Chapter 4: The protein design algorithm

This chapter has been adapted from supporting information for:

Boas FE and Harbury PB. (2008) "Physics-based design of protein-ligand binding." *Journal of Molecular Biology*. In press.

Potential energy function

Overview

The potential energy of a specific protein conformation can be partitioned into different categories. On one extreme, in a quantum calculation, all of the energy is electrostatic. On the other extreme, in an intuitive sense, the energy can be thought of in terms of hydrogen bonds, salt bridges, and steric complementarity. In between, there are molecular mechanics models that treat the protein as a collection of atoms with partial charges and van der Waals parameters, connected by springs to maintain bond lengths and angles.[103]

What is the right type of model to use for protein design? Currently, most protein design algorithms use statistical terms, derived by, for example, counting how frequently different types of hydrogen bonds and salt bridges are seen in crystal structures. The advantage of this approach is that the geometry of the interaction does not have to be exactly correct to get a reasonable energy, and it can include empirically observed phenomena that otherwise might not be modeled correctly. The

disadvantage is that you can't model cases that are missing in your training set. With more detailed sampling in conformational space, we believe it will be more accurate to directly calculate the strengths of salt bridges and hydrogen bonds from Coulomb's law and continuum electrostatics. Thus, in this paper, we have avoided statistical terms, and base all of our calculations on molecular mechanics with continuum solvent.

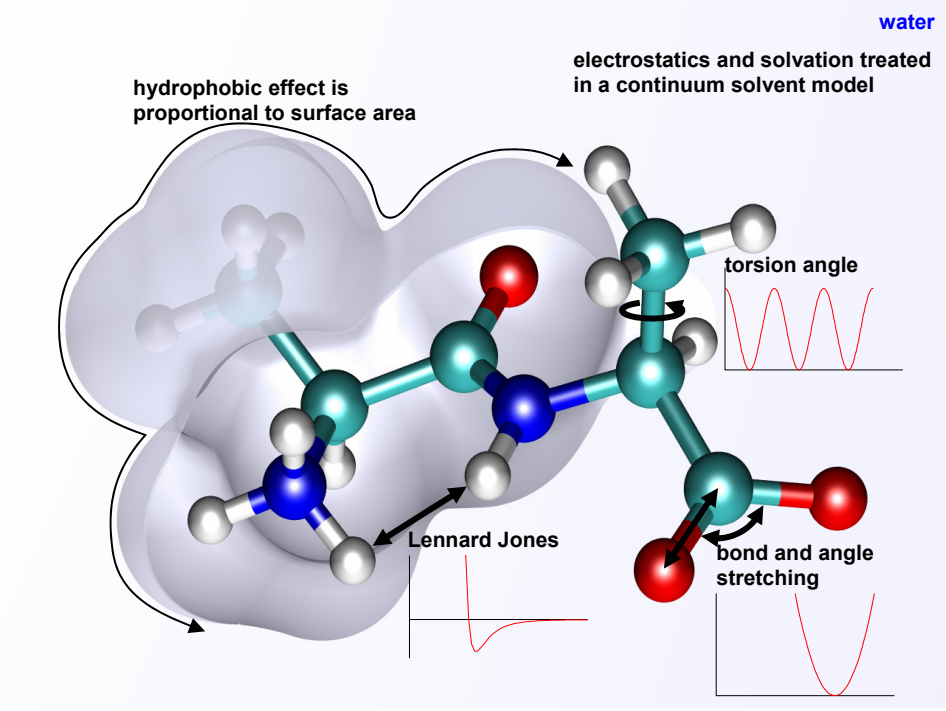
We calculate protein stability and protein-ligand dissociation energy as a difference between states:

Stability = unfolded protein energy – protein energy

Dissociation energy = protein energy + ligand energy – protein·ligand complex energy

Thus, for example, a buried salt bridge might have a Coulomb interaction energy of 100 kcal/mol, but the dissociation energy will be much less than this, because in the undocked state, those charges will have similarly favorable interactions with water. These calculations generally have a lot of large terms that almost cancel each other out, so it is important to do the calculations very carefully.

Our potential energy function (Figure 16) allows us to model effects that are typically ignored in protein design (Figure 17).



potential energy = **molecular mechanics +** (bond length + angle + torsion + LJ + Coulomb) **generalized Born +** (solvent polarization) **surface area +** ("hydrophobic" effect) **protonation energy** (pH effect)

$$= \sum_{\text{bonds}} k_b(b-b_0)^2 + \sum_{\text{angles}} (k_{UB}(S-S_0)^2 + k_\theta(\theta-\theta_0)^2) + \sum_{\text{dihedrals}} k_\chi(1+\cos(n\chi-\delta))$$

$$+ \sum_{\text{impropers}} k_\phi(\phi-\phi_0)^2 + \sum_{\substack{\text{nonbonded} \\ i<j}} E_{VDW}\left(\left(\frac{r_{min}}{r}\right)^{12} - 2\left(\frac{r_{min}}{r}\right)^{6}\right)$$

$$+332 * \sum_{\substack{\text{nonbonded} \\ i<j}} \frac{q_i q_j}{\varepsilon_{in} r} + 332 * \sum_{i,j}^{N} \text{GB}(q_i, q_j, a_i, a_j, r, \varepsilon_{in}, \varepsilon_{out}, \kappa)$$

$$+k_{SASA}\text{SASA} + \sum_{\substack{\text{deprotonated} \\ \text{amino acids}}} U_{deprot.}$$

Figure 16. Potential energy function.

All parameters[14] were from CHARMM22 except for k_{SASA} and U_{deprot}. For the generalized-Born solvation energy, a water radius of 1.4 Å was used to define the molecular surface. Distance

is in angstroms, charge is in elementary charge units, and energy is in kcal/mol. "332" is the Coulomb electrostatic constant for these units.

Variables

k_b	spring constant for bond length	E_{VDW}	van der Waals energy
b	bond length	r	inter-atom distance
b_0	equilibrium bond length	r_{min}	minimum-energy inter-atom distance
k_{UB}	Urey-Bradley constant for atoms separated by two bonds	q_i, q_j	charge on atoms i and j
		ε_{in}	protein and ligand dielectric constant = 1.0
S	distance between atoms separated by two bonds	ε_{out}	water dielectric constant = 78.4
S_0	equilibrium distance	GB()	generalized-Born solvation energy
k_θ	spring constant for bond angle	a_i, a_j	generalized-Born radii of atoms i and j
θ	bond angle	κ	inverse Debye-Hückel length (salt screening length)
θ_0	equilibrium bond angle	k_{SASA}	microscopic surface tension of water[95] = 0.0072 kcal/mol/Å^2
k_χ, n, δ	Fourier series terms for periodic barrier to rotation around bonds	$SASA$	solvent-accessible surface area (the area traced out by the center of a spherical probe touching the protein's VDW surface); calculated using a water probe radius of 1.4 Å
χ	torsion angle		
k_φ	spring constant for torsion angle to restrain planar groups		
φ	torsion angle	$U_{deprot.}$	deprotonation energy (from a thermodynamic cycle based on the pK_A's of free amino acids)
φ_0	equilibrium torsion angle		

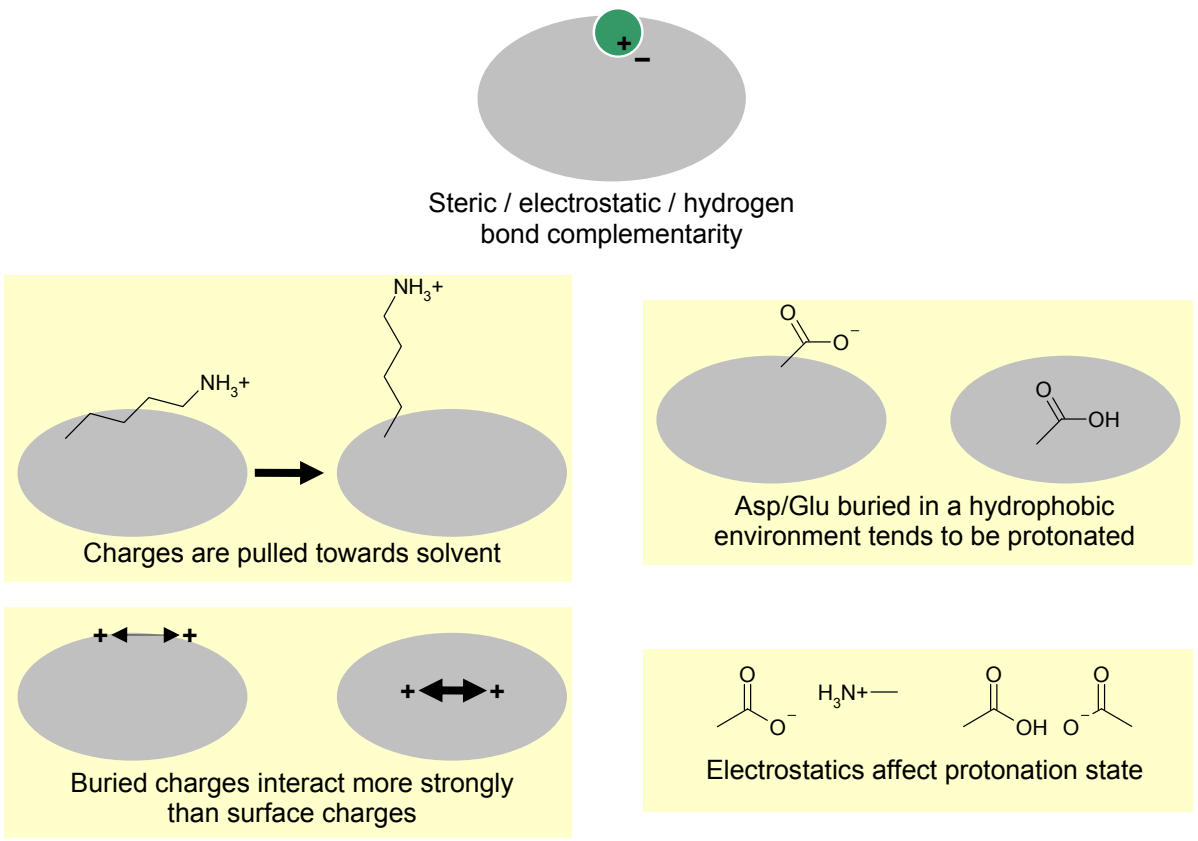


Figure 17. Examples of protein behaviors treated by our model. Factors typically ignored in design calculations are highlighted.

Notes

van der Waals energy: r_{min} for AB interaction is the arithmetic mean of r_{min} for AA and BB interactions. E_{VDW} for AB interaction is the geometric mean of E_{VDW} for AA and BB interactions. Bonded and 1,3 atoms (atoms separated by two bonds) are excluded from this sum.

Coulomb electrostatics: Bonded and 1,3 atoms are excluded from this sum.

Generalized Born solvation energy: All pairs of atoms are included in this sum (including self). Each non-self pair occurs twice in the sum.

Capping: The VDW energy was capped at 2000 kcal/mol/atom pair, and the total electrostatic energy (Coulomb plus generalized Born) was capped at ±1000 kcal/mol/atom pair to prevent floating point overflow of Boltzman weights. In well-packed structures, no interaction energies exceeded the caps.

Hydrogen bonds: These are treated as a combination of electrostatics and van der Waals interactions.

Distance cutoff: None.

Critical parameters

van der Waals: These should not be modified from the values in CHARMM22. In the design literature, van der Waals parameters are frequently stretched or scaled so as not to penalize small steric clashes resulting from limited sampling resolution. However, we've found that this has the side effect of making hydrogen bonds and salt bridges appear stronger than they actually are (see Figure 28 and Table 10).

Internal dielectric constant: We found that an internal dielectric constant of 1.0, which is the default for CHARMM22, produced the most accurate energies. In the design literature, the internal dielectric constant is frequently set to values between 4 and 20, to account for rearrangements in the rest of the protein, or to scale down the Coulomb energy in the absence of a good solvent model. We've found that this is unnecessary (and actually harmful) when the relevant residues are modeled explicitly, and a good solvent model is used.

Water

Our model treats water as a continuum dielectric with salt and surface tension. In a vacuum, the partial charges on the protein atoms interact through Coulomb's law. When we put the protein in water, there is an additional solvation energy term. Part of the solvation energy is roughly proportional to the surface area: VDW interactions with solvent, and the entropy and enthalpy of rearrangment of water molecules at a surface (the hydrophobic effect). The rest of the solvation energy is due to partial charges in the protein interacting with induced surface charges and ion clouds in the solvent ($\Delta G_{polarization}$). Charged atoms closer to the protein's surface have more favorable solvation energy and smaller charge-charge interactions. These energies are calculated using the generalized Born equation.

Generalized-Born energy

Atomic partial charges in a protein reorient water dipoles, inducing surface charges that interact favorably with the partial charges in the protein, and that screen

Coulombic interactions within the protein. Salt forms a counter-ion atmosphere around the protein that neutralizes charge over the Debye-Hückel length. We calculated the interaction energy of the protein with these induced solvent charges using the generalized-Born equation,[15] which provides an approximate solution to the Poisson-Boltzmann differential equation.[44]

The generalized-Born approach requires the calculation of generalized-Born radii for each atom (Figure 18). The manuscript compares two numerical approaches for obtaining the radii. In the first approach, generalized-Born radii are computed on the basis of an r^{-4}-weighted spatial integral (Figure 19):

$$a_i = 4\pi\left(\int_{solvent}\frac{1}{r^4}dV\right)^{-1}.$$

Here r is the distance from the atom center to each volume element in the integrand. The $1/r^4$ in this equation comes from the fact that the energy of a charge–induced dipole interaction (partial charge in the protein interacting with water) is $1/r^4$. Alternatively, more accurate radii are obtained from an empirical sum of r^{-4}- and r^{-5}-weighted spatial integrals:[16]

$$a_i = 4\pi\left(-\int_{solvent}\frac{1}{r^4}dV + P\sqrt{4\pi\int_{solvent}\frac{1}{r^5}dV}\right)^{-1},$$

where P=3.0. The integrals were performed on a rectangular grid (0.5 Å resolution) with the dielectric boundary defined as the molecular surface. Grid points were assigned to solvent if they were contained within a solvent sphere (1.4 Å) centered on a grid point outside the solvent-accessible volume of the protein. For design calculations, the molecular surface was initialized using the crystal structure of the scaffold protein, and was iteratively updated using an average of the currently optimal structures. Final energy evaluations on minimized structures used the exact molecular surface. Formulas in the Appendix give values for the spatial integrals from the grid boundary to infinity. A simpler alternative for integrating the solvent on a grid might be to analytically integrate outside the spherical atom, then subtract the protein regions on the grid outside the atom.

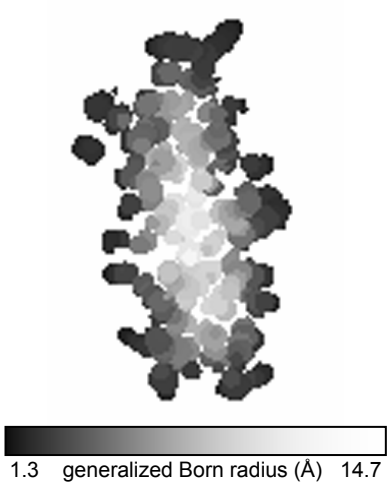


Figure 18. Slice through ribose binding protein, showing generalized Born radii. The radii correlate with atom burial.

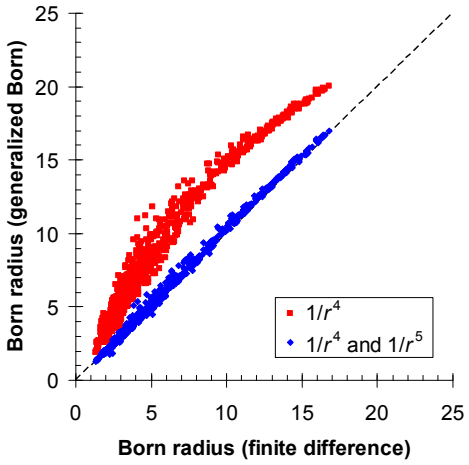


Figure 19. Comparison of generalized Born radii for protein tyrosine phosphatase 1B calculated using an integral formula (y-axis) with radii calculated using a finite-difference approach (x-axis).
Similar results were reported in [16].

After calculating generalized Born radii for each atom, we can calculate the solvation energy using the generalized Born equation:

$$\Delta G_{polarization} = \sum_{i,j} \mathrm{GB}(q_i, q_j, a_i, a_j, r, \varepsilon_{in}, \varepsilon_{out}, \kappa) = -\frac{1}{2}\left(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{out}}\right)\sum_{i,j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + a_i a_j e^{-r_{ij}^2/(4a_i a_j)}}}$$

This equation gives exact answers for the limiting cases of very close and very distant charges, and interpolates between these two extremes. In the limit of $a_1 = a_2 >> r$, the generalized Born equation calculates a solvation energy of:

$$\Delta G_{polarization} = -\frac{q^2}{2a}\left(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{out}}\right),$$

which is called the Born equation. And in the limit of $r >> a$, the generalized Born equation plus the Coulomb term gives an interaction energy of $\frac{q_1 q_2}{\varepsilon_{out} r}$, which is also correct.

The generalized Born equation can be modified to handle salt as well:[104]

$$\sum_{i,j} \mathrm{GB}(q_i, q_j, a_i, a_j, r, \varepsilon_{in}, \varepsilon_{out}, \kappa) = -\frac{1}{2}\sum_{i,j}\left(\frac{1}{\varepsilon_{in}} - \frac{e^{-\kappa\sqrt{r_{ij}^2 + a_i a_j e^{-r_{ij}^2/(4a_i a_j)}}}}{\varepsilon_{out}}\right)\frac{q_i q_j}{\sqrt{r_{ij}^2 + a_i a_j e^{-r_{ij}^2/(4a_i a_j)}}},$$

with $\kappa = \frac{\sqrt{I/\varepsilon_{out}}}{0.343}$ at 25°C.

Variables:

κ	inverse Debye-Hückel length in $Å^{-1}$
I	ionic strength in mol/l

A salt concentration of 100 mM was used for the calculations reported here (Figure 20).

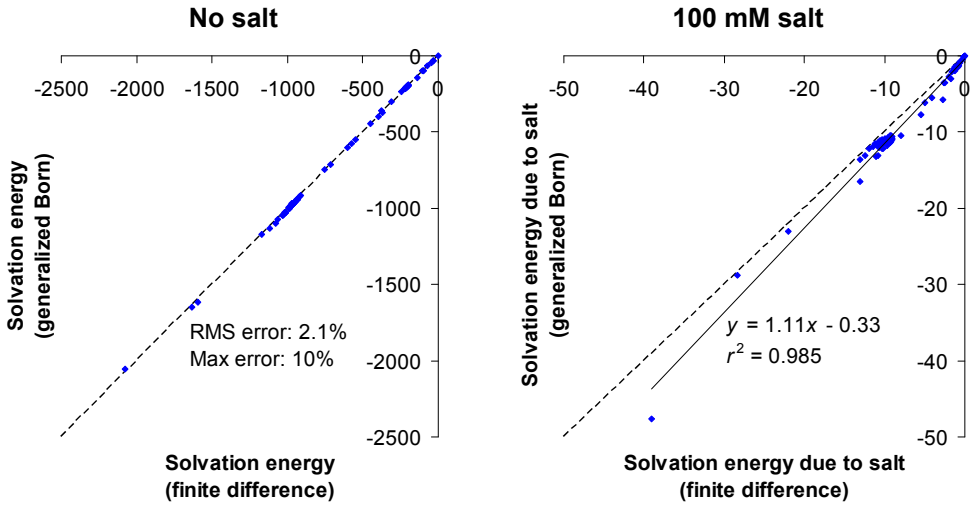


Figure 20. Comparison of solvent polarization energies for a set of small molecules, peptides, and proteins calculated using the generalized-Born approach (y-axis) with values calculated using a finite-difference approach (x-axis).

Pairwise approximation of solvent accessible surface area (SASA)

Following Street and Mayo,[40] we approximated the total SASA as the sum of accessible surface areas for each amino acid within the context of the fixed structural

elements of the design, less the probability weighted sum of the pairwise surface areas buried by each variable structural element of the design (for example a rotamer or a ligand pose). The pairwise surface areas are scaled to correct for over-counting, which occurs when multiple variable structural elements simultaneously bury one surface patch. The scaling factors were determined by a linear regression that optimized agreement between the pairwise approximation and the exact solvent accessible surface areas of 100,000 random conformations of the protein with random sequences present at the design positions. Optimal values of the scaling factors are highly under-constrained, due to correlations between the various area terms. To address this issue, we used a singular value decomposition[105] to perform the linear regression. Any scaling factors greater than 100 or less than –100 were set to 0, and the regression was repeated without them.

$$\text{SASA (linear regression form)} = \sum_{\substack{i \in \text{variable} \\ \text{position}}} t_i A_i - \sum_{\substack{i \in \text{variable} \\ \text{position}}} s_i \sum_{j \neq i} A_{i,j} - \sum_{\substack{i \in \text{fixed} \\ \text{position}}} s_i \sum_{\substack{j \in \text{variable} \\ \text{position}}} A_{i,j} + C$$

Here, variable positions included the repacked residues and the ligand. The fixed positions were the residues in the protein whose identity and conformation were held fixed during the design. This linear regression form can be rearranged into a pairwise factorable form.

SASA (pairwise form) =

$$C + \sum_{\substack{i \in \text{variable} \\ \text{position}}} \left(t_i A_i - s_i \sum_{\substack{j \in \text{fixed} \\ \text{position}}} A_{i,j} - \sum_{\substack{j \in \text{fixed} \\ \text{position}}} s_j A_{j,i} \right) - \sum_{\substack{i \in \text{variable} \\ \text{position}}} \sum_{\substack{j \in \text{variable} \\ \text{position}, j<i}} \left(s_i A_{i,j} + s_j A_{j,i} \right)$$

C — additive constant

$t_i A_i$ — SASA of rotamers and ligand poses

$s_i \sum_{j \in \text{fixed position}} A_{i,j}$ — less the pairwise area buried at interfaces with fixed structural elements

$\sum_{j \in \text{fixed position}} s_j A_{j,i}$ — pairwise area buried at interfaces between variable structural elements

Variables:

A_i the accessible surface area of a rotamer, pose or fixed conformation at position i within the context of the fixed structural elements of the design.

t_i scaling factors for accessible surface areas of rotamers or poses

$A_{i,j}$ The portion of A_i buried by the variable rotamer or pose at position j within the context of the fixed structural elements of the design.

s_i scaling factors for pairwise buried areas

The interfacial solvation energy is the product of the SASA and a microscopic surface tension of 7.2 cal/mol/Å^2 [95]. The "hydrophobic effect" driving aggregation of hydrophobic solutes in water increases in proportion to solute surface area with a slope[39] of 24 cal/mol/Å^2. This slope is reconciled with the 7.2 cal/mol/Å^2 microscopic surface tension by adding the van der Waals interaction energy between explicitly modeled hydrophobic solutes, which evaluates to roughly 17 cal/mol/Å^2 for CHARMM22.

Deprotonation energy

The structural calculations reported here modeled the pH- and environment-dependent titration of histidine and the acidic amino acids. The doubly protonated and two singly protonated states of histidine, and the protonated and deprotonated states of aspartate and glutamate were modeled as independent rotamers. Because molecular-mechanics potentials do not treat changes in covalent bonding, the

energy difference between protonated and deprotonated rotamers was computed using a thermodynamic cycle (Figure 21). For example, the deprotonation energy for an aspartate residue within a protein (labeled A in Figure 21) was determined indirectly by summing two transfer free energies (B and D) and the experimentally measured free energy for deprotonation of acetylated asparate amide in free solution (C). Free energies for the small-molecule aspartate derivatives were obtained by building a complete set of aspartate side-chain rotamers onto each member of an amino-acid backbone ensemble, evaluating the energy of each configuration, and computing the free energy as:

$U(\text{solution}) = -RT \ln(\text{partition sum})$.

Then:

$B = U(\text{AspH, solution}) - E(\text{AspH, protein})$

$C = -2.3RT*(\text{pH} - \text{pK}_a)$

$D = E(\text{Asp}^-, \text{protein}) - U(\text{Asp}^-, \text{solution})$

where U is free energy and E is potential energy. Adding these together:

$A = B + C + D = [E(\text{Asp}^-, \text{protein}) - E(\text{AspH, protein})] + [-U(\text{Asp}^-, \text{solution}) + U(\text{AspH, solution}) - 2.3RT*(\text{pH} - \text{pK}_a)]$

We denote the terms within the right bracket above, $-U(\mathrm{Asp}^{-}, \text{solution}) + U(\text{AspH}, \text{solution}) - 2.3RT^{*}(\mathrm{pH} - \mathrm{pK_a})$, as the deprotonation energy. It is added to the self-energy of each deprotonated rotamer to establish the appropriate energy relationship between the deprotonated and protonated forms of the amino acid (Table 4). The deprotonation energy is pH dependent, and all of the calculations reported here were performed at pH 7.0.

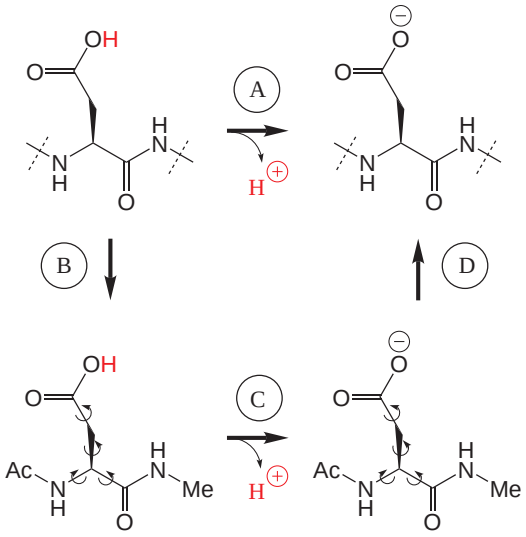


Figure 21. Thermodynamic cycle used to evaluate the deprotonation energy for aspartate (A).

The dashed lines in the top structures represent bonds to the complete polypeptide chain of the protein, which is not shown. The bottom structures depict N-acetyl, N'-methyl aspartate α-amide in its protonated and deprotonated forms. The rotational arrows on the structures at the bottom indicate that they are modeled as a structural ensemble, whereas the structures at the top are single rotamers. The deprotonation energy is calculated as the sum of two transfer energies (B and D) and the experimentally-measured free energy for protonation of the acetyl-aspartate amide (C).

Amino acid	Deprotonation energy
HSP	0
HSD	–23.19 – 1.36 (pH – 6.74)
HSE	–2.53 – 1.36 (pH – 6.14)
ASP	37.21 – 1.36 (pH – 3.71)
APP	0
GLU	41.64 – 1.36 (pH – 4.15)
GUP	0

Table 4. Deprotonation energies for the titratable amino acids in the 6028-member rotamer library.
Experimental pKa values for free amino acids are from ref. [106,107]. We did not include protonation states for CYS, TYR, LYS, or ARG because of a lack of published CHARMM22 parameters for those amino acids. 1.36 = RT ln 10 at T = 25°C.

Discrete sampling

Protein scaffold coordinates

Hydrogen coordinates were added to scaffold crystal structures using Reduce.[108]

Selection of design positions

For ABP, all side chains where the van der Waals spheres were within 1 Å of the ligand van der Waals spheres in any of four crystal structures (8ABP, 6ABP, 1ABE, 5ABP) were selected as design positions. For RBP, hydrogen bonding and hydrophobic contacts determined by the program HBPLUS[109] were selected as design positions. The resulting positions are listed in the caption to Figure 10.

For Avastin-VEGF, the repacked residues were hand picked, because with our current level of computer power, we were unable to model all interface residues at high resolution. Starting with the 6 Fab positions where mutations have been reported to improve the affinity, we then added side chains (except ALA, GLY, PRO) in Fab and VEGF contacting side chains at these 6 positions, and also included positions that showed a high conformational variability among different crystal structures (1BJ1, 1CZ8, 1FLT, 1KAT, 1QTY, 1TZH, 1TZI, 1VPP, 2VPF). The resulting positions are listed in the caption to Figure 11.

Rotamer library

A detailed rotamer library (including polar and non-polar hydrogens) was created by clustering the side chain conformations seen in high-resolution crystal structures (Table 5). Starting with the 18528 structures in Protein Data Bank Release #101 (July 2002), we removed theoretical models, structures with resolution > 1.9 Å, structures with a CAVEAT record, and structures with ≤ 10% of atoms in one of the 20 natural amino acids. This resulted in a list of 7312 structures. Hydrogens were added to each structure using Reduce[108] from the Richardson lab. The side chain conformations for each amino acid were then clustered at the resolution listed in Table 5. The clustering process involved selecting the conformation with the most close neighbors, discarding all neighbors (defined by an RMS cutoff), and repeating until a predetermined fraction of the conformations had been covered. Finally, each rotamer was locally minimized with a constraint of ± 1° on each dihedral angle. No rotamer in

the library corresponds to any of the crystallographic cooordinates of ABP, RBP, or Avastin-VEGF.

For repacking calculations, rotamers were placed at each variable position of the protein scaffold, and energy minimized using dihedral restraints and no electrostatics. The energy minimization slightly adjusted bond lengths and angles to match the equilibrium values in CHARMM22. Rotamers with energies more than 15 kcal/mol over the lowest energy rotamer of the same amino acid at the same position were eliminated.

Amino acid	Number of rotamers	Neighbor RMS cutoff (Å)	Close neighbor RMS cutoff (Å)	Coverage
ALA	3	0.5	0.3	0.999
APP	141	0.5	0.3	0.999
ARG	974	1.0	0.4	0.98
ASN	132	0.5	0.3	0.999
ASP	62	0.5	0.3	0.999
CYS	29	0.5	0.3	0.999
CYX	8	0.5	0.3	0.999
GLN	758	0.5	0.3	0.999
GLU	412	0.5	0.3	0.999
GLY	1	0.5	0.3	0.999
GUP	649	0.5	0.3	0.999
HSD	233	0.5	0.3	0.999
HSE	255	0.5	0.3	0.999
HSP	245	0.5	0.3	0.999
ILE	215	0.5	0.3	0.999
LEU	325	0.5	0.3	0.999
LYS	400	1.0	0.4	0.98
MET	181	0.8	0.4	0.99
PHE	193	0.5	0.3	0.999
PRO	8	0.5	0.3	0.999
SER	32	0.5	0.3	0.999
THR	64	0.5	0.3	0.999
TRP	238	0.6	0.3	0.99
TYR	414	0.5	0.3	0.999
VAL	56	0.5	0.3	0.999
Total	6028			

Table 5. The highest resolution rotamer library with 6028 rotamers.

APP = protonated Asp, GUP = protonated Glu, HSP = doubly protonated His, HSD = His protonated on the delta nitrogen, HSE = His protonated on the epsilon nitrogen, CYX = disulfide-bonded cysteine.

Ligand poses

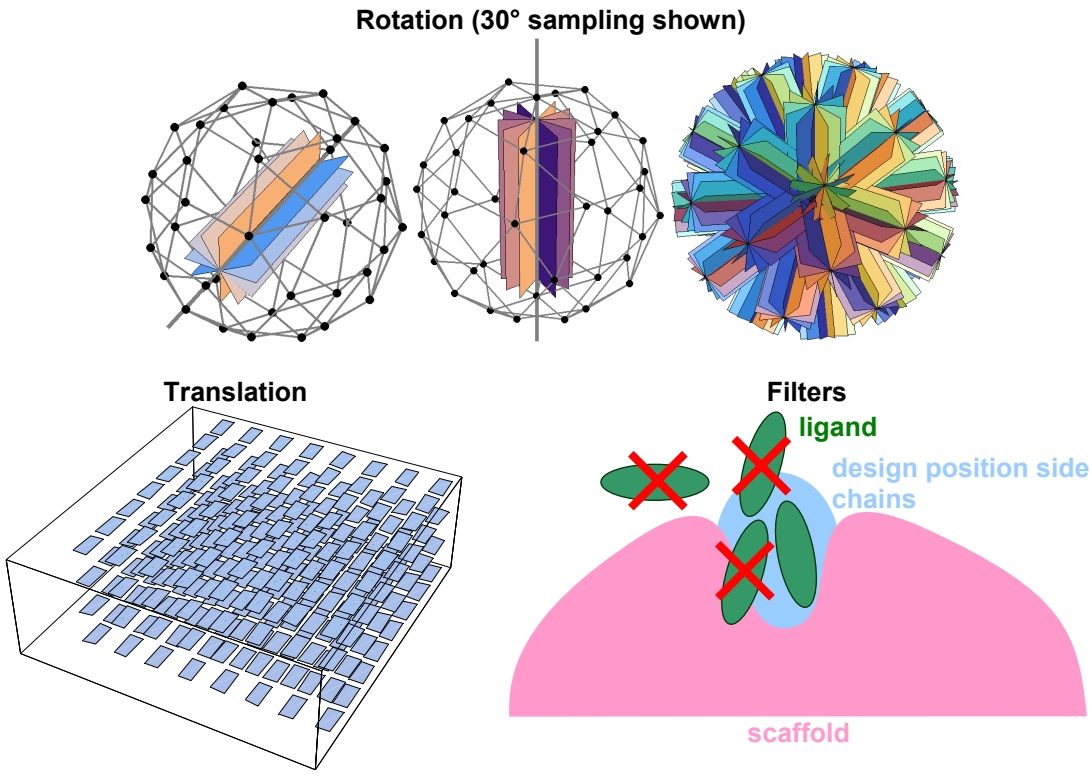


Figure 22. Ligand sampling and filters.
Ligand poses were identified by generating conformers of the ligand, and then exploring rotational and translational degrees of freedom. A series of filters was applied to identify poses that overlapped well with the-side chain regions of the design positions but not with the fixed portions of the scaffold, and that exhibited energies within 10 kcal/mol of the isolated ligand.

A series of 26 ribose and 19 arabinose conformational isomers were generated to sample the internal degrees of freedom of the two sugars. The crystal structure coordinates were not included. The 19 arabinose rotamers were generated by starting

with the two chair flip conformations of the α and β anomers of the pyranose. Each of these 4 ring conformations adopts 3^4 hydroxyl rotamers, for a total of 324 rotamers. We did not include furanose, aldehyde, or boat conformations. We calculated the CHARMM22 energy of each conformation using TINKER, including a GBSA energy term.[92] Finally, we applied a 6 kcal/mol cutoff above the lowest energy conformation, and then clustered the remaining conformations at 0.5 Å resolution. The clustering process involved selecting the lowest energy conformation, discarding all conformations within 0.5 Å RMS of this conformation, and repeating until no conformations were left. The 26 ribose rotamers were generated the same way, except that an 8 kcal/mol energy cutoff was applied.

These isomers were then rotated in 10° increments along axes defined by a triangulated icosahedron, producing 6516 rotational orientations. Using a fast Fourier transform algorithm,[110] the internal/rotational ensemble was translated along a 0.5 Å grid to find poses that overlapped well with the side-chain regions of the design positions but not with fixed regions of the scaffold. (Figure 22). The energies of poses in this subset, excluding the electrostatic energy, were evaluated. Poses with energies exceeding the energy of the isolated ligand by more than 10 kcal/mol were discarded. The remaining poses were clustered at 0.5 Å resolution to generate the set of poses used for repacking and design calculations.

Structural optimization

We optimize the bound, unbound, and unfolded states separately. The advantage of this multi-state framework[9] is that we can predict conformational changes upon binding, and also optimize for the desired combination of stability, affinity, and specificity. In contrast, algorithms that optimize a single target structure, such as dead end elimination,[34,111] do not distinguish between intramolecular and intermolecular interactions, and thus will propose mutations that stabilize the protein without improving its interaction with the ligand.[61]

We break the protein/ligand system into several parts, identify the low energy conformations of each part, then precompute all the intrinsic energies and interaction energy matrices. This allows us to quickly recalculate the binding energy for different amino acid sequences and conformations — all the energy terms have already been precomputed.

Rotamer probabilities were either initialized randomly, or set to 0's and 1's to match a single structure generated by simulated annealing or by the FASTER procedure.[112] Using a mean-field algorithm, the probabilities were then adjusted iteratively to minimize the free energy of the system.[19] New probabilities for all rotamers were first computed using the mean-field energy of each rotamer and the Boltzmann equation:

$$p_{new} = \frac{\exp(-(E_{self} + E_{interaction})/RT)}{Z}.$$

Here, Z normalizes the probabilities at a single position so that they sum to one. To prevent oscillating probabilities that do not converge, we updated probabilities with the geometric mean of the old and new values:

$$p_{updated} = \begin{cases} 0, \text{ if } p_{old} < m \text{ and } p_{new} < m \\ rp_{new}, \text{ if } p_{old} < m \\ rp_{old}, \text{ if } p_{new} < m \\ \sqrt{p_{old} p_{new}}, \text{ otherwise} \end{cases}$$

where r is a random number between 0 and 0.5, and m is the smallest positive single-precision floating point number (~1.18×10^{-38}). $p_{updated}$ must be normalized after this procedure. Alternatively, we updated one position at a time in random order, without any probability averaging. The repacking procedure was repeated 10 to 1000 times, using different initial rotamer probabilities. Two-thirds of the repacking runs used the single site update method, and the remainder were run using the simultaneous update method.

The mean field calculation is good at jumping over local barriers that stymie gradient-based minimization or molecular dynamics, because there's no barrier for sidechains flipping to completely different rotamer configurations. However, there *is* a barrier for multiple simultaneous rotamer changes, so the calculation must be repeated from different starting probabilities.

The most probable structure from the lowest energy mean-field solution was subjected to a final local minimization step. Thus, we discretely sampled a rough energy landscape to identify the lowest-lying energy well, and locally minimized to

get to its bottom (Figure 23). The calculated side-chain conformational entropy for different sequences typically varied by less than one kcal/mol, which is small relative to the other energy terms. Hu and Kuhlman also observed that side-chain conformational entropy makes small contributions in their design calculations.[113] However, it is important to note that we did not include entropy changes outside the binding site in our calculation.

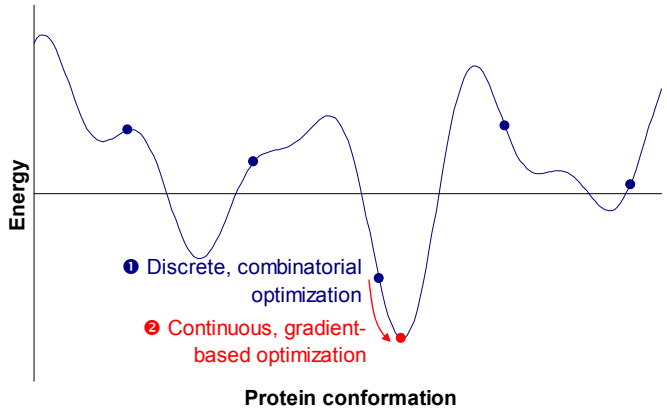


Figure 23. Discrete then continuous optimization of protein structure.

Unfolded state

The intrinsic unfolded-state chemical potential for each amino acid was determined by placing a complete rotamer set at the middle position of an ALA-ALA-ALA tripeptide library comprising multiple peptide backbone conformations with no termini (similar to the approach in [99]). The energy of each configuration was calculated, and the intrinsic unfolded-state chemical potential (Table 6) was evaluated as RT ln(partition sum).

Inter-residue electrostatic interaction energies in the unfolded state were calculated following [58], assuming that the distance distribution between residues is determined by a random walk. The total unfolded-state energy was summed as:

Unfolded state energy =

$$\overbrace{\sum_i \mu(aa_i)}^{\text{Intrinsic unfolded-state chemical potential}} + 332 * \overbrace{\sum_{i<j} \frac{q_i q_j (\sqrt{6/\pi} - \kappa d \exp(\kappa^2 d^2/6)\operatorname{erfc}(\kappa d/\sqrt{6}))}{\varepsilon_{out} d}}^{\text{Inter-residue electrostatic interaction (Gaussian chain model)}}$$

with $d = b_{eff}\sqrt{i-j} + s$.

Variables:

$\mu(aa_i)$	intrinsic chemical potential of am. acid at position i	b_{eff}	effective bond length = 7.5 Å
q_i	charge of the amino acid at position i	s	distance offset = 5 Å
d	RMS inter-residue distance	κ	inverse Debye-Hückel length in $Å^{-1}$

Amino acid	Intrinsic μ (kcal/mol)
ALA	1.39
ARG	–272.99
ASN	–78.70
ASP	–110.07
CYS	1.82
GLN	–57.51
GLU	–86.71
GLY	–8.67
HIS	–44.33
ILE	6.12
LEU	–12.49
LYS	–62.29
MET	–1.62
PHE	6.09
PRO	25.48
SER	5.38
THR	–15.83
TRP	7.68
TYR	–10.22
VAL	1.17

Table 6. Intrinsic unfolded-state chemical potentials for the amino acids in the 6028-member rotamer library.

Sequence optimization (genetic algorithm)

For sequence design, a random population of sequences was initially chosen. Putative energies and structures for each sequence were calculated as described above. The population was then ranked by computed ligand affinity, with a limit on allowable protein destabilization (10 kcal/mol in the initial generations, and 5 kcal/mol in the final generations). The top ranked sequences were mutated and recombined to generate a child population. This evolutionary procedure was iterated until functional improvements ceased to occur. (See Figure 24) We started with a high mutation rate (0.25 mutation probability per position) and low selection stringency (tournament

selection where the best of 4 randomly picked sequences is a parent for the next generation). As the population converged, we decreased the mutation rate to 0.15 and increased the selection stringency to tournament selections with 5 – 8 sequences. See Table 7 for details.

Calc phase	Generations	Seqs/gen*	Tournament	Mutation	Destab. (kcal/mol)
1	23	200	4	0.25	10
2	21	200	8	0.2	10
3	21	200	5	0.2	5
4	21	200	5	0.15	5

* The initial generation of calculation phases 1 – 3 had between 175 and 224 sequences, depending on how many top sequences were included from the previous phase. The initial generation of calculation phase 4 had 844 sequences, which included all point mutants of the top 3 sequences, double mutants of the top sequence, and random recombinants of the top sequences.

Table 7. Genetic algorithm parameters.

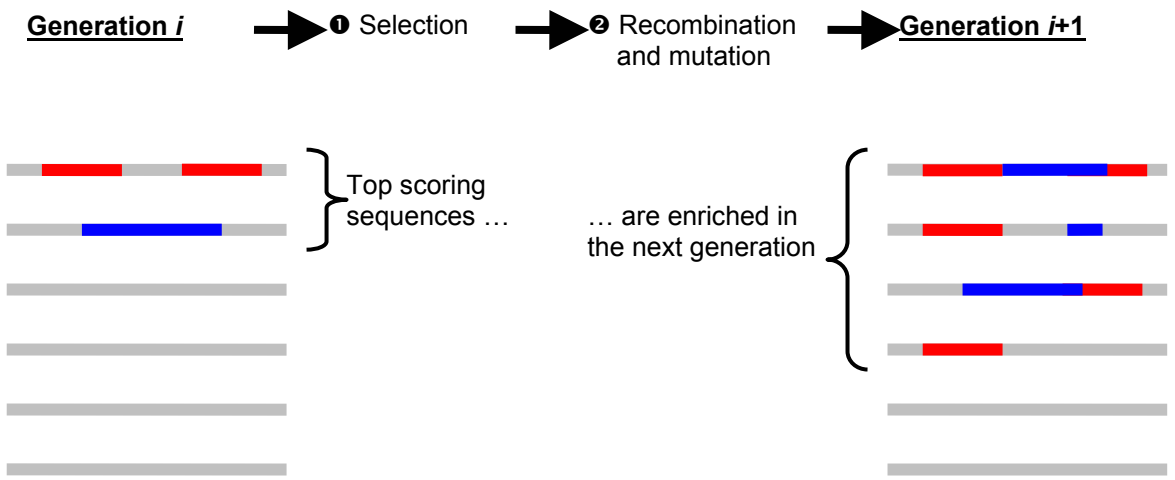


Figure 24. Genetic algorithm.

Appendix: Integrals

To calculate generalized Born radii, we integrated r^{-4} or r^{-5} outside the rectangular region $x_1 < x < x_2$, $y_1 < y < y_2$, $z_1 < z < z_2$ using these formulas:

$$\iiint_{\substack{\text{outside}\\ \text{rectangular}\\ \text{region}}} \frac{dx\,dy\,dz}{(x^2+y^2+z^2)^2} = \sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{2}\left((-1)^{j-k+1} \begin{pmatrix} g_4(x_j, y_k, z_1, z_2) & \text{if } i=1 \\ g_4(x_j, z_k, y_1, y_2) & \text{if } i=2 \\ g_4(y_j, z_k, x_1, x_2) & \text{if } i=3 \end{pmatrix} \right),$$

$$\text{with } g_4(a,b,c_1,c_2) = \frac{\sqrt{a^2+b^2}\left(\tan^{-1}\left(\frac{c_1}{\sqrt{a^2+b^2}} \right) - \tan^{-1}\left(\frac{c_2}{\sqrt{a^2+b^2}} \right) \right)}{2ab}$$

$$\iiint_{\substack{\text{outside}\\ \text{rectangular}\\ \text{region}}} \frac{dx\,dy\,dz}{(x^2+y^2+z^2)^{5/2}}$$

$$= \frac{1}{6}\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\left((-1)^{i+j+k} \frac{\sqrt{x_i^2+y_j^2+z_k^2}}{x_i y_j z_k} \right)$$

$$+ \frac{1}{6}\sum_{h=1}^{3}\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\left((-1)^{i+j+k} \begin{pmatrix} g_5(x_i, y_j, z_k) & \text{if } h=1 \\ g_5(x_i, z_j, y_k) & \text{if } h=2 \\ g_5(y_i, z_j, x_k) & \text{if } h=3 \end{pmatrix} \right),$$

$$\text{with } g_5(a,b,c) = \frac{\tan^{-1}\left(\frac{ab}{c\sqrt{a^2+b^2+c^2}} \right)}{c^2}.$$