Supplementary Materials

S.1 Features Analysis Configuration

This section describes details of how the features analysis was conducted.

General workflow

1) Preparation of Native Data
2) Generation of Relax Structures for each energy function
3) Generation of Features with Talaris2013 H-Bonds for each energy function
4) Create Features Analysis plots

S.1.1) Preparation of Native Data

a) Begin with the Top8000 set (http://kinemage.biochem.duke.edu/databases/top8000.php), curated by the Richardson Lab from structures deposited into the Protein Databank (Berman et al., 2000) filtered at the %70 percent homology level and filtered for having electron density structures deposited into the electron density server (http://eds.bmc.uu.se/eds/), this yielded ~6k protein chains, ~1.6M residues, and ~1.3M H-bonds. Since crystal structures usually lack hydrogen atoms, we used the program Reduce (Word, et al. 1999) to add them. Critically, Reduce makes minimal assumptions about where to place hydrogen atoms, defining most of their coordinates from ideal bond angles and lengths alone. For ambiguously located hydrogen atoms, e.g., the hydroxyl *HG* atom in serine, it samples a set of rotations for the hydroxyl hydrogen based on the location of nearby acceptor groups (as opposed to sampling a predetermined set of dihedral angles such as at $60°$, $-60°$, and $180°$) and scores hydrogen bond contacts using a metric based on sphere overlap for both favorable interactions between donor and acceptor atoms and unfavorable interactions between colliding atoms (as opposed to making assumptions based on charges or preferred hydrogen-bonding angles or dihedrals).

b) Convert names of hydrogens to Rosetta's naming convention:

```
python rosetta/main/tools/convert_hatm_names.py \
   --data_dir top8000_chains_eds_70 \
   --output_dir top8000_chains_eds_70_rosetta_named_hydrogens
```

c) The top8000 set is specifies a single chain from each PDB, these were extracted and put into separate pdb files prior to use.

S.1.2) Generate Relax structures for each energy function

For each Score function, we applied the FastRelax protocol to each structure to sample near native conformations. The protocol iterates between repacking sidechains and performing quasi-Newton minimization of torsional degrees of freedom while ramping in five steps the strength of the repulsive component of the Lennard-Jones term from 1/10th up to full strength, cycling from low- to high-strength repulsion three times. We chose this sampling protocol over less aggressive protocols (such as only performing minimization) to allow the structures to escape local minima in the energy landscape, and over more aggressive protocols (such as trying to fold proteins from extended chains using Rosetta's AbRelax protocol (Bradley, Misura, & Baker, 2005)) in the interest of focusing our efforts on matching distributions where native-like contacts are possible: an *ab initio* folding protocol might emphasize easy-to-form local contacts over harder-to-form long-range contacts and misrepresent the deficiencies in an energy function.

a) Prepare the following input files:

i) bsub command
ii) condor submission script
iii) top8000 pdbs and .list file (listing the relative paths to the pdb files)
iv) flags file
v) features.xml RosettaScript

In what follows, <sample_source_id> has the following form:

```
<data>_<protocol>_<sample_source>_r<revision>_<date>
```

where revision is the git revision identifier of the repository and the date format is YYMMDD so it can be sorted alphabetically. For example:

```
top8000_relax_ElecHBv2_rD28828_130301
```

i) For this study we used the KillDevil LSF cluster at UNC Chapel Hill (http://its.unc.edu/research/its-research-computing/computing-resources/killdevil/).

bsub command:

```
bsub \
    -q day \
    -n 256 \
    -J features_<sample_source_id> \
    -o features_<sample_source_id>_%J.log \
    -e features_<sample_source_id>_%J.err \
    -a mvapich mpirun \
    <path>/Rosetta/main/source/bin/rosetta_scripts.mpi.linuxgccrelease \
    -database <path>/Rosetta/main/database \
    -out:mpi_tracer_to_file features_<sample_source_id>.log \
    @flags
```

ii) condor submission script:

```
universe = vanilla
Notify_user  =
notification = Error
Log = features_<sample_source_id>.condor.log
Executable = rosetta_scripts.mpi.linuxgccrelease
Requirements = ( Memory > 512)
GetEnv = True
Error = fetures_.condor.error.log
Output = features_<sample_source_id>.condor.output.log
arguments = -database <path>/rosetta/Rosetta/main/database @flags
priority = -10
queue 256
```

iii) The following options were used (broken into sections for readability)

General flags:

```
-options:user
-remember_unrecognized_res
-remember_unrecognized_water
-no_optH
-jd2:delete_old_poses
-mute protocols.jd2
-mute core.io.pdb.file_data
```

```
-mute core.scoring.etable
-mute core.io.database
-mute core.scoring.ScoreFunctionFactory
-mute core.pack.task
-mute protocols.ProteinInterfaceDesign.DockDesign
```

Input flags:

```
-in:path top8000_chains_eds_70_rosetta_named_hydrogens
-in:file:l top8000_chains_eds_70_rosetta_named_hydrogens/all_pdbs.list
```

Output flags:

```
-out:nooutput
```

Protocol flags:

```
-parser:protocol features.xml
```

iv) The following xml script is used to apply the fast relax protocol to each input structure and extract features into the features database.

Each energy function will have different subtags for the score function s described below.

The init_struct mover is used to store a reference to the original structure for the RMSD to native features reporter.

Features.xml:

```
<ROSETTASCRIPTS>
  <SCOREFXNS>
    <s weights=score12prime>
      <Reweight scoretype=scoretype weight=weight/>
      ...
    </s>
  </SCOREFXNS>
  <MOVERS>
    <SavePoseMover name=init_struct reference_name=init_struct/>
    <FastRelax name=fast_relax scorefxn=s/>
    <ReportToDB
      name=features_reporter
      database_name="features.db3"
      database_mode=sqlite3
      database_separate_db_per_mpi_process=1
      batch_description="<data> <protocol> <sample_source>">

      <feature name=ScoreTypeFeatures/>
      <feature name=StructureScoresFeatures scorefxn=s/>
      <feature name=PoseCommentsFeatures/>
      <feature name=PoseConformationFeatures/>
      <feature name=ProteinRMSDFeatures reference_name=init_struct/>
      <feature name=RadiusOfGyrationFeatures/>
      <feature name=ResidueTypesFeatures/>
      <feature name=ResidueFeatures/>
      <feature name=PdbDataFeatures/>
      <feature name=UnrecognizedAtomFeatures/>
      <feature name=PairFeatures/>
      <feature name=ResidueBurialFeatures/>
      <feature name=ResidueSecondaryStructureFeatures/>
      <feature name=ProteinBackboneAtomAtomPairFeatures/>
      <feature name=ProteinBackboneTorsionAngleFeatures/>
      <feature name=ProteinResidueConformationFeatures/>
      <feature name=HBondFeatures scorefxn=s/>
      <feature name=HBondParameterFeatures scorefxn=s/>
      <feature name=SaltBridgeFeatures/>
      <feature name=ChargeChargeFeatures/>
    </ReportToDB>
  </MOVERS>
  <PROTOCOLS>
    <Add mover_name=init_struct/>
    <Add mover_name=fast_relax/>
    <Add mover_name=features_reporter/>
  </PROTOCOLS>
</ROSETTASCRIPTS>
```

v) At the completion of the run, there will be 265 SQLite3 database files named `features.db3_###`, each containing the features extracted from the structures run a specific node in the cluster. These were then merged together using

```
<path>/Rosetta/main/tests/features/sample_sources/merge.sh \
     features.db3 \
     features.db3_*
```

S.1.3) Generation of Features for each energy function

To make the comparison between H-bond features is important to have a consistent definition of what constitutes an H-bond. We used the ElecHBv2 definition for features and statistics described in this this paper. So to generate features using the ElecHBv2 Energy function we ran the following features extraction protocol using the features database generated in step 2 as input:

On the command line

```
# in directory containing features.db3 extracted after relaxing natives
mkdir ElecHBv2; cd ElecHBv2
<path>/Rosetta/main/source/bin/rosetta_scripts.<platform> \
    -database <path>/Rosetta/main/database @flags &> log
```

with the following support files

flags:

```
-options:user
-remember_unrecognized_res
-remember_unrecognized_water
-no_optH
-jd2:delete_old_poses
-inout:dbms:database_name ../features.db3
-in:use_database
-out:nooutput
-parser:protocol features.xml
```

features.xml: Note no specific score function weighting for flags are necessary because the ElecHBv2 is consistent with the Talaris2013 energy function (default May 2013) with the caveat that the HBond energy weight in ElecHBv2 is 80% the weight in Talaris2013, but this does this has no impact on the definition H-bonds.

```
<ROSETTASCRIPTS>
       <MOVERS>
                <ReportToDB
                        name=features_reporter
                        database_name="features.db3"
                        database_mode=sqlite3
                        database_separate_db_per_mpi_process=1
                        batch_description = "Features with Talaris2013">

                        <feature name=ScoreTypeFeatures/>
                        <feature name=StructureScoresFeatures scorefxn=s/>
                        <feature name=PoseCommentsFeatures/>
                        <feature name=PoseConformationFeatures/>
                        <feature name=RadiusOfGyrationFeatures/>

                        <feature name=ResidueTypesFeatures/>
                        <feature name=ResidueFeatures/>
                        <feature name=PdbDataFeatures/>
                        <feature name=UnrecognizedAtomFeatures/>
                        <feature name=PairFeatures/>
                        <feature name=ResidueBurialFeatures/>
                        <feature name=ResidueSecondaryStructureFeatures/>
                        <feature name=ProteinBackboneAtomAtomPairFeatures/>
```

```
                        <feature name=ProteinBackboneTorsionAngleFeatures/>
                        <feature name=ProteinResidueConformationFeatures/>
                        <feature name=HBondFeatures/>
                        <feature name=HBondParameterFeatures/>
                        <feature name=SaltBridgeFeatures/>
                        <feature name=ChargeChargeFeatures/>
                </ReportToDB>
        </MOVERS>
        <PROTOCOLS>
                <Add mover_name=features_reporter/>
        </PROTOCOLS>
</ROSETTASCRIPTS>
```

S.1.4) Create Features Analysis plots

To generate features analysis plots, we call

```
    <path>Rosetta/main/tests/features/compare_sample_sources.R \
        --config analysis_configuration.json
```

with the following configuration file:

```
analysis_configuration.json:
{
    "output_dir" : "build/general_analysis",
    "preparation_sample_source_comparisons" : [
    ],
    "sample_source_comparisons" : [
        {
            "sample_sources" : [
                {
                    "database_path" : "Native/ElecHBv2/features.db3",
                    "id" : "Native",
                    "reference" : true
                },
                {
                    "database_path" : "Elec/ElecHBv2/features.db3",
                    "id" : "Elec",
                    "reference" : false
                },
                {
                    "database_path" : "HBv1/ElecHBv2/features.db3",
                    "id" : "HBv1",
                    "reference" : false
                },
                {
                    "database_path" : "HBv2/ElecHBv2/features.db3",
                    "id" : "HBv2",
                    "reference" : false
                },
                {
                    "database_path" : "ElecHBv2/features.db3",
                    "id" : "ElecHBv2",
                    "reference" : false
                }
            ],
            "analysis_scripts" : [
```

```
"scripts/analysis/plots/hbonds/geo_dim_2d_conditional/sp2_BAH_chi_polar_density_beta_sheet_alpha_he
lix.R",
            ...

            ],
            "output_formats" : [
                "output_print_pdf",
                "output_small_pdf",
                "output_slide_pdf",
                "output_html",
                "output_csv"
            ]
        }
    ]
}
```

S.3 Supplemental Section: *HBv2* HBond Model

Figure S.3.1: *The one-dimensional BA_χ distributions for sp^2-hybridized acceptors by donor type*. Red: The native distribution; Green: *HBv1*; Blue: *HBv2*. Three-letter donor- and acceptor-type names are given in Figure S.4.1. Plots in this section include hydrogen bonds with sequence separation between donor and acceptor greater than 4 (unless otherwise noted). The BA_χ dihedral distributions for sp^2 hybridized acceptors in crystal structures broadly show a sinusoidal pattern consistent with the idea that donor H-atoms would seek out the sp^2 plane. The sinusoidal pattern is prominent in the hydroxyl-carboxyl (HXL-CXL) and hydroxyl-carboxamide (HXL-CXA) distributions, which resemble a $\cos(2\mathrm{BA}_\chi)$ function with peaks of equal height at $0°$ and $180°$ (dihedrals where the H-atom is in the sp^2 plane) and troughs of equal depth at $90°$ and $270°$ (dihedrals where the H-atom is perpendicular to the sp^2 plane), though the amplitude is greater for the carboxylate H-bonds. The distribution for carboxyl-guanidino (D/E to R) H-bonds, on the other hand, shows a large peak at $180°$, and almost no density at $0°$. The backbone/backbone distribution (which are dominated by β-sheet contacts after H-bonds with sequence separation less than five are filtered out), on the other hand, does not at all resemble a sinusoidal function.

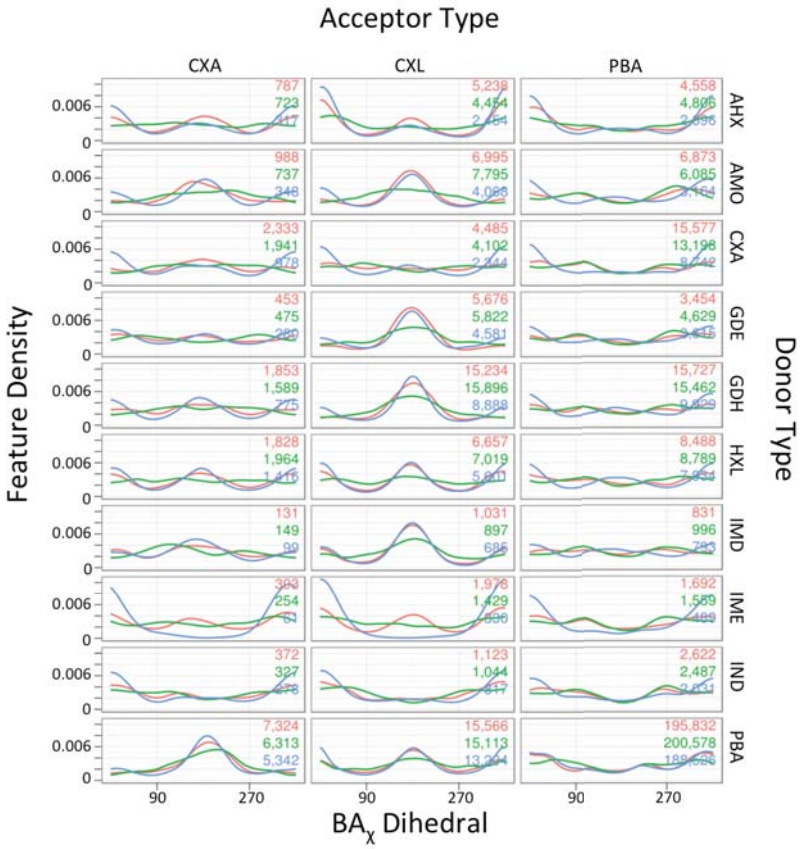


Figure S.3.2: *The Lambert-azimuthal projection.* This projection maps the surface of a sphere, parameterized by (ϕ, θ) for example (BAH, BA_χ)), onto the plane, parameterized by (x, y), via

$$x = 2\sin\left(\frac{\phi}{2}\right)\cos(\theta)$$
$$y = 2\sin\left(\frac{\phi}{2}\right)\sin(\theta)$$

For a hydrogen atom approaching a carboxyl oxygen (Left), consider the unit sphere centered on the oxygen with a positive z-axis pointed away from the base atom (C), and where the acceptor-base-base atom (R) lies in the x-z plane. The positive y-axis is going into the page. The red dot denotes the location on the surface of the sphere directly between the hydrogen and the oxygen. If this sphere were the surface of the earth (Right), then the Lambert-azimuthal projection places the North Pole in the center, and the green latitude lines appear as concentric circles. Longitude lines (not shown) would radiate outwards from the center as straight lines. If you consider the origin to lie at the North Pole ($\phi = 0°$), then the prime meridian (Longitude $= \theta = 0°$) lies along the positive x-axis. The positive x-axis corresponds to a BA_χ dihedral of 0°, the positive y-axis corresponds to a BA_χ dihedral of 90°, the negative x-axis corresponds to a BA_χ dihedral of $180°$, and the negative y-axis corresponds to a BA_χ dihedral of $270°$. The Lambert-azimuthal projection preserves the area of objects on the surface of the sphere, but it still distorts them; notice that Australia appears wider than North America. Notice also that Australia and the entire southern hemisphere are *visible* in this projection.

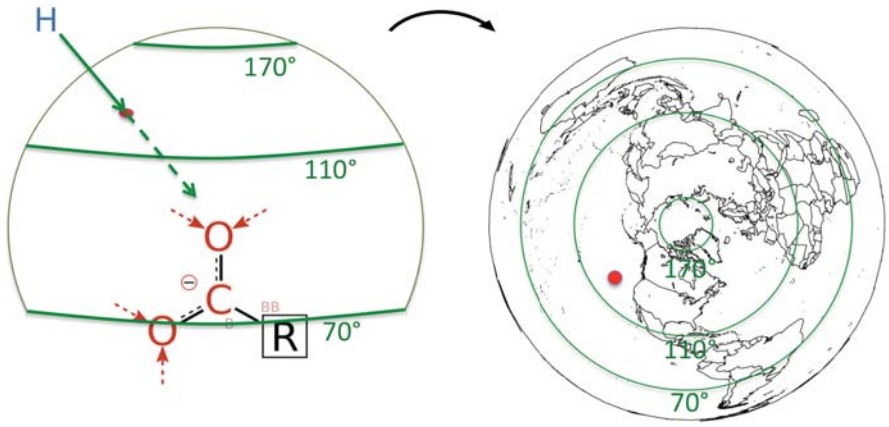


Figure S.3.3: *Lambert-azimuthal projections of Serine/Threonine hydrogen bonds to Asparatate/Glutamate.* A. Natives, B. *HBv1*, C. *HBv2*, D. *ElecHBv2*. The feature-density color mappings are individually normalized. The BA_χ dihedral angles are marked with diagonal lines; the BAH angles are shown as concentric circles radiating outwards from a BAH angle of $180°$ at the center of the plots. The *sp*^{2} lobes would be located at a BAH angle of $120°$, and at BA_χ dihedrals of $0°$ and $180°$. Indeed, in the natives, these BAH / BA_χ combinations represent the peaks of the distributions. The *HBv1* model correctly recapitulates a peak at $BAH = 120°$, but loses almost all of the BA_χ preference. The mild preference for BA_χ of $180°$ over $90°$ is likely an artifact of having started from native hydrogen bond contacts. *HBv2* and *ElecHBv2* both recover the native distribution quite well.

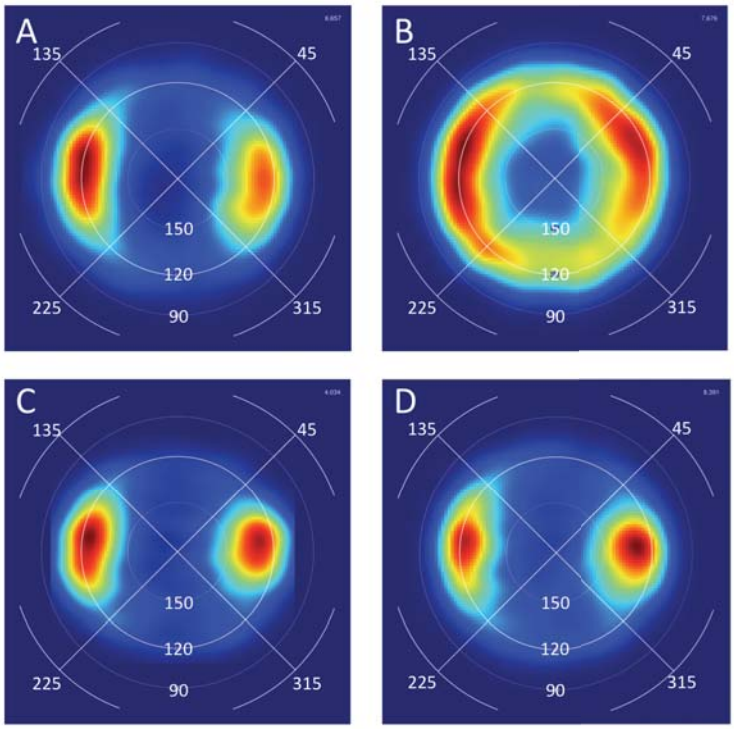


Figure S.3.4: *Lambert-azimuthal projections of lysine/backbone hydrogen bonds.* A. Natives, B. *HBv1*, C. *HBv2*, D. *ElecHBv2*. The feature-density color mappings are individually normalized. Unlike most hydrogen bonds involving sp^2-hybridized acceptors, lysine/backbone hydrogen bonds occur with significant frequency at BA_χ of $90°$. They are also much more tightly distributed towards a BAH angle of $180°$. The *HBv2* and *ElecHBv2* energy functions both produce distributions that strongly resemble the carboxyl-hydroxyl hydrogen bonds shown in Supp. Fig. S.3.3, with two well separated lobes, even if the BAH angles at the distribution peaks are $135°$ rather than $120°$.

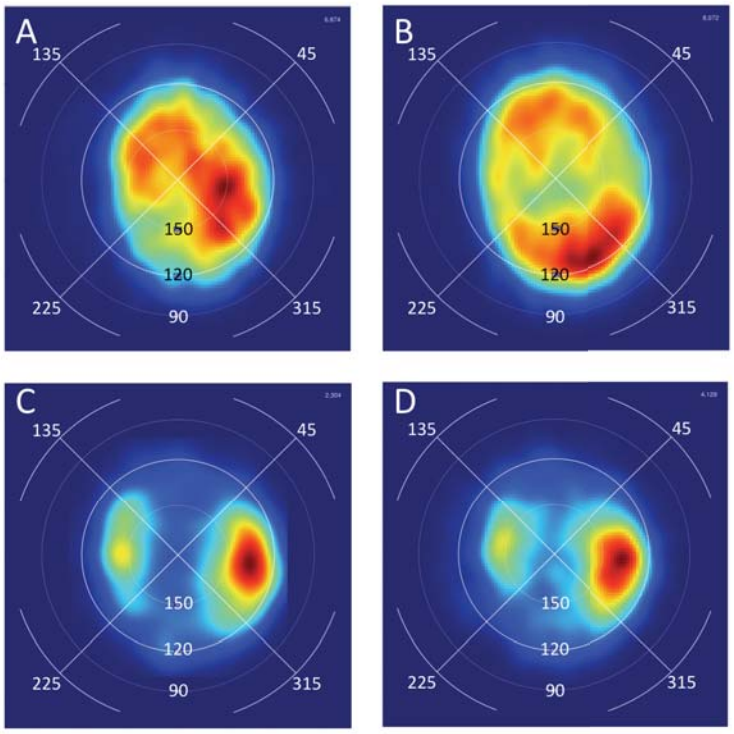


Figure S.3.5: *Lambert-azimuthal height maps of the* $h^{b}_{BAH,BA_{\chi}}$, $h^{c}_{BAH,BA_{\chi}}$ *and* $h^{d}_{BAH,BA_{\chi}}$ *potentials.* $h^{b}_{BAH,BA_{\chi}}$ is shown in A and D; $h^{c}_{BAH,BA_{\chi}}$ is shown in B. and E, $h^{d}_{BAH,BA_{\chi}}$ is shown in C and F. The functional forms for these three models are described in section S.4.2. The cross sections shown in D, E, and F cut through the potentials along the $BA_{\chi} = 90°, 270°$ axis (the y-axis). The peaks in these potentials, which are favorable in these formulations, are placed at $BAH = 120°$ and $BA_{\chi} = 0°$ and $180°$. The potentials differ in their treatment of the $BA_{\chi} = 90°, 270°$ axis: $h^{b}_{BAH,BA_{\chi}}$ is perfectly flat across this axis, $h^{c}_{BAH,BA_{\chi}}$ places a depression at $BAH = 180°$ relative to $BAH = 120°$ and $BA_{\chi} = 90°$ and $270°$, and E^{d}_{HBv2} places saddle point at $BAH = 180°$ with the potential decreasing as it approaches $BAH = 120°$ and $BA_{\chi} = 90°$ and $270°$.

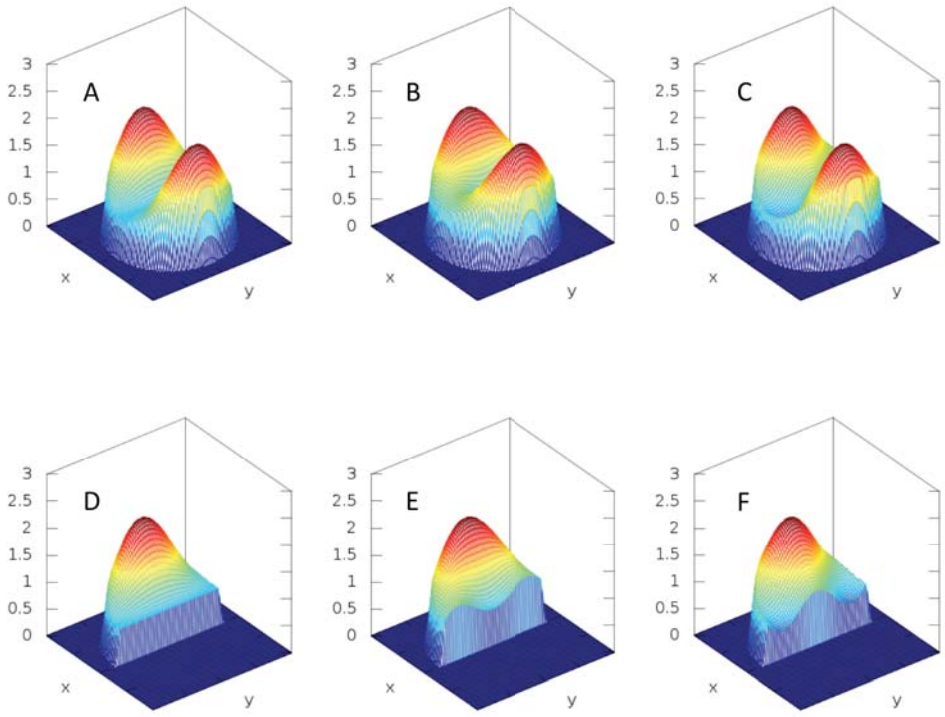


Figure S.3.6: *Lambert-azimuthal projections of long-range backbone-backbone hydrogen bonds.* A. Natives, B. E^{b}_{HBv2}, C. E^{c}_{HBv2}, and D. E^{d}_{HBv2}. The shape of these three potentials is plotted in Fig. S.3.5 and the functional form given in section S.4.2. The appearance of density in C and D in regions where its absent in the native distributions strongly suggested that the correct $BAH/\,BA_{\chi}$ potential would need to be flat through the $BA_{\chi} = 90°, 270°$ axis.

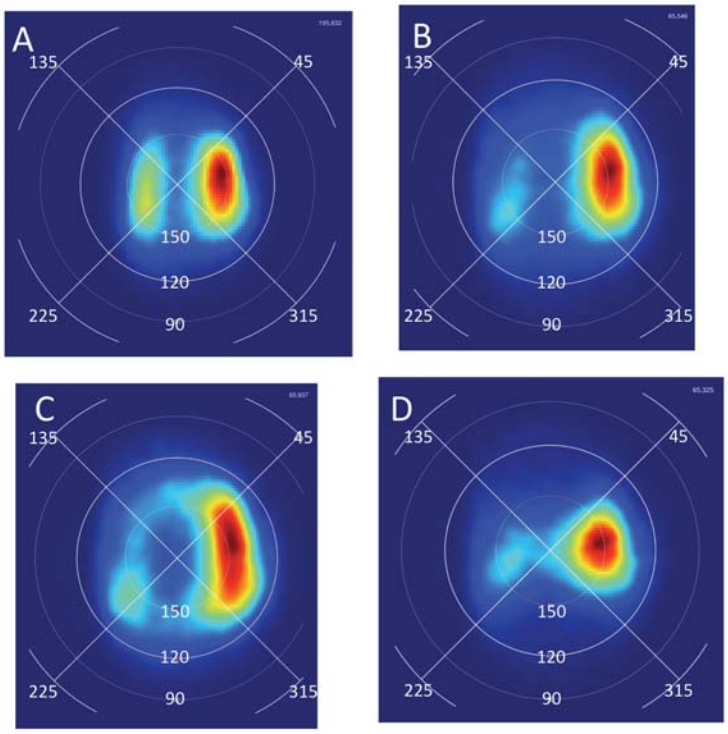


Figure S.3.7: *The joint AHD/BAH distribution of bidentate ASP/GLU vs ARG hydrogen bonds*. The native join density distribution (A) for single hydrogen bonds (red) and bidentate hydrogen bonds (cyan) is shown as a contour map. There is a single peak for bidentate hydrogen bonds at $\cos(AHD) \sim 0.83$, $\cos(BAH) = 0.77$, but the distribution is quite broad. The peak for single hydrogen bonds at $\cos(AHD) = 1$, $\cos(BAH) = 0.5$ represents $AHD = 180°$ and $BAH = 120°$. In contrast, E^{b}_{HBv2} (B) produced two peaks for bidentate hydrogen bonds; a small one very close to the peak observed for natives, and another much sharper one at $\cos(BAH) = 0.95$ and $\cos(BAH) = 0.35$. This suggested to us that the potential was emphasizing one of the two hydrogen bonds while sacrificing the other. The multiplicative structure of the E^{b}_{HBv2} potential certainly makes it possible for one high-quality hydrogen bond + one low-quality hydrogen bond to sum to a lower total energy than two medium-quality hydrogen bonds. A purely additive potential would not produce this kind of behavior, so following the observation of this distribution, we pursued a purely additive potential. *ElecHBv2* (C), was tested after updating to the new ideal bond geometries suggested by Song et al.; the bidentate distributions it produces much more closely matched that of the native's. Indeed, updating the ideal bond geometries corrected the bidentate distributions for the E^{b}_{HBv2} potential (D). We therefore have little evidence suggesting that the multiplicative functional form of E^{b}_{HBv2} is worse than the additive functional form of E_{HBv2} .

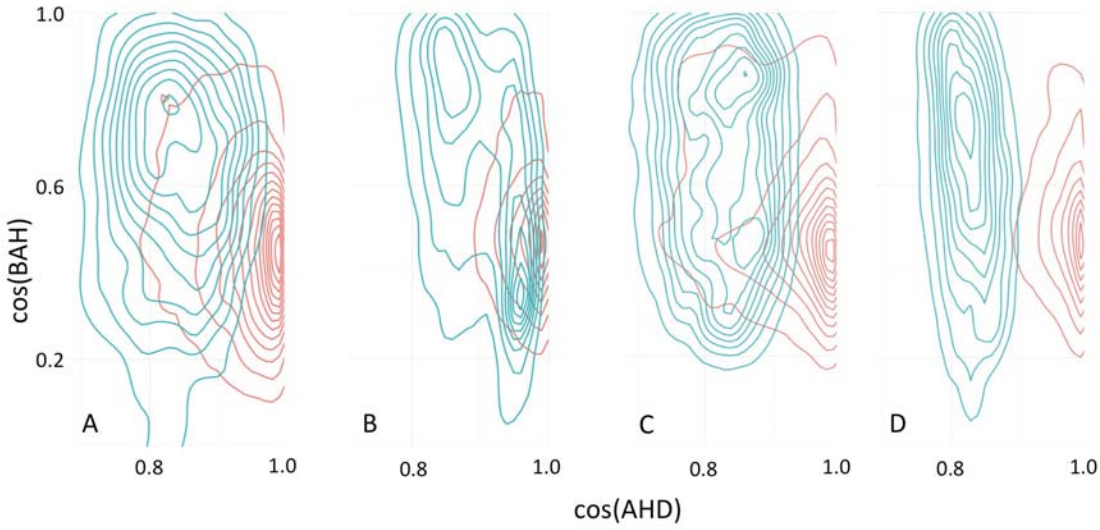


Figure S.3.8: *The BAH distribution of backbone/backbone hydrogen bonds by sequence separation*. Red: The native distribution; Green: *HBv1*; Blue: *HBv2*. The sequence separation is given in the upper left corner denoting the difference in the sequence position of the donor and the acceptor (e.g. helical hbonds have a sequence separation of $+4$). The same potential is used to describe the optimal BAH angle ($120°$) in *HBv2*, whereas three separate sets of polynomials are used to describe the BAH dependence in *HBv1*.

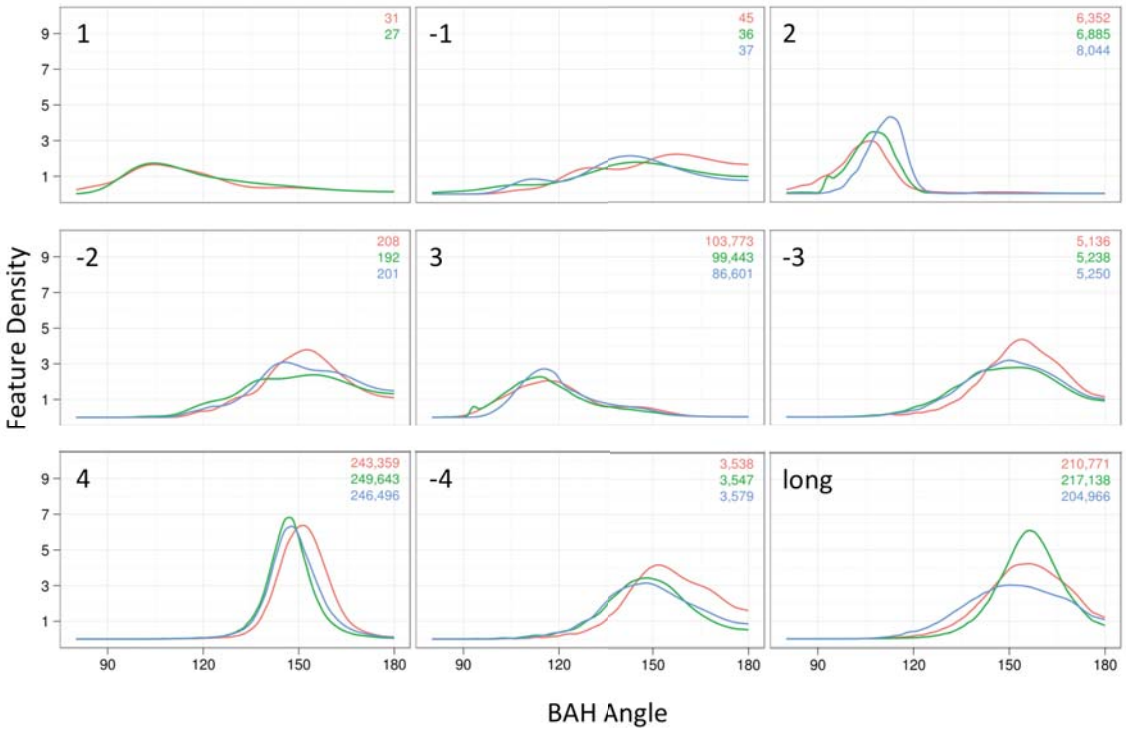


Figure S.3.9: *Score comparison for backbone/backbone hydrogen bonds between HBv1 and HBv2.* These scatter plots were generated by scoring crystal structures with both *HBv1* and *HBv2*. Each point represents a single hydrogen bond where its x-coordinate represents its weighted score from *HBv1*, and its y coordinate represents its weighted score from *HBv2*. The sequence separation is given in the upper left corner denoting the difference in the sequence position of the donor and the acceptor (e.g. helical hbonds have a sequence separation of $+4$). The line, $x = y$ is shown. Short-range hydrogen bonds are down-weighted by $\frac{1}{2}$ relative to long-range hydrogen in *HBv1*, but they are not similarly down-weighted in *HBv2*. Strikingly, long-range hydrogen bonds are quite a bit weaker in *HBv2*.

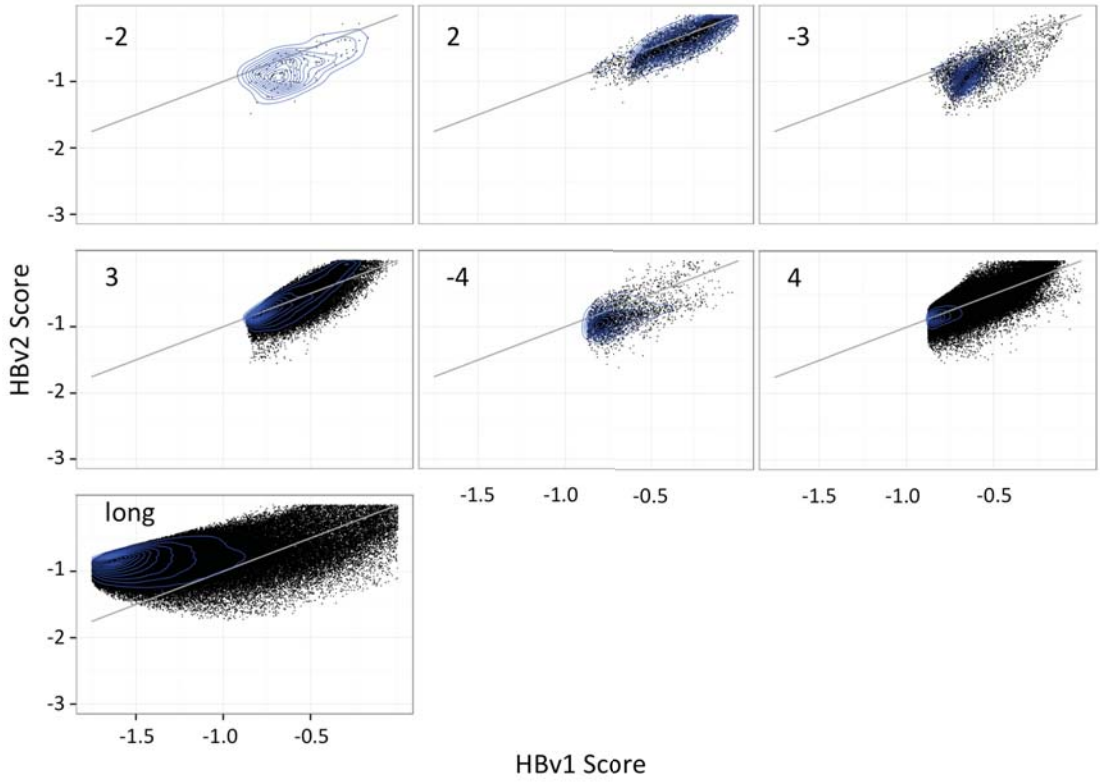


Figure S.3.10: BAH distributions for sp^3 hybridized acceptors. Red: Natives. Green: *HBv1*. Blue: *HBv2*. Aromatic hydroxyl (AHX, i.e. tyrosine) and Hydroxyl (HXL, i.e. serine and threonine) hydrogen bond acceptors both prefer to accept with BAH angles larger than the $109.5°$ angle predicted based on the location of the sp^3 lobes.

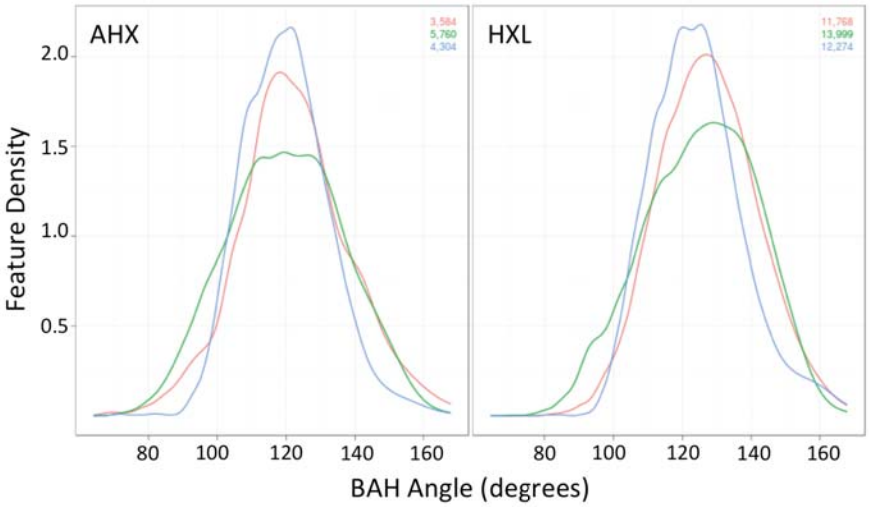


Figure S.3.11: HAH distributions for sp^3 hybridized acceptors. Red: Natives. Green: *HBv1*. Blue: *HBv2*. The HAH angle is measured from the sp^3 acceptor hydrogen, the acceptor heavyatom, and the donor hydrogen. Since the acceptor hydrogen's position has to be inferred, we consider here only the acceptors that are both accepting and donating hydrogen bonds to increase our confidence in the acceptor hydrogen's position. *HBv1* actually uses the HAH angle to evaluate the h polynomial—it treats the acceptor hydrogen as the "B" atom in defining the BAH angle. It is a little surprising, then, that the *HBv1* distribution for the HAH angle fits the native distribution so poorly. *HBv2,* on the other hand, does not explicitly model the HAH angle. Its HAH distribution, however, is influenced by both the explicitly modeled BAH angle and a sinusoidal penalty on the dihedral angle defined by the acceptor hydrogen, the acceptor-base, the acceptor and the donor hydrogen.

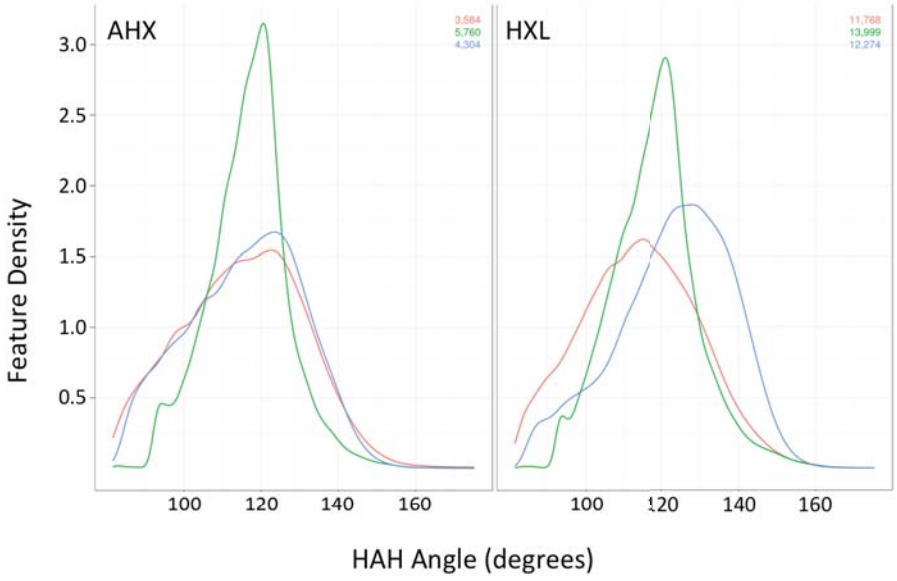


Figure S.3.12: *Hydroxyl chi angles by acceptor type*. Red: Natives. Green: Abinitio models generated w/ *HBv1*. Blue: Abinitio models generated w/ *HBv2*. Acceptor-type names are given in Figure 1. The χ_2 distribution in crystal structures does not show a preference for the staggered dihedral angles of $60°$, $-60°$, and $180°$ as might be expected. Instead, it shows peaks near $90°$ and $-90°$ and a broad plateau of density between in the range between them. A clear trough can be seen at $0°$. *HBv1* sampled χ_2 at the staggered dihedral angles only. Though it seems that in many cases the $-60°$ and $+60°$ samples minimized to the native peaks, *HBv1* did produce a non-native peak at $180°$. *HBv2* samples at $20°$ increments starting at $0°$, and produces a distribution that is a bit more native like.

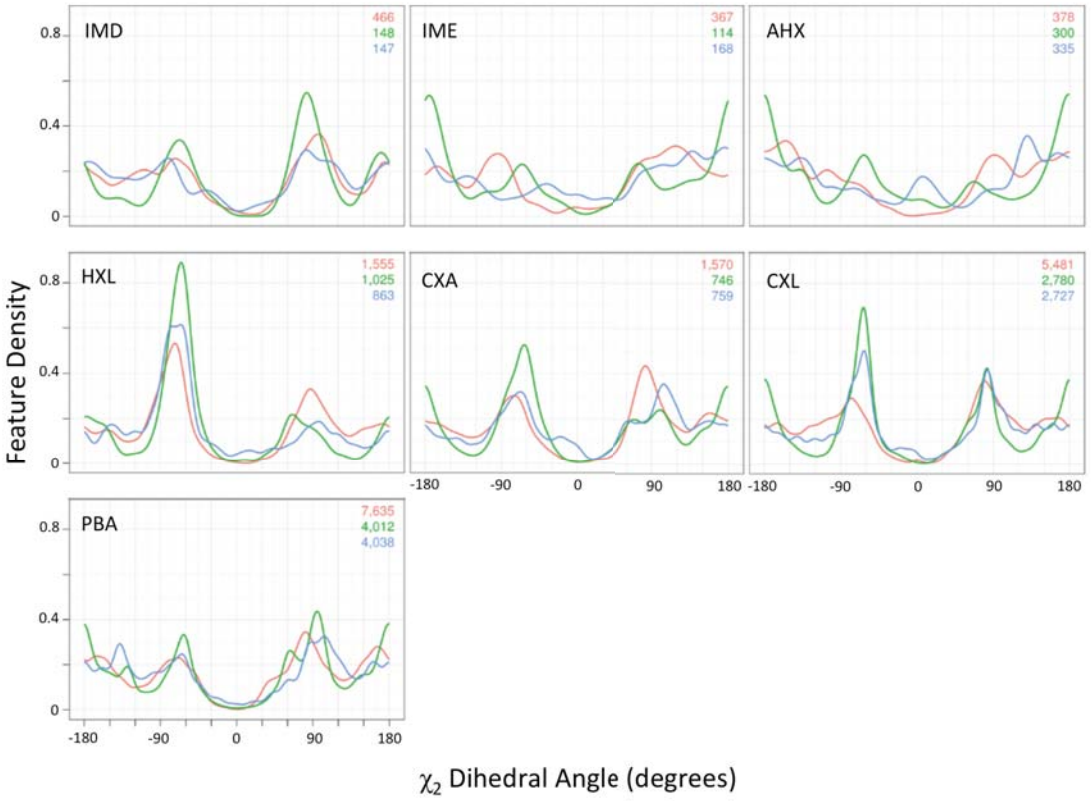


Figure S.3.13: *Acceptor-hydrogen distance for designed serines and threonines in α-helices*. When the backbone/sidechain exclusion rule is removed, Rosetta is able to design in very good hydrogen bonds between serines at residue *i*, and backbone oxygens at residues $i-3$ and $i-4$. The peak in the hydrogen-acceptor distance distribution in native structures for hydroxyl/backbone hydrogen bonds is at 1.7 Å. This small dataset was created by redesigning 38 large proteins. Counts of each kind of hydrogen bond formed are given in the upper-right corner of each plot.

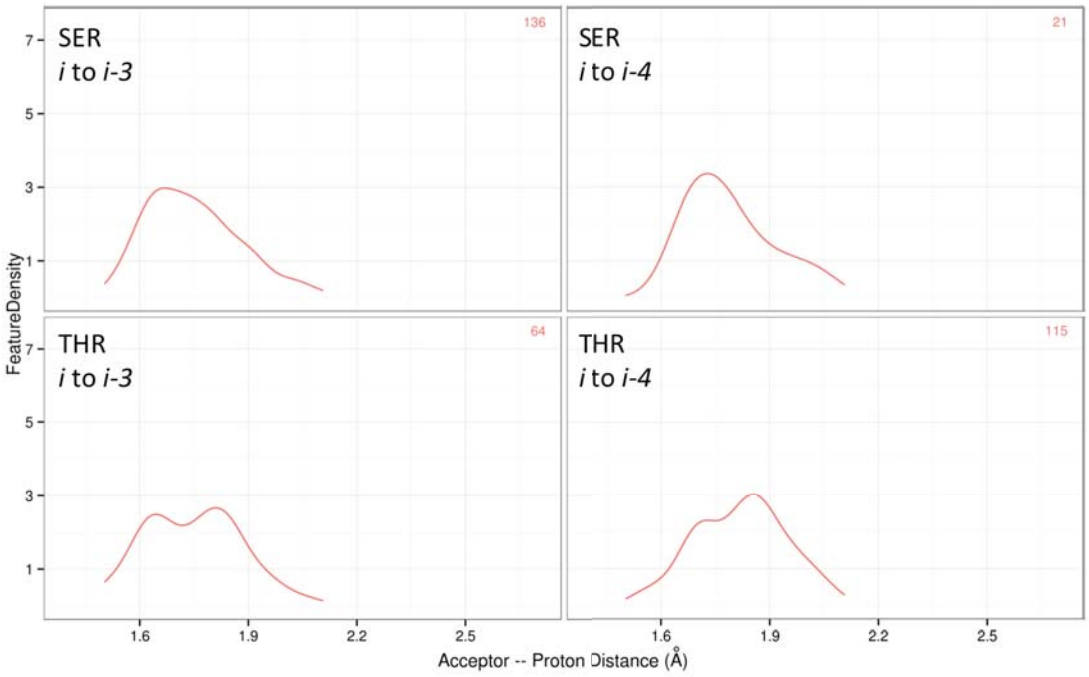


Figure S.3.14: BAH / BA_χ *Lamber-azimuthal projection for designed serines and threonines in* α*-helices*. When the backbone/sidechain exclusion rule is removed, Rosetta is able to design in very good hydrogen bonds between serines at residue i, and backbone oxygens at residues $i-3$ and $i-4$. Contacts to residue $i-3$ have a more favorable BA_χ dihedral than those formed to residue $i-4$; however, the second class of contacts still produces quite favorable energies. This small dataset was created by redesigning 38 large proteins. Counts of each kind of hydrogen bond formed are given in the upper-right corner of each plot

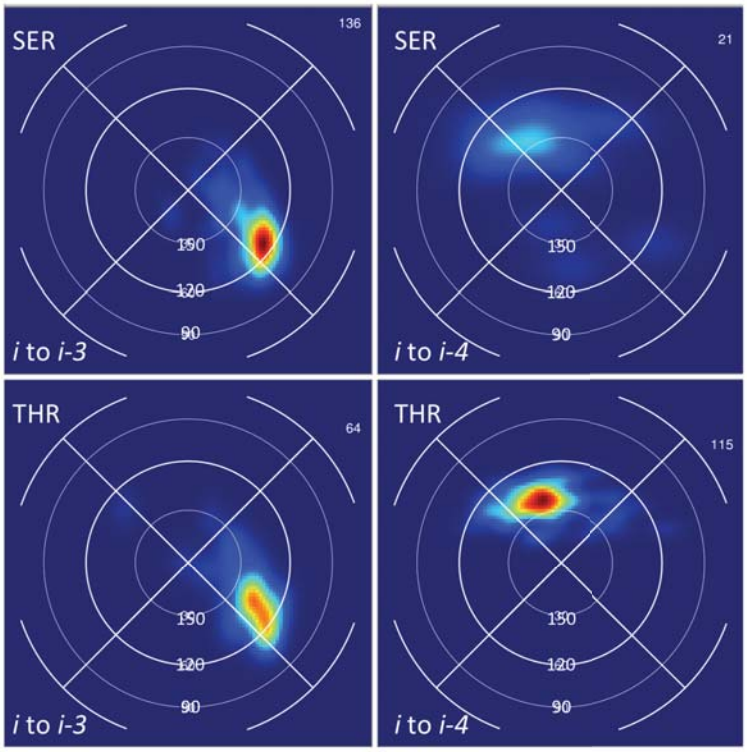


Figure S.3.15: AHD *angle for designed serines and threonines in* α*-helices*. When the backbone/sidechain exclusion rule is removed, Rosetta is able to design in very good hydrogen bonds between serines at residue *i*, and backbone oxygens at residues $i-3$ and $i-4$. This small dataset was created by redesigning 38 large proteins. Counts of each kind of hydrogen bond formed are given in the upper-right corner of each plot.

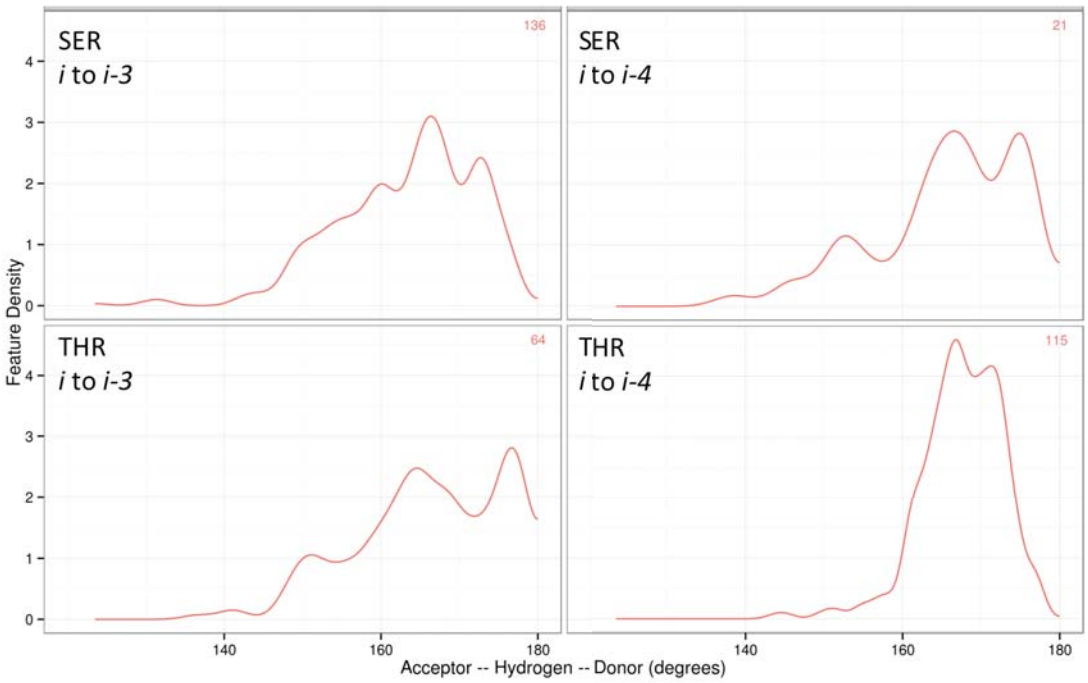


Figure S.3.16: *ElecHBv2 energies for designed serines and threonines in α-helices*. When the backbone/sidechain exclusion rule is removed, Rosetta is able to design in very good hydrogen bonds between serines at residue *i*, and backbone oxygens at residues $i-3$ and $i-4$. This small dataset was created by redesigning 38 large proteins. Counts of each kind of hydrogen bond formed are given in the upper-right corner of each plot. The maximum unweighted energy for any hydrogen bond in our models is -1.5; these hydrogen bonds are not optimal, but they are very favorable. If someone were interested in removing the backbone/sidechain exclusion rule, then they will have to devise a way overcome some very favorable contacts or otherwise face a large number of designed Ser/Thr in helices forming hydrogen bonds to carbonyl oxygens along the helix.

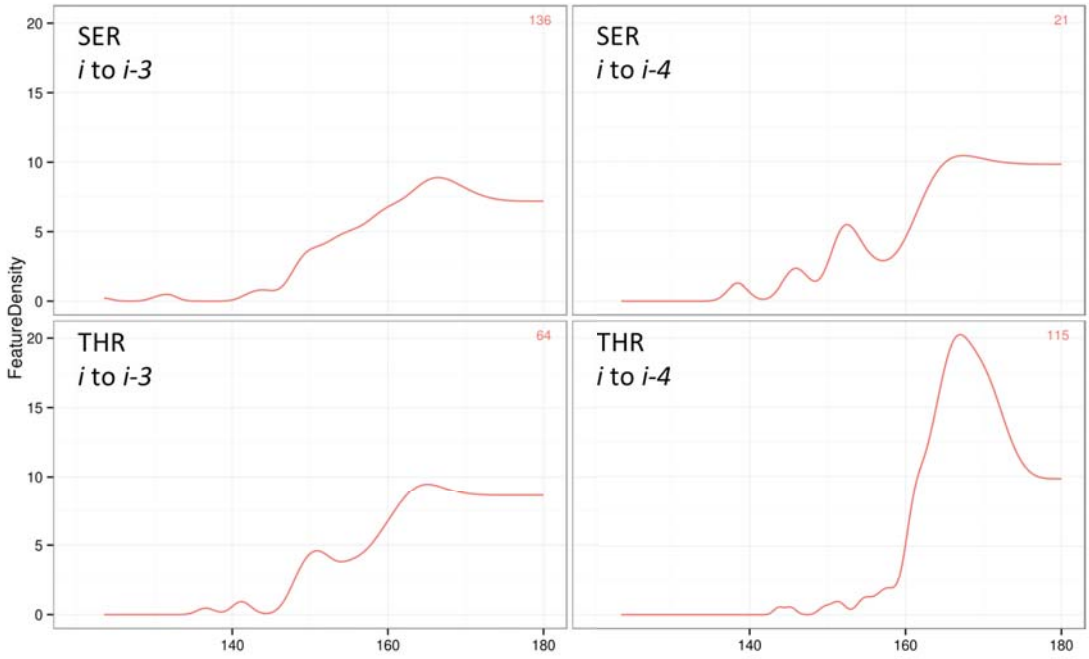


Figure S.3.17: *Problematic fitting of new polynomials for the AHD angle.* The simultaneous constraints that the derivative of the $\cos(AHD)$ polynomial be zero at $AHD = 180°$ and that the polynomial be quite steep in the region near $AHD = 180°$ cannot be reconciled with polynomials of degree 11 or less—we did not explore higher degree polynomials. Attempts to fit polynomials when the derivative was constrained to be zero at $AHD = 180°$ resulted in polynomials that were insufficiently steep. These polynomials produced AHD distributions that placed too little density at $AHD = 180°$ and too much density at $AHD = 160°$. Red: Natives. Green: *HBv1*. Blue: A prototype version on *HBv2*, which relied still on $\cos(AHD)$ polynomials. Shown here are the AHD distributions from backbone/backbone hydrogen bonds broken down by sequence separation.

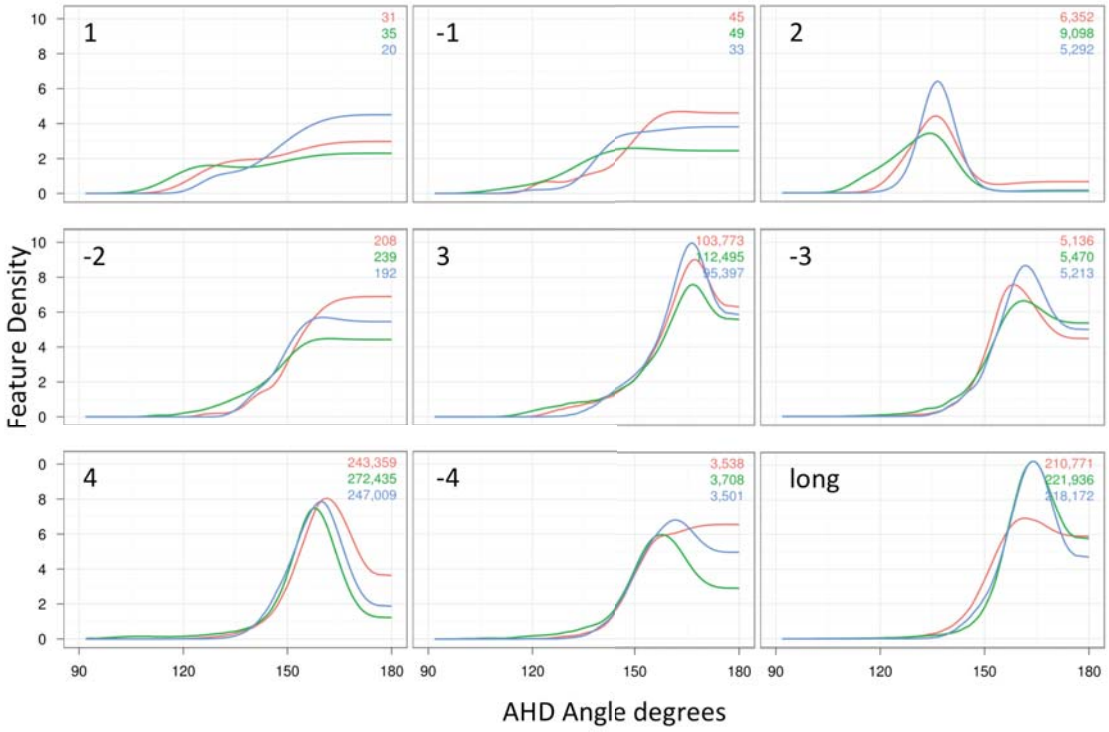


Figure S.3.18: *Fitting AHD polynomials instead of $\cos(AHD)$ polynomials alleviates the problem.* A polynomial could be fit to match the native distribution for backbone/backbone hydrogen bonds after switching from polynomials of $\cos(AHD)$ to polynomials of AHD. Red: Natives. Green: *HBv1*. Blue: *HBv2*. Shown here are the AHD distributions from backbone/backbone hydrogen bonds broken down by sequence separation.

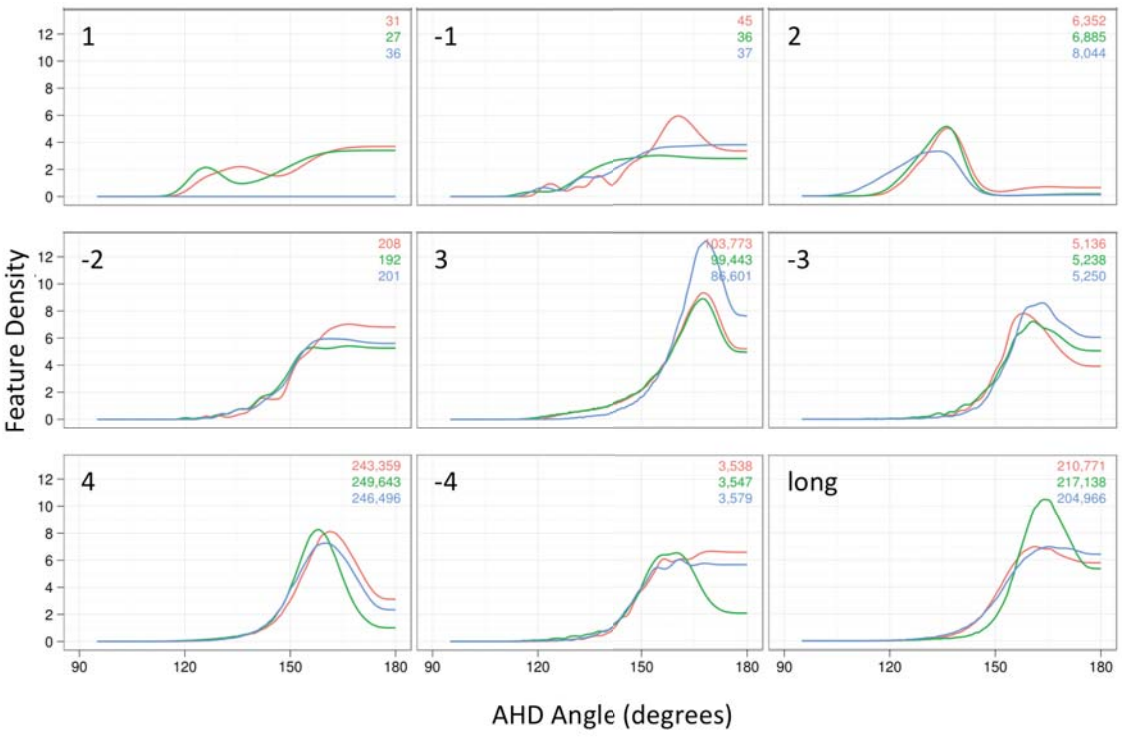


Figure S.3.19: AHD *distributions produced by HBv2 broken down by donor and acceptor type*. Red: *Natives*. Green: *HBv1*. Blue: *HBv2.*

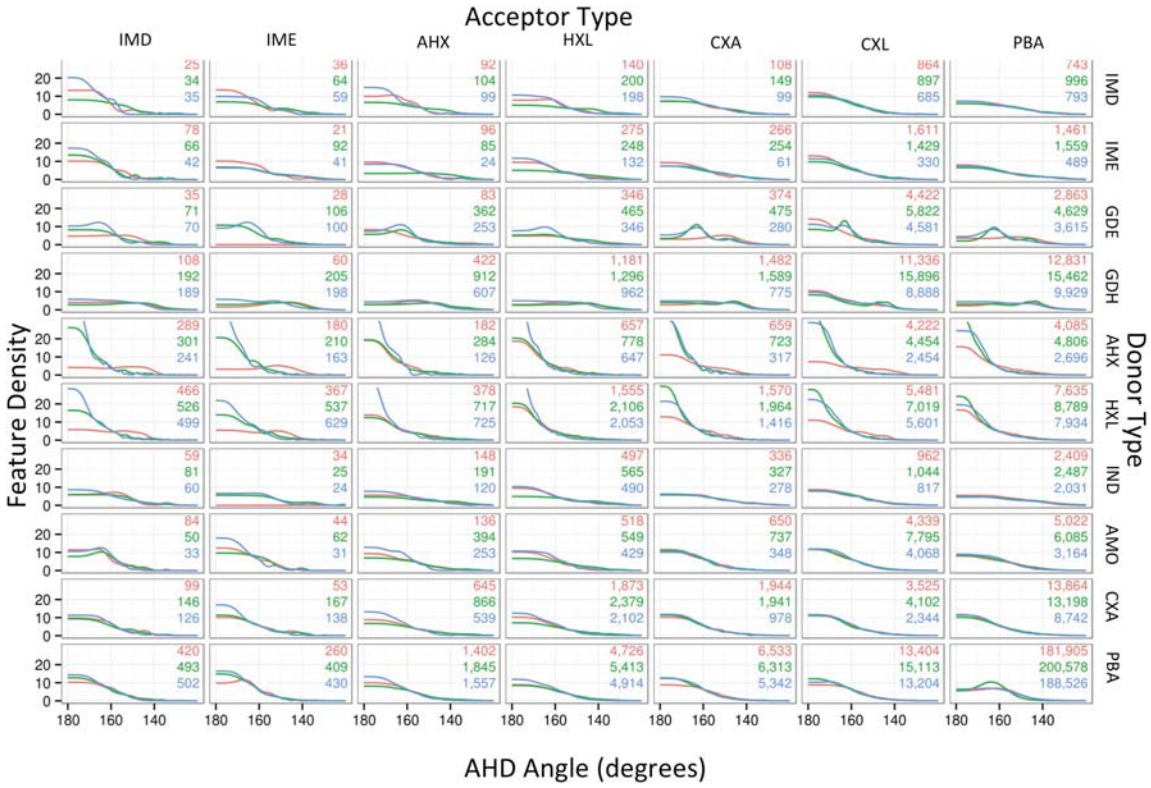


Figure S.3.20. *Derivative discontinuities in HBv1 produce significant artifacts.* This histogram, with bin width of $1/600$ Å, shows two peaks in the *HBv1* distance distribution for backbone-backbone hydrogen bonds. The labeled points on the x-axis, $(1.9, 2.1, 2.3)$ are points of derivative discontinuities in the Score12 H-Bond function due to knots in the piecewise-linear functional form of the fade functions. The Native curve is partially obscured by *ElecHBv2*. These artifacts are easily removed by eliminating the derivative discontinuities.

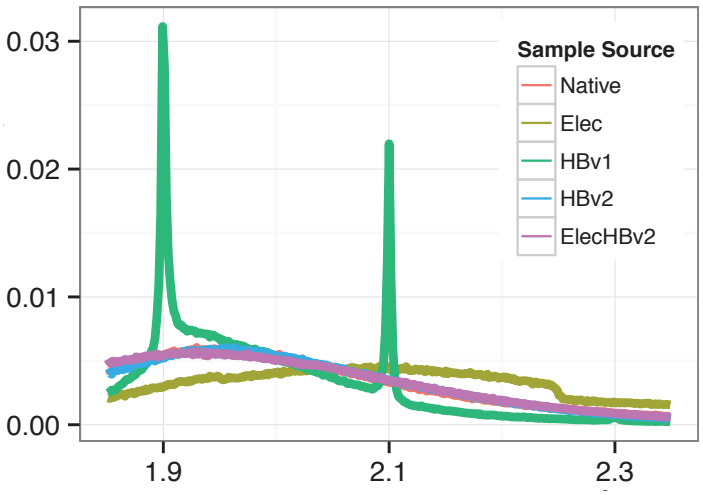


Figure S.3.21. *Apparent distance dependence of the AHD angle distribution.* This figure shows the cumulative distribution function (CDF) for the AHD angle, starting at $180°$ and decreasing towards $0°$, broken down by distance windows. The window sizes were chosen so that each window contained roughly the same number of contacts; the windows overlap so that each curve could include a greater number of counts. To generate this figure, the AHD angle for every contact with an AH_{dis} beneath 4 Å was considered instead of considering only those hydrogen bonds where their Rosetta energy was less than 0. Most contacts above 2.5 Å are not hydrogen bonds. In the natives, a trend toward decreasing stringency in the AHD angle is apparent as the AH_{dis} increases. Even if the contacts in the last two windows should not be considered hydrogen bonds, there is nonetheless a difference between the distributions for the first two. This latter difference poses this question: is there an energetic difference in the AHD angle that depends on the AH_{dis} ? *HBv1*, which incorporates a distance dependence into its angular energy component (g_s vs. g_l in Equation 7 of the main text), recapitulates the CDF differences decently well. *ElecHBv2*, which does not include a distance dependence in its angular energy component, does a better job matching the native CDFs. Our interpretation here is that the AHD angle prefers to be linear for very close contacts to minimize the electrostatic repulsion between the acceptor and donor heavy atoms; at slightly longer distances, this repulsion is less and less linear AHD angles are acceptable.

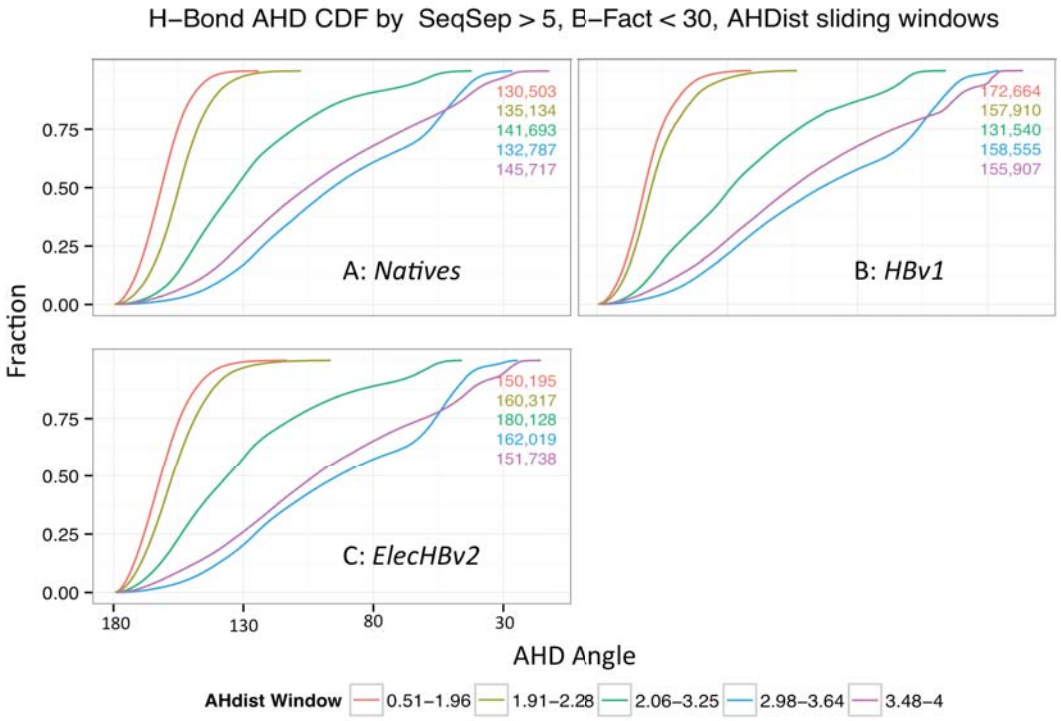


Figure S.3.22. *Distance dependence of the angular terms in the HBv1 model.* Given a particular AH_{dis}, BAH angle, and AHD angle for an amino-carboxylate (K-D/E) hydrogen bond, the *HBv1* energy can be calculated using the fifteen cells below. Each cell produces a function value. Multiply the function values across the rows then sum these products to produce the hydrogen bond energy. For this donor/acceptor pair, the long-range polynomials have a shallower well depth than the short-range polynomials for both the AHD and BAH angles. Consider a hydrogen bond with $AHD = 180°$ and $AH_{dis} = 2.3$ Å. Now, if the AH_{dis} decreases towards 1.9 Å while the AHD angle remains fixed, then the contribution from g_s increases and the contribution from g_l decreases. The same observation applies for the contribution of h_s and h_l for a BAH angle of $120°$. The distance term and the angular terms all encode a preference for short AH_{dis} values. This produces overly sharp AH_{dis} distributions in *HBv1*.

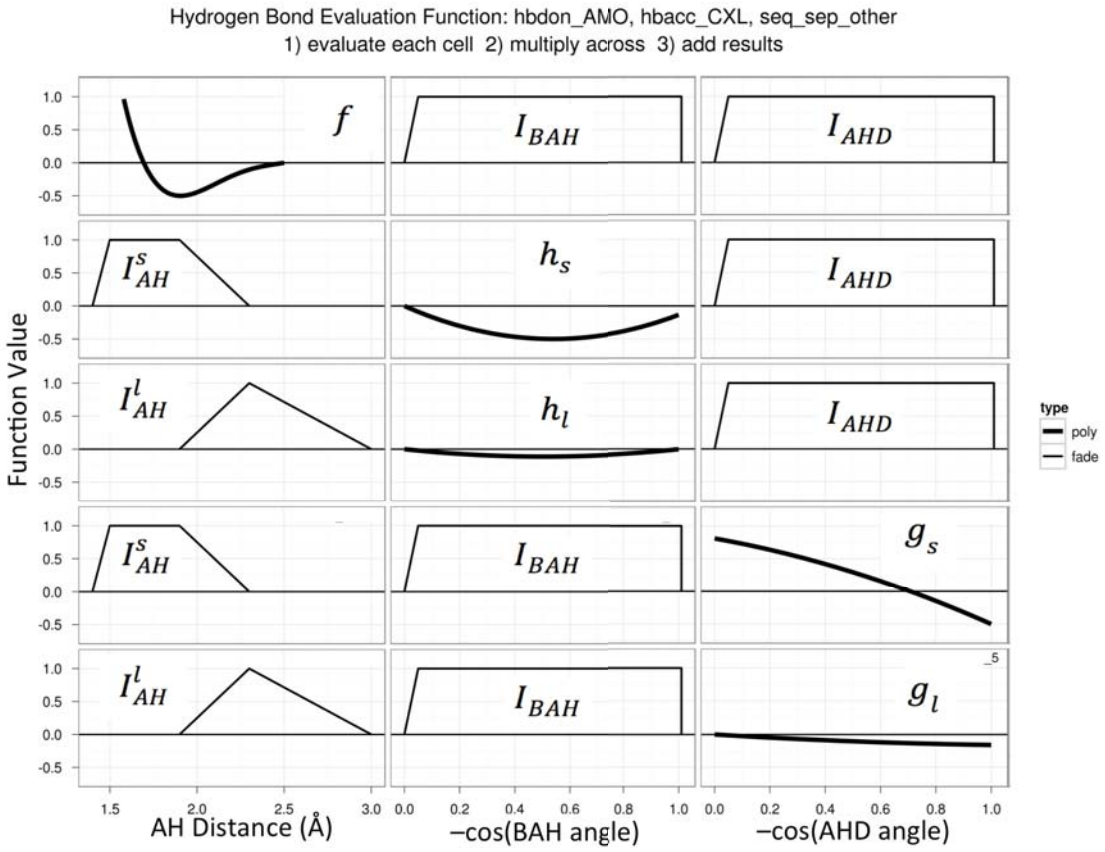


Figure S.3.23: *Backbone/backbone acceptor-hydrogen distance (AH_{dis}) distributions by sequence separation*. The sharp peaks in the *HBv1* distance distributions are the result of derivative discontinuities as described in Figure S.3.20. In *HBv2*, a single AH_{dis} polynomial was fit to describe all backbone/backbone H-bonds, whereas in HBv1 there were separate polynomials for SeqSeq $= 4$, SeqSep < 4 and SeqSep > 4. Red: *Natives*. Green: *HBv1*. Blue: *HBv2*.

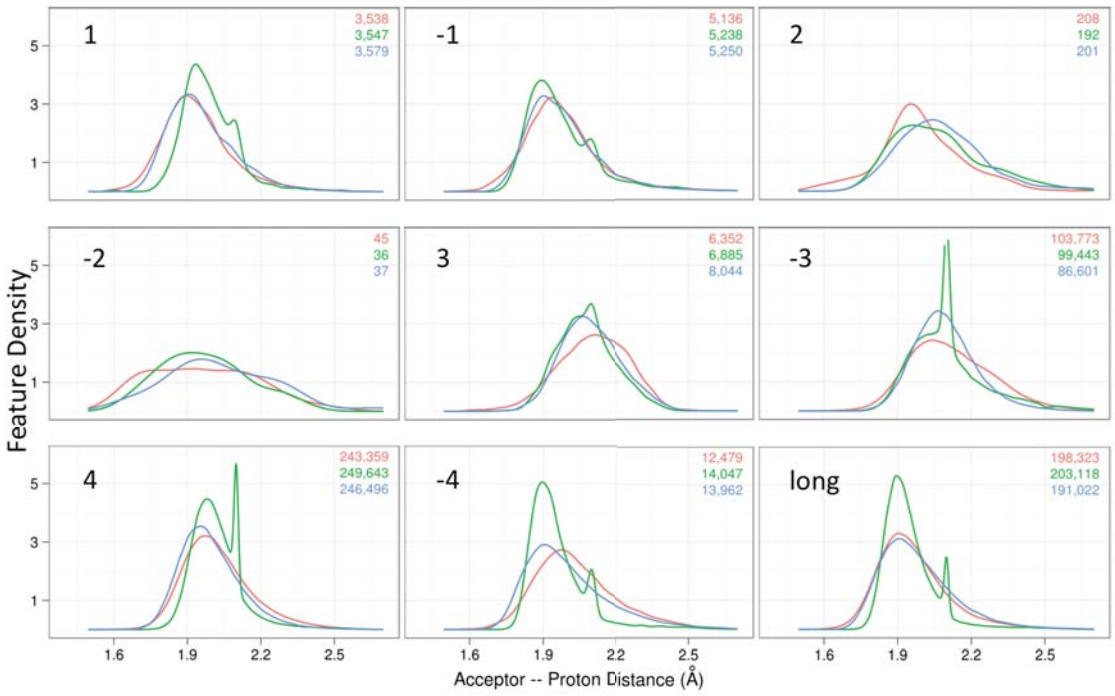


Figure S.3.24: *Acceptor-hydrogen distance (AH_{dis}) distributions broken down by donor and acceptor type.* REDUCE (Word, Lovell, Richardson, & Richardson, 1999) was used to infer hydrogen locations for the natives. Red: *Natives*. Green: *HBv1*. Blue: *HBv2.*

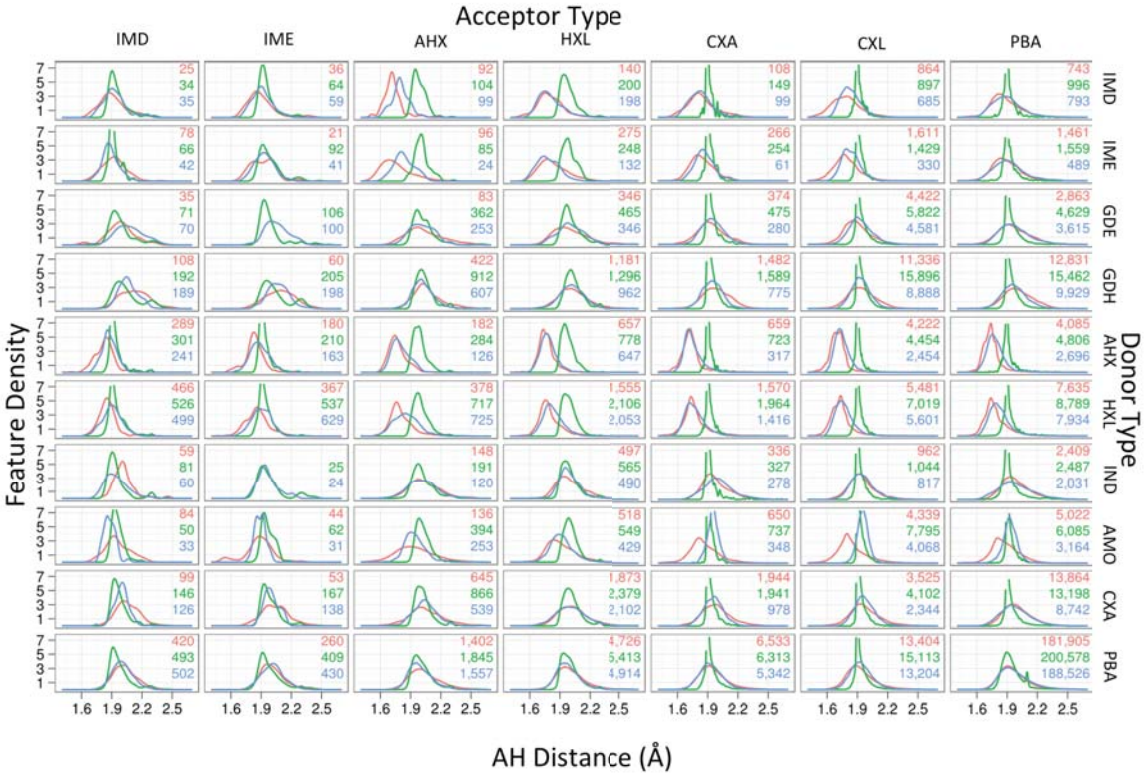


Figure S.3.25: *Acceptor to donor-heavyatom distance (AD_{dis}) distributions broken down by donor and acceptor type.* Red: *Natives*. Green: *HBv1*. Blue: *HBv2.*

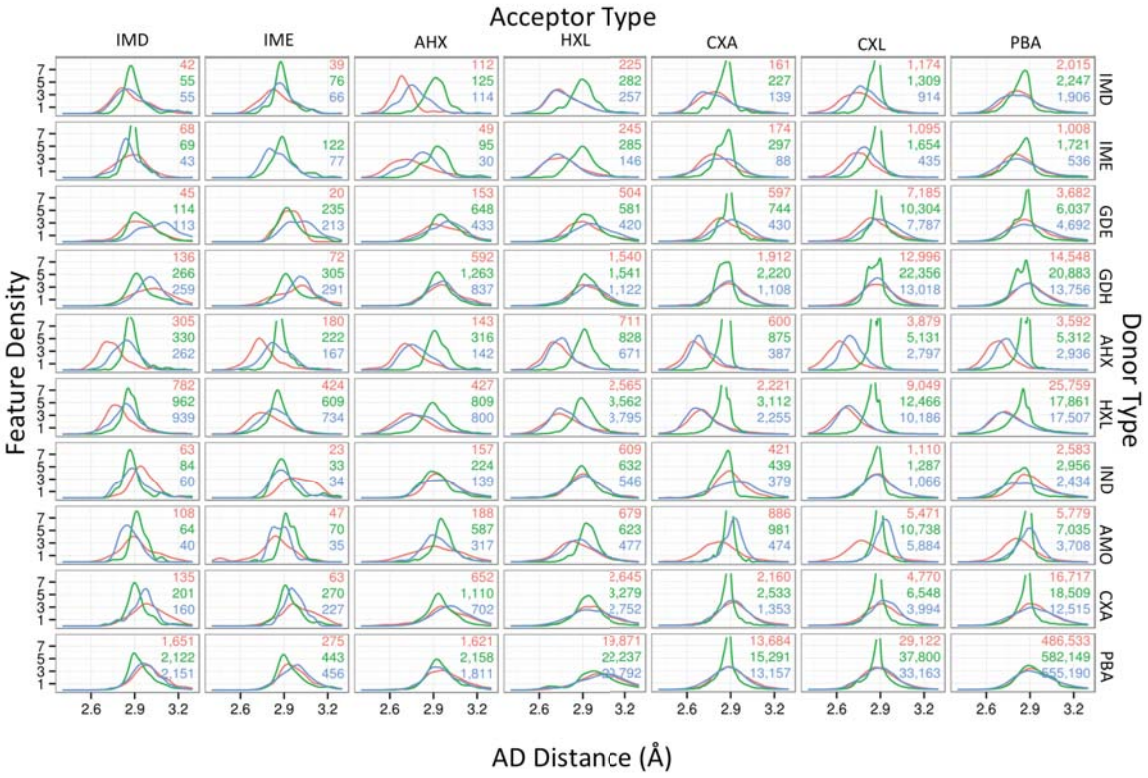
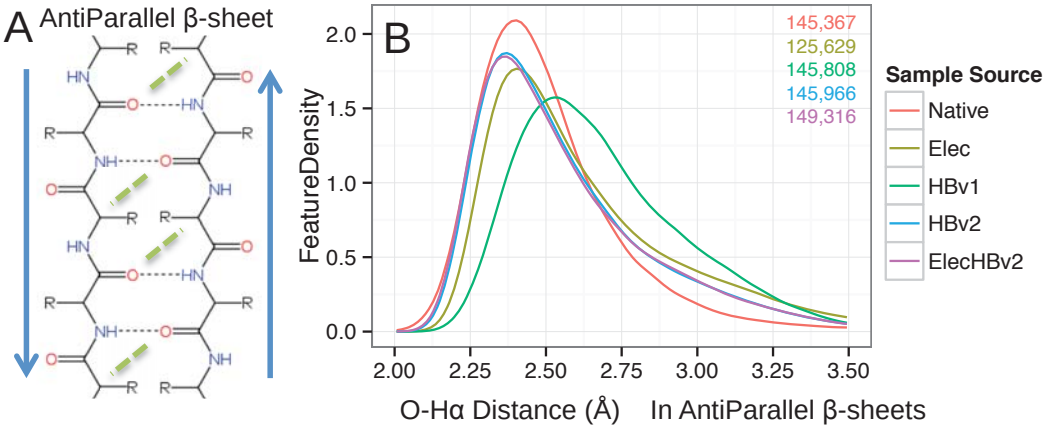


Figure S.3.26: *Recapitulation of putative carbon H-bonds in anti-parallel β-sheets* (A) In anti-parallel *β*-sheets, the close contact (H_α, O) distance feature (green dashed lines) is shown. The tight distribution for this distance has been attributed to an attractive "carbon H-bond" between these atoms. (B) Kernel density estimation of the (H_α, O) feature distribution by sample source. The *Elec, HBv2* and *ElecHBv2* sample sources recapitulate the Native distribution better than the *HBv1* sample source, indicating that modeling the sp^2 character of carbonyl acceptors and/or the electrostatic attraction between H_α is sufficient to recapitulate this feature distribution.



Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–42.

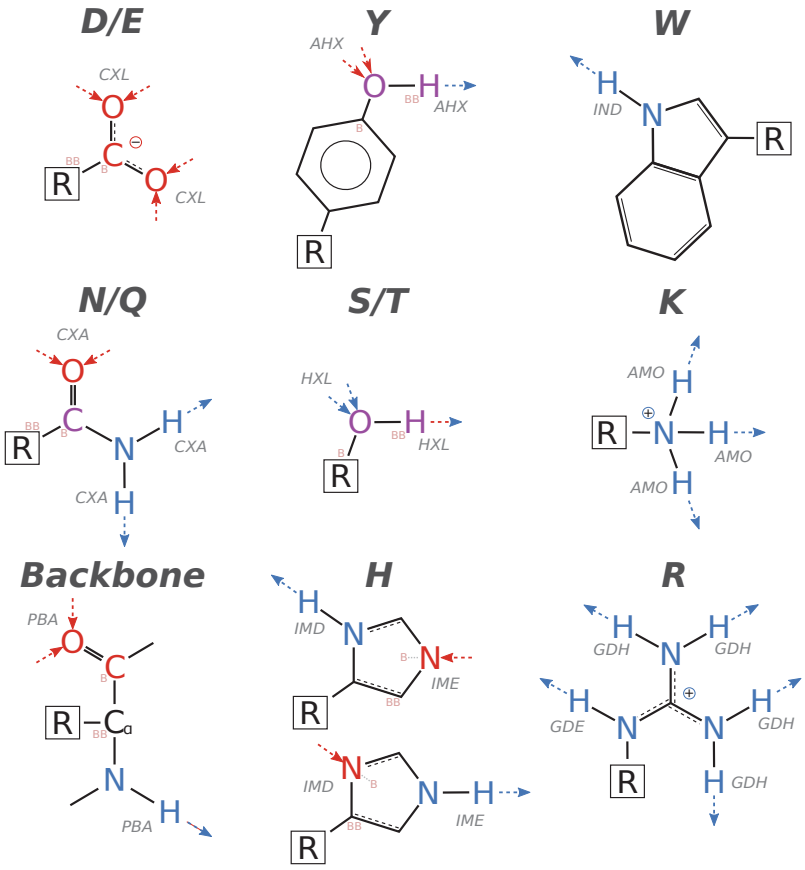
Bradley, P., Misura, K. M. S., & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science*, *309*(5742), 1868–71. doi:10.1126/science.1113801

S.4 Model Details

S.4.1 HBv2

S.4.1.1 HBv2 Chemical Types

HBv2 H-Bond Model chemical types in canonical proteins. Blue groups are donors with arrows point towards the acceptor and red groups are acceptors with arrows pointing from donors at canonical lone pairs. The base atoms (B, and BB) are labeled. Note, for Histidine, the B atom is the midpoint between the two carbons binding the nitrogen.



HBv2 H-Bond Model Acceptor Chemical Type Codes

	Code	Name	Example Rosetta Atom
1	hbacc_NONE		
2	hbacc_PBA	Protein Backbone Amide	Protein Backbone OCbb
3	hbacc_CXA	Carboxamide	ASN OD1; Gln OE1
4	hbacc_CXL	Carboxyl	ASP OD{1 2}; Glu OE{1 2}
5	hbacc_IMD	Imidazole Delta	HIS ND1 (NE2 protonated)
6	hbacc_IME	Imidazole Epsilon	HIS NE2 (ND1 protonated)
7	hbacc_AHX	Aromatic Hydroxyl	TYR OH
8	hbacc_HXL	Hydroxyl	SER OG; THR OG1

HBv2 H-Bond Model HBond Donor Chemical Type Codes

	Code	Name	Example Rosetta Atom
1	hbdon_NONE		
2	hbdon_PBA	Protein Backbone Amide	Protein Backbone Nbb--HNbb
3	hbdon_CXA	Carboxamde	ASN ND2--{1 2}HD2; Gln NE2--{1 2}HE2
4	hbdon_IMD	Imidazole Delta	HIS ND1--HE2 (NE2 not protonated)
5	hbdon_IME	Imidazole Epsilon	HIS NE2--HE2 (ND1 not protonated)
6	hbdon_IND	Indole	TRP NE1--HE1
7	hbdon_AMO	Amino	LYS NZ--{1 2 3}HZ
8	hbdon_GDE	Guanidino Epsilon	ARG NE--HE
9	hbdon_GDH	Di-Hydro Guanidino	ARG NH{1 2}--{1 2}HH{1 2}
10	hbdon_AHX	Aromatic Hydroxyl	TYR OH--HH
11	hbdon_HXL	Hydroxyl	SER OG--HG; Thr OG1--HG1
12	hbdon_H2O	Water	H2O O--H{1 2} ((TP5/TP3 models)

S.4.1.2 Other Models Considered

Before arriving at the final $h^2_{BA_\chi}$ model used for *sp*^{2} hybridized acceptors in *HBv2*, we explored several other models. The first model we considered, $h^a_{BA_\chi}$, applied a multiplicative term for all *sp*^{2} hybridized acceptors that depended on BA_χ to scale the other three terms of the energy function:

$$E^a_{\mathrm{HBv2}} = \min(0, h^a_{BA_\chi}\left(f^a_{AH_{dis}} I_{AHD} I_{BAH} + g^a_{AHD} I_{AH} I_{BAH} + h^a_{BAH} I_{AH} I_{AHD}\right)) \quad (1)$$

$$h^a_{BA_\chi}(BA_\chi) = p^a_1\ (2BA_\chi) + \ p^a_1 \quad (2)$$

with tunable parameters p^a_1 and p^a_1. (We had already removed the short- and long-range distinction from the potential, though we still relied on the fade functions). The form of $h^a_{BA_\chi}$ encodes the *sp*^{2} preference by peaks at BA_χ=0° and 180° and troughs at 90° and 270°. This functional form and the one used by Morozov *et al.* (2003)—i.e. $E_{\text{Morozov}} = f_{AH_{dis}} + g_{AHD} + h_{BAH} + i_{BA_\chi}$, where i_{BA_χ} was a knowledge-based polynomial of BA_χ—both suffer from a numerical instability in the measurement of BA_χ when BAH is near 180°: small variations in the hydrogen coordinate produce large variations in the measured BA_χ dihedral and thus large variations in the energy. BA_χ and BAH have to be considered simultaneously.

We next considered three functions, E^b_{HBv2}, E^c_{HBv2}, and E^d_{HBv2}, based on a single functional form and controlled by three parameters, p^i_1, p^i_2, and p^i_3 for $i \in \{b, c, d\}$:

$$E^i_{\text{HBv2}} = \min(\ 0, h^i_{BAH,BA_\chi} * (\ fI_{AHD}I_{BAH} + gI_{AH}I_{BAH})) \quad (3)$$

$$h^i_{BAH,BA_\chi} = \ p^i_1 + p^i_2(\frac{\cos(\pi-3BAH)+1}{2}) + \ p^i_3\frac{\cos(2BA_\chi)+1}{2}(1 - \frac{\cos(\pi-3BAH)+1}{2}) \quad (4)$$

In this model, the larger the value of h^i_{BAH,BA_χ}, the stronger the hydrogen bond. h^i_{BAH,BA_χ} combines the BA_χ preference for $0°$ or $180°$ from the $\cos(2BAH_\chi)$ term, and the BAH preference for $120°$ from the $\cos(\pi - 3BAH)$ term. As BAH approaches $180°$, the contribution from BA_χ diminishes, thereby avoiding any numerical instability. The three parameters represent three steps upward from 0; p^i_1 is a first step upwards and is applied uniformly; p^i_2 is a step upwards for values of BAH near 120°; p^i_3 is a final step upwards for values of BAH near 120° and BA_χ near either $0°$ or $180°$ (Fig. S.3.5). The h^i_{BAH,BA_χ} functions were applied to all *sp*^{2}-hybridized acceptors for all interactions, including those where the native distributions for the BAH did not show a peak at $120°$, such as the backbone/backbone H-bonds in β-sheets. This is quite different from *HBv1* where the h^1_{BAH} polynomial used to define backbone/backbone H-bonds with sequence separation less than five placed the optimal BAH angle at 150°, and the backbone/backbone H-bonds with sequence separation greater than or equal to five placed the optimal BAH angle at 158°, giving α-helical and β-sheet contacts different definitions of what is ideal. The E^b_{HBv2} function was defined by $p^b_1 = 1$, $p^b_2 = 0$, and $p^b_3 = 1$; p^b_1 and p^b_3 were fit empirically, and p^b_2 was held fixed at 0. This functional form gives no reward for having BAH near 120° if BA_χ was not also near 0° or 180° (Fig. S.3.5A&D)

E^b_{HBv2} reproduced many of the same features of the native distributions that E_{HBv2} would eventually also reproduce. It reproduced the sinusoidal behavior of BA_χ for carboxylate-hydroxyl and carboxamide-hydroxyl H-bonds and the difference in amplitudes between them. It reproduced the beetle shape in the Lambert-azimuthal projection for long-range backbone-backbone hydrogen bonds as well as the non-sinusoidal BA_χ distribution. It reproduced the strong preference for a BA_χ dihedral of 180 that carboxyl-guanidino H-bonds show.

E^b_{HBv2} failed to reproduce some of the native distributions; the BA_χ distributions for most sidechain-backbone H-bonds displayed a stronger sinusoidal behavior than is observed

in natives, such that E^{b}_{HBv2} distributions are out-of-phase with the relatively flat distributions seen in natives. For carboxyl-hydroxyl hydrogen bonds, the Lambert-azimuthal projection shows E^{b}_{HBv2} clustering density too tightly near the peak at $BAH = 120°$ and $BA_{\chi} = 0° \text{ or } 180°$, whereas the natives displayed a broader distribution that fanned out. It also failed to recapitulate the backbone-acceptor/amino-donor (i.e. lysine) distribution observed in the Lambert-azimuthal projection, which places significant density near $BAH = 180°$. To examine whether these last two deficiencies might be a result of the functional form, we tested two forms, E^{c}_{HBv2}, and E^{d}_{HBv2}, which varied in their treatment of p_2^i: E^{c}_{HBv2} set p_2^c to be positive (Fig. S.3.5B&E) to hopefully spread out the carboxyl-hydroxyl distribution (making BAH near $120°$ favorable outside of the range near $BA_{\chi} = 0 \text{ or } 180°$); E^{d}_{HBv2} set p_2^d to be negative (Supp. Fig. S.3.5C&F) to hopefully put more density near $BAH = 180°$ for the backbone-amino hydrogen bonds. Though we did not extensively tune the three parameters controlling the shape of these potentials, it was rapidly clear that they degraded the beetle distribution for backbone/backbone H-bonds in β-sheets (Fig. S.3.6). We did not to pursue E^{c}_{HBv2} and E^{d}_{HBv2} any further.

These functional forms represented a significant departure from E_{HBv1}, besides their dependence on BA_{χ}, in that they coupled multiple geometries together. In E_{HBv1}, an improvement in the AHD geometry has no effect on energetic contribution from the BAH geometry, and vice versa. In E^{b}_{HBv2}, however, an improved AHD angle puts pressure on BAH to improve; the derivatives *wrt* BAH are dependent on AHD. This coupling seemed apparent in the joint distribution of BAH and AHD angles for bidentate carboxyl-guanidino (ASP to ARG) H-bonds. In the native structures, there was a broad distribution of BAH and AHD angles with a broad peak near $AHD{\sim}145°$ and $BAH{\sim}140°$ (Supp. Fig. 3.7A). In forming a pair of hydrogen bonds, E^{b}_{HBv2} prefers to create one H-bond with nearly ideal angles—$AHD{\sim}180°$ and $BAH{\sim}120°$—and the second H-bond with worse angles – $AHD{\sim}150°$ and $BAH{\sim}150°$. This can be seen in the presence of two separate peaks (Supp. Fig. 3.7B). It appeared that E^{b}_{HBv2} was creating one strong hydrogen bond and one weak hydrogen bond. In other words, it would load all the bad geometry into one of the hydrogen bonds, and keep the other one ideal, as might be expected from its functional form. An additive functional form would allow the energy to slosh between the two hydrogen bonds formed in a bidentate interaction, but in the multiplicative form a good AHD angle has to be paired with a good BAH angle to really count. Observing the presence of the two peaks in this joint AHD/BAH distribution, we sought out a functional form, E_{HBv2}, that decoupled AHD and BAH; we would only later understand the actual reason for the distribution that E^{b}_{HBv2} produced.

E_2 did not change the distribution of BAH vs AHD for ARG/ASP bidentate hbonds, even though that is why we pursued this functional form over E^{b}_{HBv2}. It turns out that our problem stemmed from the “ideal” bond angles used to define the atomic positions in Arginine. We noticed that the bimodal distribution disappeared from the models generated by *ElecHBv2* (described in detail later), which included a fix for these angles that had been proposed earlier were but not enabled by default (Song 2011). Fixing the ideal bond geometries for *HBv2* removes the bimodal distribution and furthermore, reverting to the old (bad) angles to *ElecHBv2* reinstates the bimodal distribution. Indeed, the E^{b}_{HBv2} functional form delivers a single broad distribution when the improved bond angles are used (Fig. S.3.7). We cannot thus conclude that E^{b}_{HBv2} is a worse functional form than E_{HBv2}.

S.4.1.3 sp^2 functional form

We formulated our sp^2 potential as follows. In E_{HBv2} from Equation 2, the function $h^2_{\mathrm{BAH,BA}_\chi}$ for sp^2 hybridized acceptors is given by

$$h^2_{\mathrm{BAH,BA}_\chi} = \alpha F + (1-\alpha)G \tag{5}$$

$$\alpha = \frac{1}{2}\cos\left(2\mathrm{BA}_\chi\right) + \frac{1}{2} \tag{6}$$

$$F \begin{cases} \frac{d}{2}\cos\left(3(\pi - \mathrm{BAH})\right) + \frac{d-1}{2} & \text{if BAH} > 2\pi/3 \\ \frac{m}{2}\cos\left(\pi - \frac{\frac{2}{3}\pi - \mathrm{BAH}}{l}\right) + \frac{m-1}{2} & \text{if } 2\pi/3 \geq \mathrm{BAH} \geq \pi(2/3 - l) \\ m - \frac{1}{2} & \text{o.w.} \end{cases} \tag{7}$$

$$G \begin{cases} d - \frac{1}{2} & \text{if BAH} > 2\pi/3 \\ \frac{m-d}{2}\cos\left(\pi - \frac{\frac{2}{3}\pi - \mathrm{BAH}}{l}\right) + \frac{m+d-1}{2} & \text{if } 2\pi/3 \geq \mathrm{BAH} \geq \pi(2/3 - l) \\ m - \frac{1}{2} & \text{o.w.} \end{cases} \tag{8}$$

where the α function interpolates between F and G depending on the value of BA_χ (Figure 2).

BAH: $\mathrm{BAH} = \pi$ when the Base-Acceptor-Hydrogen atoms are co-linear and $\mathrm{BAH} = \frac{\pi}{2}$ when they form a right angle.

BA_χ: BA_χ is defined by the Abase2-Base-Acceptor-Hydrogen atoms the sp^2 orbitals are in the Abase2-Base-Acceptor plane. For backbone acceptors Abase2=Cα.

d: d parameter is the distance define by HBondOptions::sp2_BAH180_rise(), set by the -corrections::score::hb_sp2_BAH180_rise flag and defaults to 0.75.

m: m is the distance from the minimum to maximum values F and defaults to 1.6

l: l is the period/2 of the $\mathrm{BAH} = 120°$ to $\mathrm{BAH} = 60°$ piece of F, empirically fit to be 0.357

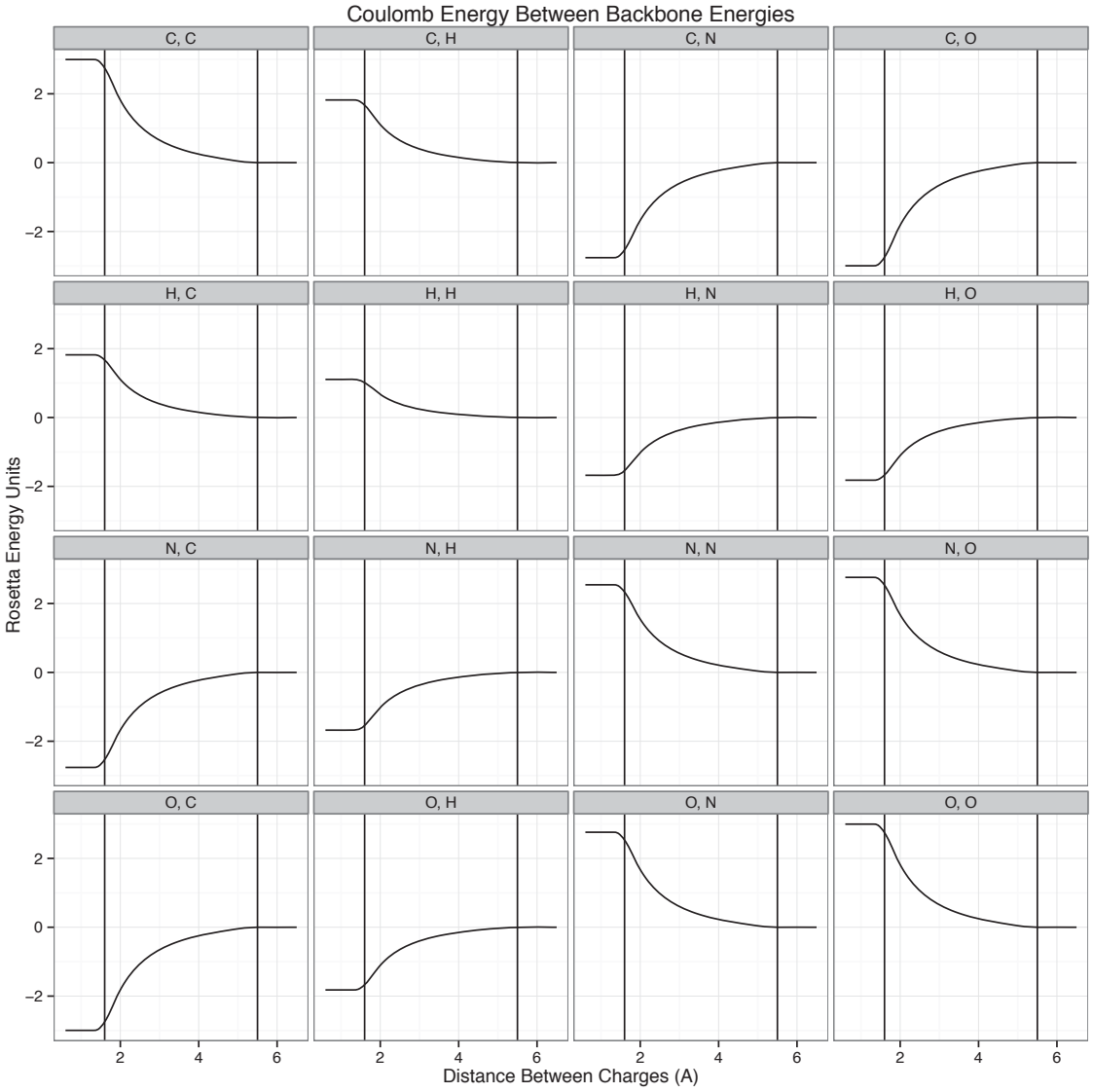
S.4.2 Electrostatics Model

Here is reference implementation of the Elec model in the R language. The atomic charges are in the residue_atoms.csv supplementary file and hyper parameters of parameters min_dist = 1.6, max_dist = 5.5 and dielectric of 10.

```
#returns:
#   f''(x1), f''(x2)
make_second_derivative <- function(
  x1, x2,
  y1, y2,  # f’(x1), f’(x2)
  yp1, yp2 # f’(x1), f’(x2)
){
  ypp1 = -0.5
  u = (3.0/(x2-x1)) * ((y2-y1)/(x2-x1)-yp1)
  qn = 0.5
  un = (3.0/(x2-x1))*(yp2-(y2-y1)/(x2-x1))
  ypp2 = (un-qn*u)/(qn*y21+1.0)
  ypp1 = ypp1 * y22 + u
  c(ypp1, ypp2)
}

# a = x1
# b = x2
# c = y2
# d = y1
# e = f''(x2)
# f = f''(x1)
cubic_polynomial_from_spline <- function(
  a, b, c, d, e, f){
  c(
    c0=((b**3*f - a**3*e)/(b-a) + (a*e - b*f) * (b-a))/6 + (b*d - a*c)/(b-a),
    c1=(3*a*a*e/(b-a) - e*(b-a) + f*(b-a) - 3*b*b*f/(b-a))/6 + (c - d)/(b-a),
    c2 = (3*b*f - 3*a*e)/(6*(b-a)),
    c3 = (e - f)/(6*(b – a)))
}


eval_cubic_polynomial <- function(x, poly){
  ((poly[4]*x+poly[3])*x+poly[2])*x+poly[1];
}




# d = distance between charges (q1, q2)
# n_bond_sep = number of covalent bonds between q1 and q2
# die = dielectric
# min_dist, max_dist = where to where to evaluate energy
# smoothed +/- .25 around min_dist and max_dist
coulomb_energy  <-  function(d,  q1,  q2,  n_bond_sep,  die=10.0,  min_dist=1.6,
max_dist=5.5){
  if(n_bond_sep < 4L){
    return(0.0)
  } else if(n_bond_sep == 5L){
    sep_weight = 0.2
  } else {
    sep_weight = 1.0
  }

  C0 <- 322.0637
```

```
  C1 <- C0 / die
  C2 <- C1 / (max_dist * max_dist)


  low_poly_start <- min_dist - 0.25
  low_poly_start_score <- C1/(min_dist**2) - C2
  low_poly_start_deriv <- 0

  low_poly_end <- min_dist + 0.25
  low_poly_end_score <- C1/(low_poly_end * low_poly_end) - C2;
  low_poly_end_deriv <- -2*C1/(low_poly_end**3);

  low_poly_deriv2 <- make_second_derivative(
    low_poly_start, low_poly_end,
    low_poly_start_score, low_poly_end_score,
    low_poly_start_deriv, low_poly_end_deriv)
  low_poly_start_deriv2 <- low_poly_deriv2[1]
  low_poly_end_deriv2 <- low_poly_deriv2[2]

  low_poly <- cubic_polynomial_from_spline(
    low_poly_start, low_poly_end,
    low_poly_end_score, low_poly_start_score,
    low_poly_end_deriv2, low_poly_start_deriv2 )

  hi_poly_start <- max_dist - 1.0
  hi_poly_start_score = C1/(hi_poly_start**2) - C2;
  hi_poly_start_deriv = -2*C1/(hi_poly_start**3);

  hi_poly_end <- max_dist
  hi_poly_end_score <- 0
  hi_poly_end_deriv <- 0

  hi_poly_deriv2 <- make_second_derivative(
    hi_poly_start, hi_poly_end,
    hi_poly_start_score, hi_poly_end_score,
    hi_poly_start_deriv, hi_poly_end_deriv)
  hi_poly_start_deriv2 <- hi_poly_deriv2[1]
  hi_poly_end_deriv2 <- hi_poly_deriv2[2]

  hi_poly <- cubic_polynomial_from_spline(
    hi_poly_start, hi_poly_end,
    hi_poly_end_score, hi_poly_start_score,
    hi_poly_end_deriv2, hi_poly_start_deriv2 );

  if (d > max_dist) {
    return(0.0);
  } else if (d < low_poly_start) {
    return(sep_weight * q1 * q2 * low_poly_start_score);
  } else if (d < low_poly_end) {
    return(sep_weight * q1 * q2 * eval_cubic_polynomial( d, low_poly ));
  } else if (d > hi_poly_start) {
    return(sep_weight * q1 * q2 * eval_cubic_polynomial( d, hi_poly ));
  } else {
    return(sep_weight * q1 * q2 * (C1/(d*d) - C2));
  }
}
```

S.4.2.1 Evaluation of the Elec model over all pairs of backbone atom types. These energies, along with energies involving sidechains are summed and linearly weighted into the overall energy function. The black vertical lines indicate the min_dist and max_dist values, where the smoothing end and begin.



S.4.3 Disulfide Model

Ideally, the energy function should ensure that optimization of near-native conformations converges to the native conformation. For this to happen, the native structures should be at the global minimum in the energy landscape—a necessary condition for that to be true is that the energy gradient of native conformations must be zero. Inspection of per-atom gradients by energy term in Rosetta-refined crystal structures (Dimaio 2013) revealed large gradients in the disulfide-bond energy term. To further investigate, we compared the distributions of disulfide-bond features (the SG-SG distance, Cβ-SG-SG angle, Cα-Cβ-SG-SG dihedral, and Cβ-SG-SG-Cβ dihedral) estimated from the Top8000, yielding 1920 disulfide bonds (*native*), with the distributions from 50 small disulfide-containing proteins containing 191 disulfide bonds that were refined with Rosetta (*old*). We observed a substantial discrepancy in both the SG-SG distance and Cβ-SG-SG angle distributions (Fig. S.4.3.1).

To improve the disulfide model, we developed a new functional form of a skewed Gaussian to model the SG-SG distance feature, and mixtures of 1, 2, and 3 von Mises functions, respectively, to model the Cβ-SG-SG angle and the Cα-Cβ-SG-SG and Cβ-SG-SG-Cβ dihedrals (Fig. S.4.3.2). Initially, we fit the potential to match the native distributions. To balance against the other terms in the energy function, we then added a constant offset to the disulfide energy. We fit this offset by refining the 50 small disulfide-containing proteins after severing their disulfide bonds, and chose the smallest offset that favored the disulfide-linked conformations; this gave a value of -2 energy units.

Figure S.4.3.1: Disulfide-bond feature distributions comparing the Top8000 chains sample source (native) against 50 small disulfide-containing conformations refined with Rosetta using the Score12 disulfide model (old) and the updated disulfide model (new). Each cell plots the estimated distributions of disulfide bond geometric features: (UPPER-LEFT) SG-SG distance (UPPER RIGHT) Cβ-SG-SG angle (LOWER LEFT) Cα-Cβ-SG-SG dihedral angle (LOWER RIGHT) Cβ-SG-SG-Cβ dihedral angle.

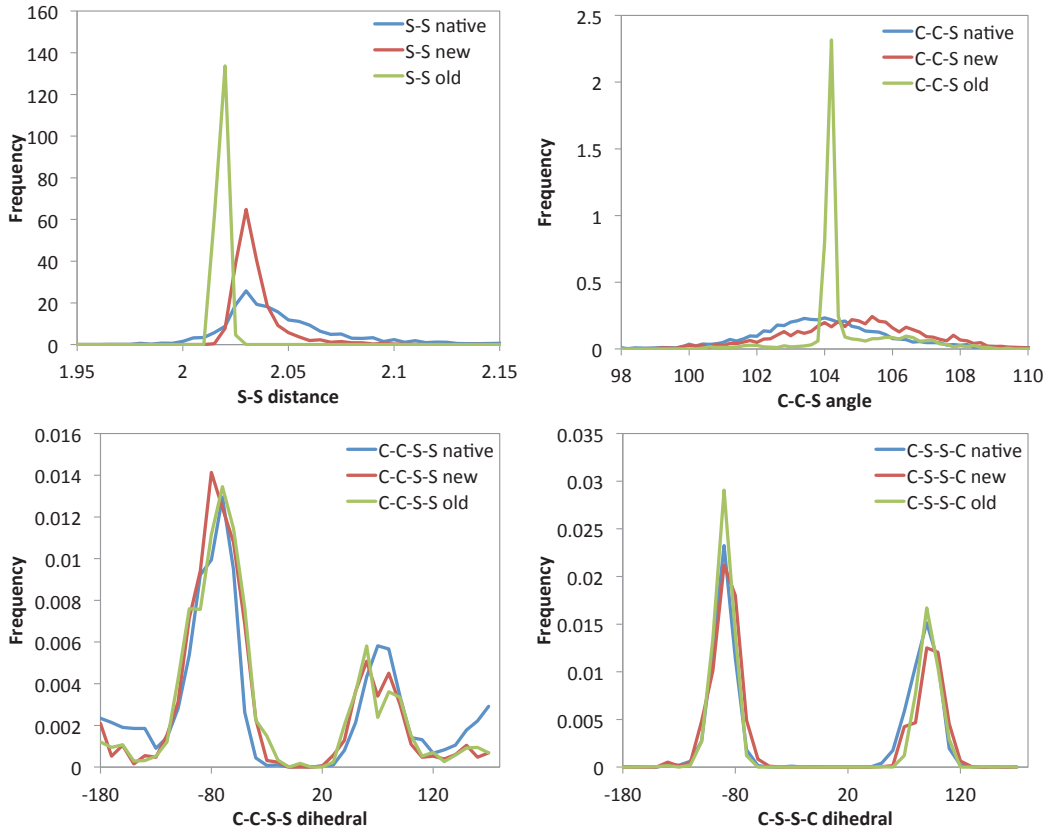
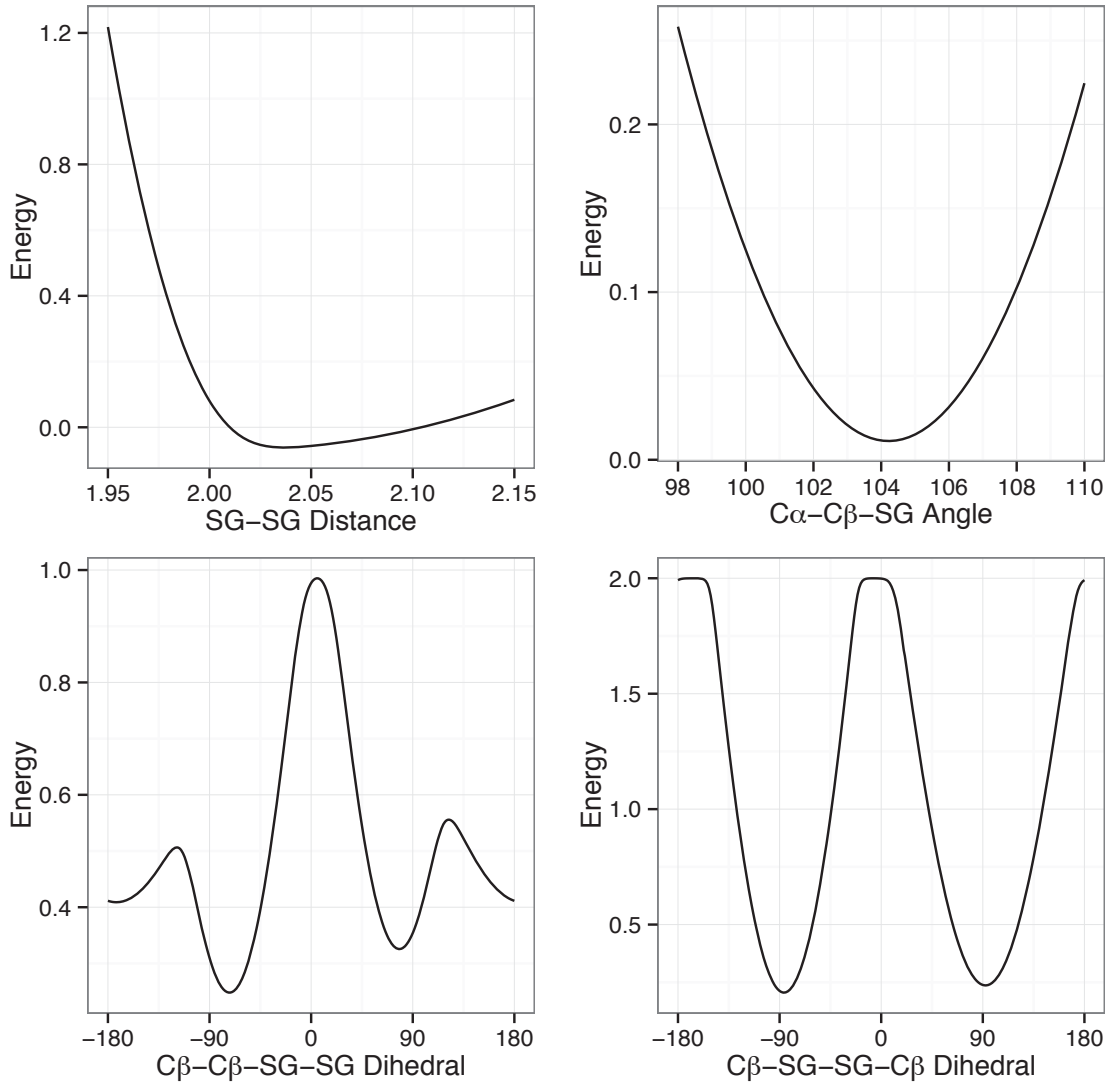


Figure S.4.3.2: Components of the new disulfide bond potential



S.5 Sample Sources and Energy Functions

S.5.1) Overview of Score Functions

Score12

Score12 was the standard Rosetta score function 2003-2013(Rohl, Strauss, Misura, & Baker, 2004).

The H-bond model is based on the model presented in Kortemme (2003)(Kortemme, Morozov, & Baker, 2003), which we call *HBv1* here. It consists of a linear combination of 1d potentials over H-bond lengths and angles derived from distributions estimated from a survey of 698 high-resolution crystal structures using a knowledge-based potential methodology. The functional form is $E_{\mathrm{HBv1}} = w_{\mathrm{env}} \min\left(0, f^1_{AH_{dis}} + g^1_{AHD} + h^1_{BAH}\right)$ (Eq. 1). The full functional form is described in (S.3.22).

Baseline Corrections on Score12 (HBv1)

1) Idealized coordinates derived from the expected value for bond lengths, angles, and dihedral angles collected from structures (Song, Tyka, Leaver-Fay, Thompson, & Baker, 2010)

2) An updated disulfide potential (S.4.3)

3) The 2010 Dunbrack rotamer library(Leaver-fay et al., 2013; Shapovalov & Dunbrack, 2011)

4) Use of bicubic interpolation of backbone dependent knowledge based potentials(Leaver-fay et al., 2013)

5) Reversion of atomic LK_DGFREE(dgfree) parameters to the EEF1 parameters(Lazaridis & Karplus, 1999):

NH20(HN2); -10 to -7.8
Narg(NC2); -11 to -10
OH(OH1); -6.77 to -6.70
ONH2(O); -10 to -5.85

6) Analytic evaluation (rather than table lookup) of the Lennard Jones (fa_rep/fa_atr) and EEF1 (fa_sol) energy terms, which allows for correct derivative computation(Leaver-fay et al., 2013).

HBv2

HBv2 includes

1) Baseline corrections

2) Rotameric sampling for hydroxyl-chi dihedral angles for (SER/THR): from (-60, 60, 180) to (0, 20, 40, …, 340),

3) Adjust of Lennard-Jones parameters between hydroxyl O and H atoms and H-Bond acceptor atoms(Leaver-fay et al., 2013),

4) *sp2* H-Bond functional form is $E_{\mathrm{HBv2}} = w_{\mathrm{env}}\, s\left(f^2_{AH_{dis}} + g^2_{AHD} + h^2_{BAH,BA_\chi}\right)$ (Eq. 2), it and the fit parameters are discussed in main document.

Elec

The *Elec* H-bond model is a simple Coulombic model of electrostatics; partial charges are assigned to atoms following CHARMM 19(Brooks et al., 1983), the dielectric is proportional to $1/r$(Hingerty, Ritchie, Ferrell, & Turner, 1985; Warshel, Russell, & Churg,

1984) over the range, and both short- and long-range cutoffs are smoothed using spline interpolation.

Elec includes
1) Baseline corrections
2) Smoothed distance dependent dielectric Coulomb potential (S.4.2).
3) No HBond term

ElecHBv1
ElecHBv1 includes
1) Baseline corrections
2) Elec corrections
3) HBv1 term

ElecHBv2
ElecHBv2 includes
1) Baseline corrections
2) Elec electrostatics model
3) HBv2 H-bond model
4) The HBv2 parameters were refit so relaxed natives recapitulate the native distribution, described in the section (2.11).

S.5.2) Flags and Score Weight files for Energy functions

Reference weights for each weight set have been refit using the OptE protocol (Leaver-fay et al., 2013), valid for Rosetta (Jul 2014).

Score12
Flags:
```
-restore_pre_talaris_2013_behavior
```

Weights:
```
METHOD_WEIGHTS ref 0.102507 -0.0316374 -0.504387 -0.629043
1.47152 -0.406181 1.01085 0.401181 -0.473385 0.187405 -0.203416 -
0.776415 -1.00694 -0.795115 -0.772679 -0.336971 -0.244449
0.311957 1.571 1.1242
fa_atr 0.8
fa_rep 0.44
fa_sol 0.65
pro_close 1
fa_pair 0.49
hbond_sr_bb 0.585
hbond_lr_bb 1.17
hbond_bb_sc 1.17
hbond_sc 1.1
dslf_ss_dst 1
dslf_cs_ang 1
dslf_ss_dih 1
```

```
dslf_ca_dih 1
rama 0.2
omega 0.5
fa_dun 0.56
p_aa_pp 0.32
ref 1
```

Elec

Flags:

```
-restore_pre_talaris_2013_behavior
-corrections::chemical:icoor_05_2009
-analytic_etable_evaluation
-use_bicubic_interpolation
-dun10
-set_atom_properties fa_standard:ONH2:LK_DGFREE:-5.85
fa_standard:NH2O:LK_DGFREE:-7.8 fa_standard:Narg:LK_DGFREE:-10.0
fa_standard:OH:LK_DGFREE:-6.70
-smooth_hack_elec
-hackelec_min_dis 2.0
-hackelec_r_option false
```

Weights:

```
METHOD_WEIGHTS ref 0.102507 -0.0316374 -0.504387 -0.629043
1.47152 -0.406181 1.01085 0.401181 -0.473385 0.187405 -0.203416 -
0.776415 -1.00694 -0.795115 -0.772679 -0.336971 -0.244449
0.311957 1.571 1.1242
fa_atr 0.8
fa_rep 0.44
fa_sol 0.75
pro_close 1
hack_elec 0.70
hbond_sr_bb 0
hbond_lr_bb 0
hbond_bb_sc 0
hbond_sc 1.1
dslf_fa13 1.0
rama 0.2
omega 0.5
fa_dun 0.56
p_aa_pp 0.32
ref 1
```

HBv1

Flags:

```
-restore_pre_talaris_2013_behavior
-corrections::chemical:icoor_05_2009
-analytic_etable_evaluation
-use_bicubic_interpolation
-dun10
```

```
-set_atom_properties fa_standard:ONH2:LK_DGFREE:-5.85
fa_standard:NH2O:LK_DGFREE:-7.8 fa_standard:Narg:LK_DGFREE:-10.0
fa_standard:OH:LK_DGFREE:-6.70
```

Weights:

```
METHOD_WEIGHTS ref  0.16 1.7 -0.67 -0.81 0.63 -0.17 0.56 0.24 -
0.65 -0.1 -0.34 -0.89 0.02 -0.97 -0.98 -0.37 -0.27 0.29 0.91 0.51
fa_atr 0.8
fa_rep 0.44
fa_sol 0.75
fa_intra_rep 0.004
fa_pair 0.49
fa_plane 0
fa_dun 0.56
ref 1
hbond_lr_bb 1.17
hbond_sr_bb 0.585
hbond_bb_sc 1.17
hbond_sc 1.1
p_aa_pp 0.32
dslf_fa13 1.0
pro_close 1.0
omega 0.5
rama 0.2
```

HBv2

Flags:

```
-restore_pre_talaris_2013_behavior
-corrections::chemical:icoor_05_2009
-analytic_etable_evaluation
-use_bicubic_interpolation
-dun10
-set_atom_properties fa_standard:ONH2:LK_DGFREE:-5.85
fa_standard:NH2O:LK_DGFREE:-7.8 fa_standard:Narg:LK_DGFREE:-10.0
fa_standard:OH:LK_DGFREE:-6.70
-hbond_params sp2_params
-hb_sp2_chipen
-hb_sp2_BAH180_rise 0.75
-hbond_measure_sp3acc_BAH_from_hvy
-lj_hbond_hdis 1.75
-lj_hbond_OH_donor_dis 2.6
-expand_st_chi2sampling
```

Weights:

```
METHOD_WEIGHTS ref 0.242542 0.0525932 -0.49322 -0.640135 1.13617
-0.345874 0.871098 0.441403 -0.429774 0.203928 -0.127723 -0.74838
-0.593685 -0.768373 -0.84356 -0.269817 -0.17786 0.370444 1.29093
0.829296
fa_atr 0.8
fa_rep 0.44
```

```
fa_sol 0.75
fa_intra_rep 0.004
pro_close 1
fa_pair 0.49
hbond_sr_bb 1.17
hbond_lr_bb 1.17
hbond_bb_sc 1.17
hbond_sc 1.1
dslf_fa13 1.0
rama 0.2
omega 0.5
fa_dun 0.56
p_aa_pp 0.32
yhh_planarity 0.5
ref 1
```

ElecHBv2

Flags:

```
# This is Talaris2013 is the default
```

Weights:

```
METHOD_WEIGHTS ref 0.592942 0.354993 -1.28682 -1.55374 0.43057
0.140526 0.357498 0.831803 -0.287374 0.602328 0.158677 -0.94198 -
0.219285 -1.17797 -0.14916 0.176583 0.16454 0.744844 0.92933
0.131696
fa_atr 0.8
fa_rep 0.44
fa_sol 0.75
fa_intra_rep 0.004
hack_elec 0.7
pro_close 1
hbond_sr_bb 1.17
hbond_lr_bb 1.17
hbond_bb_sc 1.17
hbond_sc 1.1
dslf_fa13 1.0
rama 0.2
omega 0.5
fa_dun 0.56
p_aa_pp 0.32
yhh_planarity 0.5
ref 1
```

References:

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, *4*(2), 187–217. doi:10.1002/jcc.540040211

Hingerty, B. E., Ritchie, R. H., Ferrell, T. L., & Turner, J. E. (1985). Dielectric effects in biopolymers: The theory of ionic saturation revisited. *Biopolymers*, *24*(3), 427–439. doi:10.1002/bip.360240302

Kortemme, T., Morozov, A. V., & Baker, D. (2003). An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein–Protein Complexes. *Journal of Molecular Biology*, *326*(4), 1239–1259. doi:10.1016/S0022-2836(03)00021-4

Lazaridis, T., & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins Structure Function and Genetics*, *35*(2), 139–45.

Leaver-fay, A., O'Meara, M. J., Tyka, M., Jacak, R., Song, Y., Kellogg, E. H., … Kuhlman, B. (2013). *Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement* (Vol. 523). doi:10.1016/B978-0-12-394292-0.00006-0

Rohl, C., Strauss, C., Misura, K., & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods in Enzymology*, *383*, 66–93.

Shapovalov, M. V, & Dunbrack, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure (London, England : 1993)*, *19*(6), 844–58. doi:10.1016/j.str.2011.03.019

Song, Y., Tyka, M., Leaver-Fay, A., Thompson, J., & Baker, D. (2010). Structure guided forcefield optimization. *Proteins: Structure, Function, and Bioinformatics*.

Warshel, a, Russell, S. T., & Churg, a K. (1984). Macroscopic models for studies of electrostatic interactions in proteins: limitations and applicability. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(15), 4785–9.

S.6 Benchmark Details

S.6.1 Relax Native Runtime

To test the impact of the score functions studied in this work on the run-time we created a relax-native benchmark. Structures were selected from the Protein Databank April 19'th 2014 with the following criteria: deposited since 2012, resolution < 1.5 Å, no ligands, no modified residues, has electron density link in the EDS, zero disulfide bonds, and filtered to have at most 40% sequence identity. This returned a list of 35 native structures.

3VQF	4CIL	4GAI	4JMI
3VZ6	4DMV	4GEI	4JZ5
3W24	4DQ7	4GMQ	4JZQ
3ZBD	4DT4	4GS3	4KEE
3ZNY	4EEW	4HJP	4LJ1
4B6G	4EF0	4HS5	4LTT
4B89	4EZA	4I84	4M9K
4B9G	4G3N	4IC4	4NI6
4B9I	4G8D	4J5Q	

Each native was relaxed 5 times with the FastRelax protocol, with a separate execution of Rosetta per target. The mean and standard deviation of the total runtime for each target was computed and is shown in the following table:

Score Function	**Mean seconds**	**StdDev seconds**
Score12'	2,801.6	4,974.8
HBv1	2,776.7	5,016.0
HBv2	1,867.1	2,786.3
Elec	3,302.6	5,876.7
ElecHBv2	2,499.5	3,394.9

The FastRelax protocol includes stages of full atom gradient-based minimization (Minimize) and fixed-backbone stochastic discrete-sampling of sidechain conformations (Repack). Factors that contribute to changes in run time include the computational cost of each energy function evaluation and the "smoothness" of the energy function affecting the rate of convergence during each minimization. These results indicate the *ElecHBv2* is not slower than *Score12'*.

S.6.2 Scientific Benchmark Methods

S.6.2.1 Relax Native Recovery

We use the FastRelax protocol to sample near native conformations that are optimized for the Rosetta energy function. The FastRelax protocol iterates between repacking sidechains and performing quasi-Newton minimization of torsional degrees of freedom while ramping in five steps the strength of the repulsive component of the Lennard-Jones term from $1/10^{th}$ up to full strength, cycling from low- to high-strength repulsion three times(Khatib et al., 2011).

We chose this sampling protocol over less aggressive protocols (such as only performing minimization) to allow the structures to escape local minima in the energy landscape, and over more aggressive protocols (such as trying to fold proteins from extended chains using Rosetta's AbRelax protocol(Bradley, Misura, & Baker, 2005)) in the interest of focusing our efforts on matching distributions where native-like contacts are possible: an *ab initio* folding protocol might emphasize easy-to-form local contacts over harder-to-form long-range contacts and misrepresent the deficiencies in an energy function.

Having a structure that is locally optimal for the energy function is often important starting point for structure design tasks(Nivón, Moretti, & Baker, 2013). While it is possible to resort to explicit structural restraints to remain close to the native structure, ideally the energy function should move the native conformation as little as possible. Additionally having the Rosetta energy function consistent with native conformations is useful for and solving crystal structures(DiMaio et al., 2013).

S.6.2.2 Monomer Sequence Recovery

This protocol first builds rotamers, taken from 2010 Dunbrack rotamer library supplemented with extra samples for the first two χ dihedrals taken at plus and minus one standard deviation from the mean for the rotameric χ (those dihedrals where the two center atoms that define it are both sp3 hybridized), and halfway between the mean χ angle for each rotamer bin and either bin boundary for the non-rotameric χ angles (those dihedrals where one of the two center atoms that define it is sp2 hybridized, e.g. χ_2 from phenylalanine). Following rotamer creation, the protocol employs the multi-cool simulated annealer(Leaver-Fay, Jacak, Stranges, & Kuhlman, 2011) to optimize the rotamer and amino acid assignment. This annealer, like Rosetta's default annealer(Kuhlman & Baker, 2000), considers random rotamer substitutions, evaluates the change in energy induced by the rotamer substitution, and then uses the current "temperature" and the Metropolis criterion to either accept or reject the substitution. The multi-cool annealer has an altered temperature schedule from Rosetta's default annealer and spends more time at very low temperatures

S.6.2.3 Interface Sequence Recovery

The complexes were chosen from X-ray crystal structures deposited in the Protein Databank having resolution less than 2 Å, bond length outliers in less than 5% of the residues (as defined by MolProbity(Chen et al., 2010)), a MolProbity score less than 2.0,

run.sh

```
#!/usr/bin/env bash
```

```
set -e

rosetta_path=$HOME/GIT/Rosetta/main

basename=$PWD
jobname=$(basename $PWD)
outdir="outputs"

nprocs=173
exepath="${rosetta_path}/source/bin/rosetta_scripts.mpi.linuxgccr
elease"

queue="day"
bsubcmd="bsub -q $queue -n $nprocs -J $jobname -o stdout.%J -e
stderr.%J -a mvapich mpirun"

cmd="$bsubcmd $exepath @score.flags @protocol.flags \
  -mpi_tracer_to_file $jobname.tracers"

[ -e "$PWD/${outdir}/" ] && rm -r "$PWD/${outdir}/"
mkdir $outdir
cd $outdir
cp ../$0 launcher_used
cp ../*flags .
cp ../*xml .
cp ../*list .
echo $cmd > cmd_run
$cmd
```

protocol.flags

```
-database Rosetta/main/database
-l kan_plus.list
-nstruct 1
-jd2:ntrials 20
-ignore_unrecognized_res
-linmem_ig 10
-skip_set_reasonable_fold_tree
-out:pdb_gz true
-overwrite
-options:user
-parser:protocol interface_packmin.xml
```

interface_packmin.xml

```
<ROSETTASCRIPTS>
      <SCOREFXNS>
            <talaris_optE weights=KH_talaris_kanplus_20130602 />
      </SCOREFXNS>
      <TASKOPERATIONS>
            <InitializeFromCommandline name=init />
            <RestrictToInterfaceVector name=interface_vector
jump=1 />
```

```
	</TASKOPERATIONS>

	<MOVERS>
		<AtomTree name=docking_tree docking_ft=1 />

		<MinMover name=min_all
			scorefxn=talaris_optE
			chi=1
			bb=1
			jump=1 />

		<PackRotamersMover name=pack_inter
			scorefxn=talaris_optE
			task_operations=init,interface_vector />

		<ParsedProtocol name=bind>
			<Add mover=pack_inter />
			<Add mover=min_all />
		</ParsedProtocol>

		<InterfaceAnalyzerMover name=interface_analyzer
			scorefxn=talaris_optE
			packstat=1
			pack_input=0
			pack_separated=0
			jump=1 />
	</MOVERS>

	<APPLY_TO_POSE>
	</APPLY_TO_POSE>
	<PROTOCOLS>
		<Add mover_name=docking_tree />
		<Add mover_name=bind />
		<Add mover_name=interface_analyzer />
	</PROTOCOLS>
</ROSETTASCRIPTS>
```


Bradley, P., Misura, K. M. S., & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science*, *309*(5742), 1868–71. doi:10.1126/science.1113801

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. a, Immormino, R. M., Kapral, G. J., … Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica. Section D, Biological Crystallography*, *66*(Pt 1), 12–21. doi:10.1107/S0907444909042073

DiMaio, F., Echols, N., Headd, J. J., Terwilliger, T. C., Adams, P. D., & Baker, D. (2013). Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nature Methods*, (September), 1–5. doi:10.1038/nmeth.2648

Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popovic, Z., … Players, F. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(47), 18949–53. doi:10.1073/pnas.1115898108

Kuhlman, B., & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(19), 10383–10388.

Leaver-Fay, A., Jacak, R., Stranges, P. B., & Kuhlman, B. (2011). A generic program for multistate protein design. *PloS One*, *6*(7), e20937. doi:10.1371/journal.pone.0020937

Nivón, L., Moretti, R., & Baker, D. (2013). A Pareto-optimal refinement method for protein design scaffolds. *PloS One*, *8*(4), 1–5. doi:10.1371/Citation