101]
Homology Modeling	Structure prediction incorporating homologous information from templates	[102]
Thermodynamic properties	Recapitulation of thermodynamic properties of protein side-chain analogues	[17]
Recapitulation of Xtal structure geometries	Recapitulation of features (e.g. atom-pair distance distribution) from high-resolution protein crystal structures	[49]

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 energies can be applied to detailed interpretations of structural models. In this section, we discuss two common structure calculations: (1) estimating the free energy change ($\Delta\Delta G$) of mutation[97] and (2) modeling the structure of a protein-protein interface.[101]

$\Delta\Delta G$ of mutation. The first example demonstrates how Rosetta can be used to estimate and rationalize thermodynamic parameters. Here, we present an example $\Delta\Delta G$ of mutation calculation for the T193V mutation in the RT-RH derived peptide bound to HIV-1 protease (PDB 1kjg, **Fig. 8A**).[103] The details of this calculation are provided in the Supporting Information.

Rosetta calculates the $\Delta\Delta G$ of the T193V mutation to be -4.95 kcal/mol, and the experiment[103] measured -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[104] to identify important energetic contributions to the total $\Delta\Delta G$. First, we decomposed the $\Delta\Delta G$ into individual energy terms and observe the balance of terms, both favorable and unfavorable, that sum to the total (**Fig. 8B**). 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 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 (**Fig. 8C**).

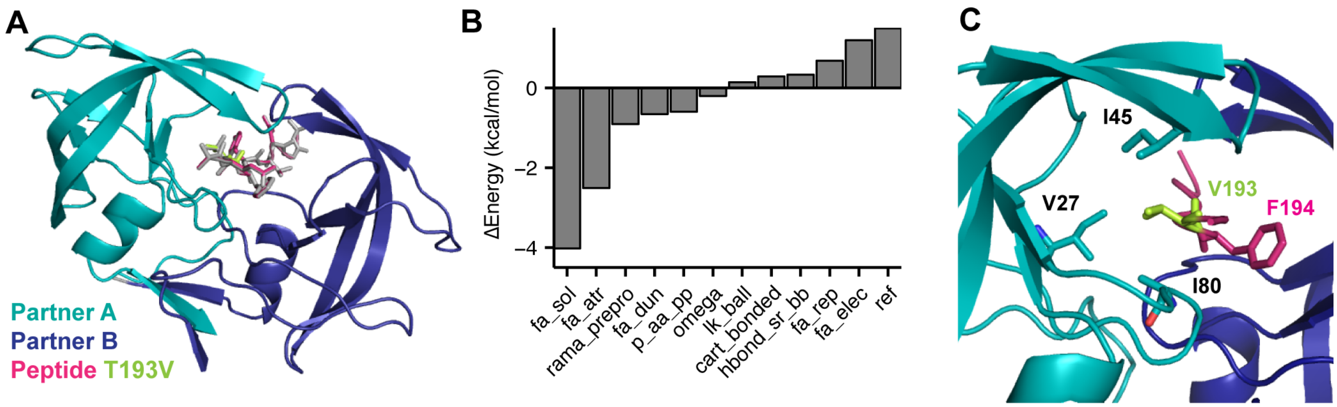


Figure 8: Structural model of the HIV-1 protease bound to the T4V mutant RT-RH derived peptide
(A) Structural model of the native HIV-1 peptidase (teal and dark blue), bound to the native peptide (gray) superimposed onto the T4V mutant peptide (magenta). (B) Contributions greater than ± 0.1 kcal/mol to the ΔΔG of mutation for T4V. The remaining contributions are: `dslf_fa13` = 0 kcal/mol, `hbond_lr_bb` = -0.09 kcal/mol, `hbond_bb_sc` = -0.05, `hbond_sc` = -0.0104, `fa_intra_rep` = 0.01, `fa_intra_sol` = -0.07, and `yhh_planarity` = 0. (C) Hydrophobic patch of residues surrounding position four on the RT-RH peptide.

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

Table 3: Change in atom pair energies between I80 and T4 versus V4 in kcal/mol

T193→V193 Atoms	I80 Atoms			
	CB	CG1	CG2	CD1
N	0.000	0.000	0.000	0.000
CA	0.000	0.000	0.000	0.004
C	0.000	0.000	0.000	0.008
O	0.000	0.000	0.000	-0.010
CB	0.000	0.054	0.000	-0.002
OG1 → CG1	0.008	-0.054	-0.316	-0.398
CG2 → CG2'	0.000	0.000	0.001	0.020

Protein-protein docking. The second example shows how the Rosetta energies of an ensemble of models can be used to discriminate between models and investigate the characteristics of a protein–protein interface. Below, we investigate docked models of West Nile Virus envelope protein and a neutralizing antibody (PDB 1ztx; **Fig. 9A**).[105] Calculation details can be found in the Supporting Information.

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

Like the previous example, we decompose the energies to yield information about the nature of interactions at the interface. Here, we observed 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` (**Fig. 9B**). 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 (RMS→0), the favorable, attractive energy (`fa_atr`) decreases whereas the unfavorable, repulsive energy (Δ`fa_rep`) increases. Change in the isotropic solvation energy (`fa_sol`) is positive (unfavorable), indicating that upon interface formation, polar residues are buried. Balancing the desolvation penalty, the change in polar solvation energy (`lk_ball_wtd`) and electrostatics (`fa_elec`) is negative due to polar contacts forming at the interface. Finally, 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.

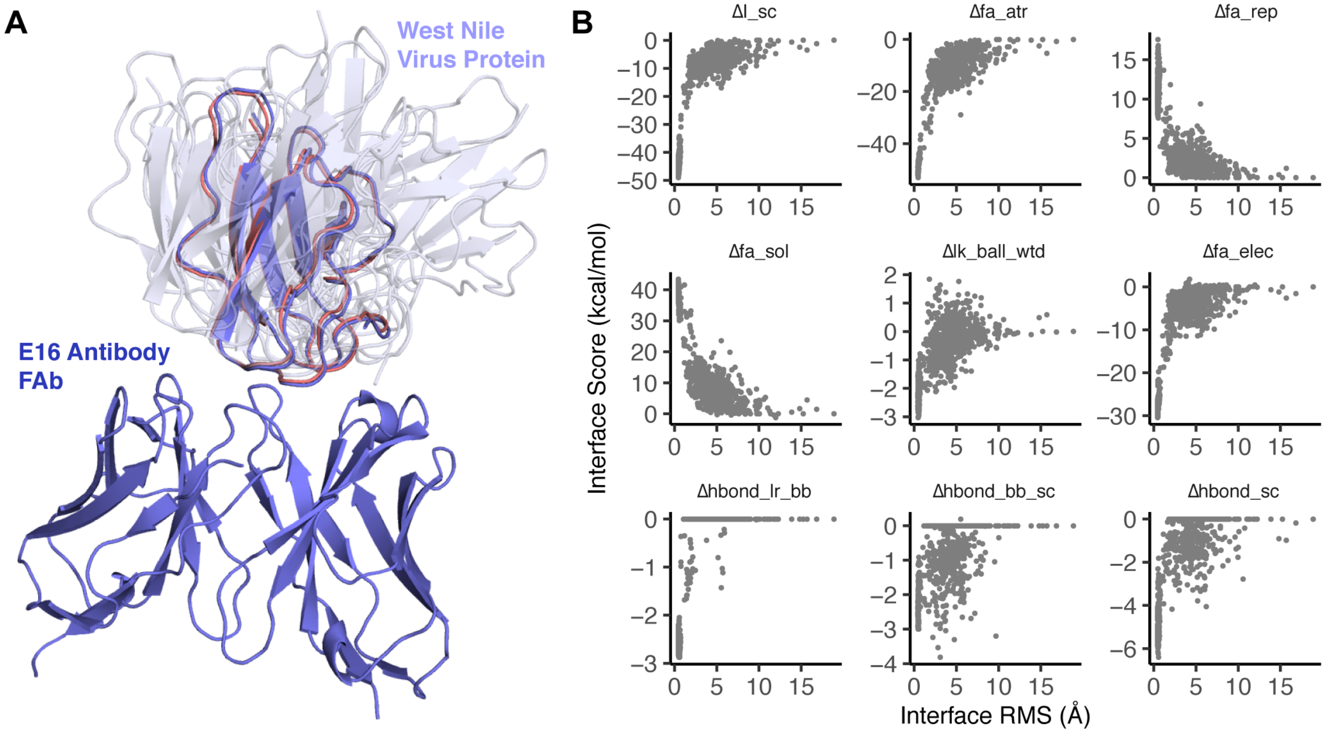


Figure 9: Using energies to discriminate docked models of West Nile Virus and the E16 neutralizing antibody

(A) Comparison of the native E16 antibody (purple) docked to the lowest RMS model of the West Nile Virus envelope protein and several other random models of varying energy to show sampling diversity (gray, semi transparent). (B) Change in the interface energy relative to the unbound state versus RMS to native. Models at low RMS to the native interface have a low overall interface energy due to favorable van der Waals contacts, electrostatic interactions, and side-chain hydrogen bonds, as reflected by the Δ`fa_atr`, Δ`fa_elec`, and Δ`hbond_sc` energy terms.

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; yet, energy functions are far from perfect. Compared to the first torsional potentials, 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 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 design of synthetic macromolecules to predicting 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 non-canonical amino acids and non-protein biomolecules (**Table S5**). Recently, Bhardwaj, Mulligan & Bahl *et al.*[67] adapted the `rama_prepro`, `p_aa_pp`, `fa_dun`, `pro_close`, `omega`, `dslf_fa13`, `yhh_planarity` and `ref` terms to be compatible with mixed-chirality peptides. Several of Rosetta's statistical potentials are validated against quantum mechanical calculations for evaluating for non-protein models (**Table 4**). The first non-protein terms were added by Havranek *et al.*[106] and Yu *et al.*[107] who modified the hydrogen bonding potential to capture planar hydrogen bonds between protein side chains and nucleic acid bases. Renfrew *et al.*[65,108] added molecular mechanics torsions and Lennard-Jones terms to model proteins with non-canonical amino acids, oligosaccharides, β-peptides, and oligo-peptoids.[66] Labonte *et al* [68] implemented Woods' CarboHydrate-Intrinsic (CHI) function[109,110] which evaluates glycan geometries given the axial-equatorial character of the bonds. In addition, Das *et al.* added a set of terms to model Watson-Crick base pairing, π-π interactions in base stacking, and torsional potentials important for predicting and designing RNA structures.[61,111–113] These terms are presented in detail in the Supporting Information.

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

Table 4: New energy terms for biomolecules other than proteins

Biomolecule	Term	Description	Unit	Ref.
Non-Canonical Amino Acids	`mm_lj_intra_rep`	Repulsive van der Waals energy between two atoms from the same residue	kcal/mol	[65]
	`mm_lj_intra_atr`	Attractive van der Waals energy between two atoms from the same residue	kcal/mol	[65]
	`mm_twist`	Molecular mechanics derived torsion term for all proper torsions	kcal/mole	[65]
	`unfolded`	Energy of the unfolded state based on explicit unfolded state model	AU*	[65]
	`split_unfolded_1b`	One-body component of the two-component reference energy, lowest energy of a side chain in a dipeptide model system	AU	In SI
	`split_unfolded_2b`	Two-body component of the two-component reference energy, median two-body interaction energy based on atom type composition	AU	In SI
Carbohydrates	`sugar_bb`	Energy for carbohydrate torsions	kcal/mol	[68]
DNA	`gb_elec`	Generalized Born model of the electrostatics energy	kcal/mol	[106]
RNA	`fa_stack`	π-π stacking energy for RNA bases	kT	[112]
	`stack_elec`	Electrostatic energy for stacked RNA bases	kT	[113]
	`fa_elec_rna_phos`	Electrostatic energy (`fa_elec`) between RNA phosphate atoms	kT	[61]
	`rna_torsion`	Knowledge-based torsional potential for RNA	kT	[61]
	`rna_sugar_close`	Penalty for opening an RNA sugar	kT	[61]

* AU, arbitrary units

Capturing the intra- and extra-cellular environment

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

Currently, Rosetta includes two groups of energy terms to model alternate environments (**Table 5**). Kilambi *et al.*[59,114] implemented a method to account for alternate protonation states due to pH changes. In addition, Rosetta implements Lazaridis' Implicit Membrane Model for modeling proteins in a lipid bilayer environment.[36,60,115,116] While these improve structure prediction accuracy, both models require more computation time. This trade-off between the need for detail and computational complexity will be evaluated as Rosetta aims to model more complicated biological systems and contexts.

Table 5: Energy terms for structure prediction in different contexts

Context	Term	Description	Unit	Ref.
Membrane Environment	`fa_mpsolv`	Solvation energy dependent on the protein orientation relative to the membrane	kcal/mol	[115,117]
	`fa_mpenv`	One-body membrane environment energy dependent on the protein orientation relative to the membrane	kcal/mol	[115,117]
pH	`e_pH`	Likelihood of side chain protonation given a user-specified pH	kcal/mol	[114]

The origin of energy models: top-down versus bottom-up development

Traditionally, energy functions are developed using a bottom-up approach: experimental observables serve as building blocks to parameterize physics-based formulas. The advent of powerful optimization techniques and artificial intelligence recently empowered the top-down category where numerical methods are used to derive models and/or parameters. Top-down approaches have been used to solve problems in various fields including structural biology and bioinformatics. Recently, top-down development was also applied to optimizing the Lennard-Jones, Lazaridis-Karplus, and Coulomb parameters in the Rosetta energy function (parameters in **Table S4-S6**).[49,92]

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 models, there is a risk of over-fitting as validation via structure prediction datasets 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 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 individual benefits.

A highly interdisciplinary endeavor

The Rosetta energy function has advanced rapidly due to the Rosetta Community: a highly-interdisciplinary collaboration between scientists with diverse backgrounds located in over 50 labs 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,[47] 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 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 fifteen years. Still, the energy function is far from complete and will continue to evolve long after this publication. Thus, we hope 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.

Supporting Information

Supporting Information File: RosettaEnergyFunctionReview_Alford_etal_SupportingInfo.pdf

Author Information

Corresponding Author

Jeffrey J. Gray
Email: jgray@jhu.edu
Department of Chemical and Biomolecular Engineering
3400 N Charles Street
Baltimore, Maryland 21218 United States

Author Contributions

Wrote the manuscript: RFA, JRJ, ALF, JJG
Analysis Scripts and Examples: RFA, JRJ, MSP, JJG
Expertise on protein energy terms and editing: ALF, FPD, MJO, HP, PB, MVS, RLD, BK, JJG
Expertise and description of non-protein energy terms and editing: PDR, KK, VKM, JWL, RB

Funding Sources

RFA is funded by a Hertz Foundation Fellowship and an NSF Graduate Research Fellowship. JRJ and JJG are funded by NIH GM-078221. ALF, JJG and BK are funded by NIH GM-73151. MJO is funded by NSF GM-114961. PDR and RB are funded by the Simons Foundation. MVS and RLD are funded by NIH GM-084453 and NIH GM-111819. MSP is funded by NSF BMAT 1507736. JWL is funded by NIH F32-CA189246. KK is funded by an NSF Graduate Research Fellowship and an SGF Galiban Fellowship. DB, HP and VKM are funded by NIH GM-092802. TK is funded by NIH GM-110089 and GM-117189.

Acknowledgements

While our author list reflects main contributors to this paper, the Rosetta energy function work would not be possible without the collaboration of the entire Rosetta Commons: a community of scientists, engineers, and software developers that have worked together for almost 20 years. We thank the community for their continued dedication to empowering users to ask and answer interesting biological questions with Rosetta. We thank Sergey Lyskov for development of the benchmark server that enables continuous and transparent energy function testing. We also thank Morgan Nance and Henry Lessen for helpful comments on the manuscript.

References

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

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

(3) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K. W.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y.-E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol.* **2011**, *Volume 487*, 545–574.

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

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

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

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

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

(9) Westheimer, F. . Calculation of the Magnitude of Steric Effects. *Steric Eff. Org. Chem.* **1956**, 523–555.

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

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

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

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

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

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

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

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

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

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

(20) Sun, H. COMPASS: An Ab Initio Force-Field Optimized for Condensed-Phase ApplicationsOverview with Details on Alkane and Benzene Compounds. **1998**.

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

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

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

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

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

(26) Bowie, J. U.; Lüthy, R.; Eisenberg, D. A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science (80-.).* **1991**, *253* (5016), 164–170.

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

(28) Skolnick, J.; Kolinski, A. Simulations of the Folding of a Globular Protein. *Science (80-.).* **1990**, *250* (4984), 1121–1125.

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

(30) Arieh Warshel, *; Mitsunori Kato, and; Pisliakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. **2007**.

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

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

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

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

(35) Neria, E.; Fischer, S.; Karplus, M. Simulation of Activation Free Energies in Molecular Systems.

J. Chem. Phys. **1996**, *105* (5), 1902.

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

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

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

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

(40) Bradley, P.; Misura, K. M. S.; Baker, D. Toward High-Resolution de Novo Structure Prediction for Small Proteins. *Science (80-.).* **2005**, *309* (5742).

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

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

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

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

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

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

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

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

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

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

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

(52) DiMaio, F.; Song, Y.; Li, X.; Brunner, M. J.; Xu, C.; Conticello, V.; Egelman, E.; Marlovits, T. C.; Cheng, Y.; Baker, D. Atomic-Accuracy Models from 4.5-Å Cryo-Electron Microscopy Data with

Density-Guided Iterative Local Refinement. *Nat. Methods* **2015**, *12* (4), 361–365.

(53) Fleishman, S. J.; Whitehead, T. A.; Ekiert, D. C.; Dreyfus, C.; Corn, J. E.; Strauch, E.-M.; Wilson, I. A.; Baker, D. Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. *Science (80-.).* **2011**, *332* (6031), 816–821.

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

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

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

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

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

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

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

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

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

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

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

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

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

(67) Bhardwaj, G.; Mulligan, V. K.; Bahl, C. D.; Gilmore, J. M.; Harvey, P. J.; Cheneval, O.; Buchko, G. W.; Pulavarti, S. V. S. R. K.; Kaas, Q.; Eletsky, A.; Huang, P.-S.; Johnsen, W. A.; Greisen, P. J.; Rocklin, G. J.; Song, Y.; Linsky, T. W.; Watkins, A.; Rettie, S. A.; Xu, X.; Carter, L. P.;

Bonneau, R.; Olson, J. M.; Coutsias, E.; Correnti, C. E.; Szyperski, T.; Craik, D. J.; Baker, D. Accurate de Novo Design of Hyperstable Constrained Peptides. *Nature* **2016**, *538* (7625), 329–335.

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

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

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

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

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

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

(74) Warshel, A.; Russell, S. T. Calculations of Electrostatic Interactions in Biological Systems and in Solutions. *Q. Rev. Biophys.* **2009**, *17* (3), 283.

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

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

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

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

(79) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K. W.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y.-E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. In *Methods in enzymology*; 2011; Vol. 487, pp 545–574.

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

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

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

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

(84) Shortle, D. Propensities, Probabilities, and the Boltzmann Hypothesis. *Protein Sci.* **2003**, *12* (6),

1298–1302.

(85) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The Penultimate Rotamer Library. *Proteins* **2000**, *40* (3), 389–408.

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

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

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

(89) Inc, A. Insight II. San Diego 2000.

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

(91) Renfrew, P. D.; Butterfoss, G. L.; Kuhlman, B. Using Quantum Mechanics to Improve Estimates of Amino Acid Side Chain Rotamer Energies. *Proteins Struct. Funct. Bioinforma.* **2007**, *71* (4), 1637–1646.

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

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

(94) Song, Y.; Tyka, M.; Leaver-Fay, A.; Thompson, J.; Baker, D. Structure-Guided Forcefield Optimization. *Proteins* **2011**, *79* (6), 1898–1909.

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

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

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

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

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

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

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

(102) Song, Y.; DiMaio, F.; Wang, R. Y.-R.; Kim, D.; Miles, C.; Brunette, T.; Thompson, J.; Baker, D. *High-Resolution Comparative Modeling with RosettaCM*; 2013; Vol. 21.

(103) Altman, M. D.; Nalivaika, E. A.; Prabu-Jeyabalan, M.; Schiffer, C. A.; Tidor, B. Computational Design and Experimental Study of Tighter Binding Peptides to an Inactivated Mutant of HIV-1 Protease. *Proteins Struct. Funct. Bioinforma.* **2007**, *70* (3), 678–694.

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

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

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

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

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

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

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

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

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

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

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

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

(116) Yarov-Yarovoy, V.; Schonbrun, J.; Baker, D. Multipass Membrane Protein Structure Prediction Using Rosetta. *Proteins* **2006**, *62* (4), 1010–1025.

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