**Supplementary Information**

**Increasing the Efficiency and Accuracy of the ABACUS Protein Sequence Design Method**

Peng Xiong1, Xiuhong Hu1, Bin Huang1, Jiahai Zhang1, Quan Chen\*,1, Haiyan Liu\*,1,2,3

1School of Life Sciences, University of Sciences and Technology of China;

2Hefei National Laboratory for Physical Sciences at the Microscale;

3School of Data Science, University of Sciences and Technology of China, Hefei, Anhui 230026, China.

\*To whom correspondence should be addressed.

E-mails: hyliu@ustc.edu.cn, chenquan@ustc.edu.cn

**1. Supplementary Methods**

**1.1. The ABACUS2 energy function**

**(1) Interpolation schemes to obtain the single residue energy and representative points for backbone positions**

For an actual site *p* in a design target, its single position residue type energy is determined from these predetermined tables through interpolation, namely,

, (S

1)

in which *r* is the index of a representative position with structural features and, is a subset of representative points that contains the 6 nearest neighbors of *p* in the structural feature space (1 *SS* type 2 nearest *SAI* values 3 nearest points on the *RAMA* plane), and a weighting factor that inversely correlates with the similarity between the actual site *p* and the representative site *r*, which has been defined as

. (S

2)

The overall set of representative points in the space of local structural features comprise all possible combinations of one the three possible *SS* types (namely, helix, strand and coil, which are determined from backbone hydrogen bond patterns[1]), every member of a set of representative points on the *φ-ψ* plane, and every member of a set of *SAI* values covering the range from 0 to 1 ( a *SAI* value of 0 corresponds to the smallest solvent accessibility and 1 corresponds to the largest solvent accessibility). The representative *φ-ψ* set contains 200 points (Supplementary Figure S1), which have been chosen by minimizing the RMS error of approximating the actual *φ* and *ψ* angles in native proteins with the respective nearest representative points. With the set of points given in Supplementary Figure S1, the final RMS error is 3.84 degree, the *φ-ψ* regions that are more densely populated by native proteins also more densely covered. The set of representative *SAI* values contains 18 values, with the first two being 0.1 and 0.2 and the subsequent 16 values evenly distributed from 0.25 to 1.0. All three structural features combined, the total number of representative points to cover the local structure feature space is .

For each representative position *r*, an ensemble of backbone positions that are structurally similar to *r* have been collected from the training proteins. Then and subsequently have been estimated on the basis of the native residue types at these positions. The criteria for local structure similarity include the same secondary structure type as well as sufficiently small differences in *φ*, *ψ* and in *SAI*. The thresholds for maximum *φ*, *ψ* and *SAI* differences have been chosen adaptively, with the initial threshold for the distance on the *φ-ψ* plane being 7°, and for the difference in *SAI* being 0.1. If for any representative point the resulting number of structurally similar template positions was less than 400, the thresholds for that point were gradually increased until more than 400 template positions could be retrieved. With this adaptive approach, we expect to reach a proper balance between the relevance of the retrieved template positions with respect to the target position and the statistical uncertainty in the estimated conditional probabilities.

**(2) Interpolation schemes to obtain the residue pair energy and representative points for backbone position pairs**

For a representative pair of positions comprising positions *s* and *t*, has been estimated by first collecting positions pairs in training proteins that match the respective *SS* types and *SAI* values of both *s* and *t* as well as *SEPst*. We note such a matching training position pair as *mst* comprising positions and . The criteria for *SAI* matching are for both and . Then the following adaptive kernel approach has been used to take into account the extent of matching between and ,

, (S

3)

in which is the total number of training pairswith their *SS* and *SAI* matching those of *s* and *t*, is 1 if and refer to the same residue type and 0 otherwise. The refers to the root mean square deviations of atomic positions between the pair *st* and the pair *msmt*. The atoms used for the RMSD calculation include backbone atoms N, Cα, and C, and Cβ as well the pseudo Cγ atoms (see the pseudo sidechain model below). The kernel function *h(RMSD)* takes the following form,

(S

4)

The parameter is 0.05 Å for local position pairs and 0.08 Å for nonlocal position pairs. The parameter *RMSD0* is adaptively changed, being initially 0.02 Å for local pairs and 0.5 Å for nonlocal pairs, and gradually increased until the denominator in the right part of formula (S3) gets larger than 4000.

The interpolation formula to obtain *e2* for an actual site pair comprising sites *p* and *q* is

. (S5)

Again, is defined to be inversely correlated with the extent of similarity between the representative site pair *st* and the actual site pair *pq*. The subset contains the representative points that are the 12 nearest neighbors of *pq* (the 3 nearest points on the 2-dimensional *SAI* plane the 4 nearest points in the *RGEOM* space). Besides the exact matches of *SS* types and of sequence separations (for the local pairs) for to be nonzero, also depends on the differences in *SAI* and *RGEOM* between *st* and *pq* in the following form,

*.*  (S

6)

To choose the representative backbone site pairs, there are 9 possible SS type combinations and 5 types of sequence separations. In addition, 12 points have been chosen on the two dimensional plane of the SAI values (Supplementary Figure S2). These points have been chosen in the same spirit as the choice of representative points on the φ-ψ plane, namely, to minimize the RMS error of approximating the SAI values of the actual site pairs in training proteins by those of the representative points. For each given SS type combination together with a given sequence separation, a set of backbone position pairs representing the relative geometry space have been selected to minimize the total RMSD of approximating the RGEOM of actual position pairs in training proteins by their respective nearest neighbors in the representative set. Briefly, an initial large set of backbone position pairs with non-redundant relative geometries (mutual RMSD larger than 0.15, 0.2, 0.25, 0.3 and 0.5 Å for sequence separations of 1,2,3,4 and non-local pairs, respectively) have been extracted from the training proteins. Then a subset of a given number of pairs (see Table S1 for the actual numbers) have been selected to represent the entire set with minimum RMSD. The selection has been carried out using a Monte Carlo (MC) procedure, in which members in the current tentative subset are randomly substituted by members not belonging to the subset. The associated change in RMSD is employed to determine whether to accept or reject the substitution. The MC steps have been repeated until the best results could not be improved further in a large number of MC steps. The subset that lead to the smallest RMSD are accepted as the final result. Table S1 gives the averaged RMSD of approximating the actual position pairs by their respective nearest neighbors in the final representative sets. In Table S1, the numbers of the representative site pairs for different SS type combinations at different sequence separations have been chosen to balance between accuracy and efficiency.

**(3) The interpolation scheme to compute the backbone dependent rotamer energy**

The *φ-ψ* plane is represented using 200 points (Supplementary Figure S1). For a point at, refers to the conditional probability of observing rotamer of residue type *a*. This probability has been estimated as

, (

S7)

in which the index for training residues, the number of rotamers, and the number of training residues, all for residue type *a*. The weight is given by

(

S8)

in which is the RMSD of sidechain atom positions between and . The values for the parameter are 0.1 Å for ILE, VAL, SER and PRO, 0.15 Å for LEU and THR, 0.2 Å for CYS, ASP, PHE, HIS, ASN, TRP and TYR, 0.3 Å for GLU and GLN, 0.4 Å for LYS and MET, and 0.5 Å for ARG. The is an adaptive parameter, its value for a given residue type *a* and a representative point *r* being 5 degree initially, and gradually increased until, among the romaters of *a*, the largest value of the numerator in formula (S7) exceeds 100.

On the basis of the conditional probabilities estimated using formula (S7), the rotamer energy is defined as

. (

S9)

By this definition, the rotamer energy is always nonnegative, with the lowest rotamer energy for any given residue type being zero. Again, for an actual backbone site *p*, the corresponding rotamer energy is obtained by interpolation using the pre-calculated energies at nearby representative *φ-ψ* points.

**(4) Statistical analyses of native protein structures to derive the atomic packing energy terms**

To define an appropriate functional form and determining the parameters, we first divide the inter-residue atomic contacts (distance below 7 Å) into direct ones and indirect ones (Supplementary Figure S3). The inter-atomic radial distribution functions were determined from the direct contacting distances for different types of atom pairs. The resulting distributions are bell-shaped or unimodal, the peak positions considered to be the optimum packing distances between atom pairs of corresponding types. From these pair-wise distances, atom type-specific half packing distances for individual atom types were derived so that the sum of the atomic half packing distances (noted as for an atom *i* of a given type) can reproduce the pairwise optimum distances with the smallest RMS errors. This process has been repeated using 5 different independent sets of training proteins, with the standard deviations of the estimated atomic half packing distances being relatively small and ranging from 0.004 to 0.044 Å (Table S2).

In analyzing the optimum packing distances, it has been observed that for atoms contained in an aromatic ring, contacting atoms located approximately in the ring plane can approach them at notably larger distances (by c.a. 0.25~0.3 Å) than contacting atoms located out of the ring plane. Thus, the direct contacting distances involving aromatic atoms have been separated into approximately horizontal (in-plane) and approximately vertical (out-of-plane) cases to obtain two sets of optimum of packing distances, namely, and (Table S3). For an aromatic ring atom *i* packing against another atom *j*, a tilt angle *θ* is calculated as the angle (ranged between 0 to 90o) between the vector connecting the two atoms and the normal vector of the aromatic plane, and the actual is determined as

**.** (S

10)

Besides the packing involving aromatic atoms, a different set of optimum packing distances have been defined for hydrogen bond pairs (Table S4) on the basis of the hydrogen bond distance distributions in the training proteins.

The actual depth of the attractive well for the packing interaction between atoms *i* and *j* is , in which to compensate for the somewhat overestimated attractive packing involving larger sidechains relative to smaller ones, the empirical factor was introduced to downscale the attractive interactions if any of the two interacting atoms has more than one covalently bonded neighboring atoms, namely,

(S

11)

with

(S

12)

in which and are the numbers of atoms covalently bonded to atoms *i* and *j*, respectively. The remaining parameters and , which respectively determine the widths of the well on the shorter distance side and on the longer distance side, depend on through the following respective formulae,

, and . (S

13)

The values of the length scaling parameters and (Table S5) have been initially chosen by fitting the repacking energy to the negative logarithm of the aforementioned unimodal radial distributions, and then refined based on the results of the sidechain repacking tests (see below).

When using packing energies treated as sum over atom pairs to carry out sequence design, we found that residue types of larger sidechains (i.e., those containing more number of sidechain atoms) tend to be over-favored at surface positions and to a lesser extent at intermediately exposed positions. To compensate for this effect, the solvent accessibility-dependent weighting factor has been introduced in formula (5) of the main text. It takes the following form,

,

(S14)

in which is the minimum weight for pairs of surface residues, and the maximum weigh for pairs of core residues. Their values and those of and are given in Table S5. These values have been refined according to results of the sequence design computational experiments on training backbones.

When ABACUS2 is used for sequence design, no disulfide bond will be introduced. When the accuracy of the method was tested by considering repacking of sidechains in the native proteins that contain disulfide bonds, the following empirical term is considered between the two cysteine sidechains that form a disulfide bond,

(S15)

in which *d* is the distance (in Å) between the two sulfur atoms, and are the respective bond angles and torsional angles (in degrees) involved in the disulfide bond.

**(5) The determination of *SAI* using a refined pseudo sidechain model**

In ABACUS2, a re-refined pseudo sidechain type is used. The *SAI* for a backbone position has been determined by considering a fixed set of surface points that are evenly spread over the potentially solvent accessible part of the corresponding pseudo sidechain’s atomic surfaces. The fraction of points that remain solvent-exposed in the presence of the backbone and the pseudo sidechain atoms at all other positions is determined. The results are used to rank all backbone positions in the training protein structures, in the order of ascending solvent accessibility. The ranks are then rescaled to values between 0 and 1, to obtain a function that maps the fraction of solvent exposed pseudo sidechain surface points to the *SAI* value.

The pseudo sidechain model used in ABACUS2 has been empirically defined and heuristically optimized on the basis of the mutual information between the residue type and calculated *SAI* in native proteins.[2] Briefly, 5 candidate pseudo sidechain models with varied numbers of sidechain atoms and sidechain geometries have been considered (see PSA to PSE in supplementary Figure S4). Within each candidate model, parameters including internal geometries, atomic radius, and the parts of atomic surfaces for solvent accessibility calculations have been varied to find parameter combinations that yield the largest mutual information between the calculated *SAI* and the residue type in native proteins. The model that leads to the most informative *SAI* values is PSD (see supplementary Figure S4, with the surface points used for *SAI* calculations shown in supplementary Figure S5), which has been employed in ABACUS2.

**1.2. Protein sets for training and testing**

Four sets of training protein structures have been considered in deriving the ABACUS2 model. The first training set (TRN21031) contains 21031 protein chains with resolution higher than 2.5 Å and mutual sequence identity lower than 50%.[3] This training set has been used to train the single residue and the residue pairwise statistical energy terms. The second training set (TRN7258) contains 7258 protein chains with resolution higher than 1.8 Å and mutual sequence identity below 30%.[3] This training set has been used to derive the pseudo sidechain model, the rotamer library, and the rotamer energy and atomic packing radial distributions. The third training set (TRN200) contains 200 protein chains and is a subset of TRN7258. It has been used in the sidechain repacking computational experiments to optimize parameters in the packing energy functions. The fourth training set (TRN40) contains 40 proteins of 76 to 200 residues in length. They have been used to evaluate the errors in the interpolation-based statistical energies in comparison with the directly estimated statistical energies. They have also been used in the amino acid sequence redesign computational experiments to optimize the weights of the packing energies and the secondary structure type-dependent reference energies.

The set of test proteins (TST40) contains 40 protein chains of lengths from 100 to 200. Unless specified otherwise, the test results shown in Results and Discussions have been obtained on TST40. A complete list of PDB IDs in TRN200, TRN40 and TST40 are given in supplementary Table S6.

**1.3. The computational protocol for sidechain optimization**

Given the backbone structure and an amino acid sequence, the optimum rotamer sequence has been determined using Mont Carlo simulated annealing. Each MC run starts from a random initial rotamer sequence and a temperature of 10.0 (in the energy unit, and the same below). Then, for a target sequence of length *L*, the MC temperature is multiplied by 0.9 after every 1000*L* MC steps until the temperature is lower than 0.01. The rotamer sequence generated by MC simulated annealing is further optimized by two rounds of quenching. In the first round, pairs of interacting backbone positions are scanned. For each pair, all allowed rotamer combinations are considered and the rotamer sequence is updated with the combination that lead to the lowest total energy. In the second round, individual backbone positions are scanned. For each position, the rotamer that leads to the lowest total energy is selected. In the above protocol, the lengths and parameters of the MC simulated annealing have been empirically chosen so that runs starting from different initial rotamer sequences can converge to highly similar results. The final quenching rounds lead to only small adjustments of the result and do not need to be repeated. Started from different initial rotamer sequences, the *RMSD* between the final structures are within 0.4 Å and the energy differences within 0.5 unit.

**1.4. The computational protocol for amino acid sequence design**

For a given backbone target, amino acid sequences are designed using the following steps to minimize the ABACUS2 total energy. In the first step, an allowed set of rotamer states is determined for each backbone position. Excluded from these sets are rotamers that sterically clash with the fixed backbone or that have backbone-dependent rotamer energies above 4.0 unit. In the second step, the single position energies are computed for each allowed rotamer at each position and the pairwise energies for each allowed rotamer pair at each position pair. The calculated values are stored as look-up tables. In the third step, the rotamer sequence is optimized by Monte Carlo simulated annealing to lower its total energy, which comprises sums of entries from the pre-computed energy tables. Each MC run starts from a random initial rotamer sequence and an initial temperature of 100.0. In each MC step, the rotamer sequence is changed by random single-site substitution and the change is either accepted or rejected on the basis of the energy change and by the Metropolis criterion. The temperature is multiplied by 0.95 after every 1000 MC steps (*L* is the length of the target chain) until the temperature is lower than 0.001. With this protocol, the averaged overall identity of the sequences generated by different MC runs are above 85%. As those levels of sequence identity are much higher than those observed for native proteins that fold into highly similar structures, we will treat these sequences as equally good designs for a given target. For each target in TST40, 5 sequences have been designed, the results are averaged over these sequences.

**1.5. Expression and characterization of proteins designed using ABACUS2**

For each backbone target, three genes encoding designed sequences (see supplementary Table S8) have been synthesized. We would like to note that after the sequences for experimental test had been designed, there was a small adjustment of the ABACUS2 code, which involved using special optimum packing distances to treat the inter-atomic packing between hydrogen bond donor and acceptor atoms. Thus the version used to design the sequences in Table S8 slightly differed from the final version described here. As the sequences in Table S8 are highly similar (sequence identity >75%) to those designed using the final version, we did not try to repeat experiments on sequences designed using the final version.

The synthesized DNA sequences were cloned into a pET-22b(+) vector by using the NdeI and XhoI sites. Proteins were expressed in *Escherichia coli* BL21(DE3) cells with 1mM IPTG induction. Cells were cultured at 16℃ for 20 h and proteins purified using a Ni2+-NTA resin followed by gel filtration in a Superdex 75 column with the ÄKTA purifier system. Uniformly 15N-labeled proteins were prepared by growing the bacteria in SV40 medium using 15NH4Cl (0.5 g/L) as stable isotope sources. NMR experiments have been performed at 298 K on a Bruker DMX600 spectrometer equipped with triple resonances, self-shielded z axis gradient probes. Data were processed using the programs NMRDraw/NMRPipe. Spectra were analyzed using the program SPARKY 3. The circular dichroism spectra of purified proteins have been obtained on a Jasco J-810 spectrometer using a 0.5 mm path length quartz cuvette at room temperature. Protein samples were prepared in 10 mM KH2PO4, 100 mM KCl buffer and protein concentrations were adjust to 0.3mg/ml before measurement. An increasing temperature was applied to the protein samples for denaturation studies. The thermal denaturation CD data were obtained by measuring ellipticity every 5 °C of temperature increasing from 20 °C to 95 °C at λ=218 nm.

**2. Supplementary Tables**

**Table S1.** Discretization scheme of pair geometry

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *SS*b | *SEP*c*=1* | | *SEP=2* | | *SEP=3* | | *SEP=4* | | *SEP>4* | |
| *N*d | *RMSD* | *N* | *RMSD* | *N* | *RMSD* | *N* | *RMSD* | *N* | *RMSD* |
| HH | 100 | 0.043 | 100 | 0.072 | 200 | 0.082 | 100 | 0.108 | 5000 | 0.328 |
| HE | 100 | 0.058 | 100 | 0.149 | 100 | 0.21 | 200 | 0.222 | 5000 | 0.312 |
| HC | 200 | 0.049 | 400 | 0.108 | 400 | 0.138 | 800 | 0.145 | 6000 | 0.354 |
| EH | 100 | 0.082 | 100 | 0.196 | 100 | 0.268 | 200 | 0.271 | 5000 | 0.313 |
| EE | 100 | 0.075 | 400 | 0.122 | 200 | 0.248 | 200 | 0.29 | 3000 | 0.224 |
| EC | 200 | 0.081 | 600 | 0.155 | 600 | 0.226 | 800 | 0.245 | 5000 | 0.327 |
| CH | 200 | 0.077 | 400 | 0.15 | 400 | 0.199 | 800 | 0.217 | 6000 | 0.355 |
| CE | 200 | 0.08 | 600 | 0.151 | 600 | 0.223 | 800 | 0.266 | 5000 | 0.338 |
| CC | 200 | 0.082 | 800 | 0.159 | 1000 | 0.215 | 1500 | 0.249 | 8000 | 0.360 |

aThe RMSD are calculated over the backbone atoms and the pseudo Cγ atoms of the two positions, and averaged over backbone position pairs contained in the TRN21031 set.

bSecondary structure types of the two backbone positions. H stands for helix, E for strand, and C for coil.

cThe sequence separation between the two backbone positions.

dThe number of points used to represent the relative geometry.

**Table S2.** The half optimum packing distances (in Å) of different atom types.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atom** |  | **SDa** | **Atom** |  | **SD** | **Atom** |  | **SD** |
| N | 1.764 | 0.004 | LYS- Cα | 1.823 | 0.009 | GLN- Cγ | 1.855 | 0.006 |
| PRO-N | 1.707 | 0.018 | LYS- Cβ | 1.909 | 0.013 | GLN- Cδ | 1.688 | 0.034 |
| C | 1.647 | 0.004 | LYS- Cγ | 1.863 | 0.010 | GLN-Oε1 | 1.653 | 0.006 |
| O | 1.648 | 0.003 | LYS- Cδ | 1.881 | 0.008 | GLN-Nε2 | 1.699 | 0.011 |
| ALA-Cα | 1.825 | 0.009 | LYS-Cε | 1.907 | 0.015 | ARG- Cα | 1.854 | 0.020 |
| ALA-Cβ | 1.868 | 0.005 | LYS-Nζ | 1.921 | 0.025 | ARG- Cβ | 1.855 | 0.021 |
| CYS- Cα | 1.843 | 0.022 | LEU- Cα | 1.820 | 0.008 | ARG- Cγ | 1.844 | 0.011 |
| CYS- Cβ | 1.843 | 0.017 | LEU- Cβ | 1.917 | 0.007 | ARG- Cδ | 1.827 | 0.015 |
| CYS-Sγ | 1.902 | 0.007 | LEU- Cγ | 1.877 | 0.005 | ARG-Nε | 1.665 | 0.021 |
| ASP- Cα | 1.766 | 0.010 | LEU- Cδ1 | 1.931 | 0.003 | ARG-Cζ | 1.627 | 0.014 |
| ASP- Cβ | 1.848 | 0.005 | LEU- Cδ2 | 1.931 | 0.003 | ARG-Nη | 1.706 | 0.012 |
| ASP-Cγ | 1.664 | 0.024 | MET- Cα | 1.783 | 0.030 | SER- Cα | 1.799 | 0.018 |
| ASP-Oδ | 1.672 | 0.011 | MET- Cβ | 1.854 | 0.013 | SER- Cβ | 1.847 | 0.007 |
| GLU- Cα | 1.831 | 0.010 | MET- Cγ | 1.842 | 0.009 | SER-Oγ | 1.651 | 0.011 |
| GLU- Cβ | 1.835 | 0.007 | MET- Sδ | 1.904 | 0.009 | THR- Cα | 1.827 | 0.042 |
| GLU-Cγ | 1.857 | 0.005 | MET-Cε | 1.882 | 0.009 | THR- Cβ | 1.853 | 0.011 |
| GLU-Cδ | 1.688 | 0.013 | ASN- Cα | 1.749 | 0.021 | THR-Oγ1 | 1.649 | 0.008 |
| GLU-Oε | 1.707 | 0.017 | ASN- Cβ | 1.824 | 0.025 | THR-Cγ2 | 1.894 | 0.009 |
| PHE- Cα | 1.823 | 0.014 | ASN- Cγ | 1.658 | 0.012 | VAL- Cα | 1.933 | 0.035 |
| PHE- Cβ | 1.814 | 0.009 | ASN- Oδ1 | 1.620 | 0.013 | VAL- Cβ | 1.872 | 0.008 |
| GLY- Cα | 1.804 | 0.007 | ASN- Nδ2 | 1.707 | 0.035 | VAL- Cγ1 | 1.925 | 0.004 |
| HIS- Cα | 1.771 | 0.013 | PRO- Cα | 1.876 | 0.007 | VAL- Cγ2 | 1.917 | 0.002 |
| HIS- Cβ | 1.819 | 0.015 | PRO- Cβ | 1.876 | 0.005 | TRP- Cα | 1.722 | 0.030 |
| ILE- Cα | 1.954 | 0.017 | PRO- Cγ | 1.887 | 0.009 | TRP- Cβ | 1.806 | 0.011 |
| ILE- Cβ | 1.912 | 0.011 | PRO- Cδ | 1.869 | 0.009 | TYR- Cα | 1.814 | 0.022 |
| ILE- Cγ1 | 1.900 | 0.003 | GLN- Cα | 1.815 | 0.044 | TYR- Cβ | 1.818 | 0.017 |
| ILE- Cδ1 | 1.915 | 0.002 | GLN- Cβ | 1.827 | 0.008 | TYR-Oη | 1.611 | 0.006 |
| ILE- Cγ2 | 1.912 | 0.003 |  |  |  |  |  |  |

aThe standard deviations have been estimated by partitioning the training data in TRN7258 into 5 sets, estimating the half distances of packing using each sets, and then determining the standard deviations of the estimated values.

**Table S3**. The half distances (in Å) for the optimum in-plane () and out-of-plane () packing of aromatic atoms.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atom** |  |  | **Atom** |  |  | **Atom** |  |  |
| **PHE-Cγ** | 1.715 | 1.980 | HIS-Nε2 | 1.590 | 1.900 | TRP-Cε3 | 1.700 | 1.977 |
| **PHE-Cδ** | 1.715 | 1.980 | TRP-Cγ | 1.650 | 1.900 | TRP-Cη2 | 1.700 | 1.977 |
| **PHE-Cε** | 1.715 | 1.980 | TRP-Cδ1 | 1.635 | 1.972 | TYR- Cγ | 1.670 | 1.980 |
| **PHE-Cζ** | 1.715 | 1.980 | TRP-Cδ2 | 1.580 | 1.950 | TYR- Cδ | 1.670 | 1.980 |
| **HIS-Cγ** | 1.610 | 1.905 | TRP-Nε1 | 1.575 | 1.940 | TYR- Cε | 1.670 | 1.980 |
| **HIS-Nδ1** | 1.590 | 1.900 | TRP-Cε2 | 1.580 | 1.950 | TYR-Cζ | 1.670 | 1.980 |
| **HIS-Cδ2** | 1.610 | 1.905 | TRP-Cε3 | 1.665 | 1.948 |  |  |  |
| **HIS-Cε1** | 1.610 | 1.905 | TRP-Cζ1 | 1.650 | 1.977 |  |  |  |

**Table S4.** Optimum packing distances (in Å) for hydrogen-bonding heavy atom pairs.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Donor atom | Acceptor atom | | | | | |
| O | ASP-Oδ  GLU-Oε | ASN- Oδ1  GLN-Oε1 | THR-Oγ1  SER-Oγ | TYR-Oη | HIS-Nδ1  HIS-Nε2 |
| HIS-Nδ1, HIS-Nε2 | 2.93 | 2.89 | 2.9 | 3.05 | 2.98 | 3 |
| ARG-Nη | 2.81 | 2.74 | 2.8 | 2.75 | 2.71 | 3.09 |
| ARG-Nε | 2.9 | 2.88 | 2.89 | 2.94 | 2.95 | 3.04 |
| LYS-Nζ | 2.86 | 2.85 | 2.84 | 2.9 | 3.08 | 2.98 |
| TRP-Nε1 | 2.82 | 2.78 | 2.81 | 2.86 | 2.94 | 2.89 |
| THR-Oγ1 | 2.89 | 2.86 | 2.88 | 2.91 | 2.93 | 2.99 |
| SER-Oγ | 2.73 | 2.66 | 2.69 | 2.74 | 2.71 | 2.76 |
| TYR-Oη | 2.67 | 2.61 | 2.66 | 2.71 | 2.71 | 2.71 |
| ASN- Nδ2, GLN-Nε2 | 2.93 | 2.92 | 2.93 | 2.95 | 2.98 | 3.02 |

**Table S5.** Parameters for calculating the atomic packing energies.

|  |  |
| --- | --- |
| Parameter | Value |
| in formula (21) | 0.095 (0.1 for hydrogen bonding atomic pairs) |
| in formula (21) | 0.28 (0.15 for hydrogen bonding atomic pairs) |
| in formula (18) | 1.3, if both atoms are nonpolar; |
| 0.7, if only one atom is polar; |
| 0, if both atoms are polar and cannot form hydrogen bond; |
| 1.5, if the two atoms can form hydrogen bond; |
| 1.4, if one atom is aromatic, while the other is non-polar. |
| in formula (22) | 0.05 |
| in formula (22) | 0.75 |
| in formula (22) | 0.65 |
| *σ* in formula (22) | 0.07 |

**Table S6.** PDB IDs of proteins in the training and test sets.

|  |
| --- |
| **The TRN200 set:** 1b0b, 1b0u, 1c0p, 1d02, 1d0d, 1d0q, 1e0c, 1f0l, 1i0r, 1j0p, 1k0i, 1m0k, 1m0w, 1n08, 1n0w, 1n0x, 1o06, 1p0h, 1p0z, 1q0p, 1q0r, 1r0m, 1r0u, 1s0a, 1t07, 1t0b, 1t0p, 1t0t, 1u07, 1u0k, 1v05, 1v0w, 1w0h, 1w0n, 1w0p, 1x0t, 1y02, 1y07, 1y0b, 1y0h, 1y0k, 1y0u, 1z0j, 1z0n, 1z0p, 1z0s, 1z0w, 2a0b, 2b06, 2b0a, 2b0t, 2b0v, 2c0a, 2c0c, 2e0t, 2f01, 2f0c, 2g0w, 2i02, 2i0o, 2j05, 2j0a, 2o0a, 2o0j, 2o0m, 2o0q, 2p09, 2p0b, 2p0n, 2p0s, 2q03, 2q0s, 2q0t, 2q0y, 2r01, 2r0c, 2r0x, 2r0y, 2v03, 2v05, 2v0p, 2w0i, 2y0o, 2z08, 2z0j, 2z0q, 2z0t, 2z0x, 3a02, 3a07, 3a09, 3a0s, 3a0y, 3b0f, 3b0g, 3b0p, 3b0t, 3b0x, 3c0f, 3d01, 3d02, 3d06, 3d0f, 3d0j, 3e03, 3e05, 3e0e, 3e0x, 3e0z, 3f0d, 3f0h, 3f0p, 3g02, 3g0k, 3g0m, 3g0o, 3g0t, 3h05, 3h09, 3h0n, 3h0o, 3h0u, 3i09, 3i0w, 3i0z, 3k01, 3k05, 3k06, 3k0b, 3l00, 3l0f, 3l0q, 3m0f, 3m0m, 3m0z, 3n01, 3n08, 3n0r, 3n0u, 3n0x, 3o0d, 3o0q, 3o0y, 3p02, 3p0b, 3p0f, 3p0k, 3p0y, 3q0h, 3r0n, 3r0v, 3s0a, 3t0o, 3v0d, 3v0s, 3w06, 3w07, 3w0e, 3w0k, 3w0o, 3w0t, 3x0f, 3x0i, 3x0t, 3x0u, 4a02, 4a0d, 4b0h, 4b0m, 4b0z, 4c08, 4c0n, 4d05, 4d0p, 4d0q, 4f01, 4f03, 4f06, 4f0j, 4f0r, 4f0w, 4f0z, 4g0x, 4h08, 4h0c, 4i0n, 4i0w, 4j0d, 4j0e, 4j0w, 4k0n, 4l05, 4l07, 4l0c, 4l0n, 4m0n, 4m0w, 4n01, 4n02, 4n03 |
| **The TRN40 set:** 1q1f, 1l3p, 2e1f, 2pvb, 2a0b, 1cy5, 1k04, 1r7j, 1v2z, 1x91, 1x8q, 1i8a, 2nvh, 1txl, 1w0n, 1v05, 2q3w, 1h4a, 1v70, 1v2x, 2a4v, 1m7b, 2a1i, 1j24, 2fwh, 2d59, 1r26, 1dz3, 1qzm, 1z2u, 1nwz, 1twu, 1i0v, 1r29, 1q2y, 1a1x, 1ubq, 1g2r, 1ew4, 1f7l |
| **The TST40 set:** 1f4p, 1l1q, 1mba, 1naz, 1p90, 2a15, 2fr2, 2g40, 2g64, 2h8e, 2j8h, 2o0q, 2o7a, 2w1r, 2wtg, 2xdh, 2xki, 2xov, 2xy1, 2z5w, 3bt5, 3h79, 3m1x, 3p2h, 3p9n, 3q4o, 3r87, 3zja, 4bja, 4gco, 4l8a, 4o6u, 4p82, 4ums, 4ymy, 5e16, 5fui, 5hdw, 5i9p, 5j1z |

**Table S7.** The secondary structure type-dependent residue type-specific reference energies ( in formula (6) of the main text.

|  |  |  |  |
| --- | --- | --- | --- |
| Residue type | Secondary structure-dependent reference energy | | |
| helix | strand | coil |
| ALA | 0.1 | 0 | 0.1 |
| CYS | 0 | 0 | 0 |
| ASP | -0.2 | -0.3 | 0 |
| GLU | 0.1 | 0 | 0 |
| PHE | -0.2 | -0.2 | -0.2 |
| GLY | 0 | 0 | 0 |
| HIS | -0.5 | -0.5 | -0.5 |
| ILE | 0.5 | 0.3 | 0.3 |
| LYS | 0.1 | 0 | 0 |
| LEU | 0.6 | 0.5 | 0.25 |
| MET | 0 | 0 | 0 |
| ASN | -0.3 | -0.3 | -0.1 |
| PRO | 0.5 | 0.5 | 0.5 |
| GLN | -0.3 | -0.3 | -0.3 |
| ARG | -0.2 | 0 | 0 |
| SER | -0.3 | -0.2 | -0.2 |
| THR | -0.2 | 0.2 | 0 |
| VAL | 0.2 | 0.2 | 0.2 |
| TRP | 0 | 0 | 0 |
| TYR | 0 | 0 | 0 |

**Table S8.** Proportion of residues with χ1 correctly predicted within 40°.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ABACUS1 | **ABACUS2** | SCWRL4 | Rosetta default | Rosetta ex1 | Rosetta ex2 | Rosetta ex3 | Rosetta ex4 |
| CYS | 0.883 | **0.951** | 0.934 | 0.934 | 0.934 | 0.951 | 0.918 | 0.934 |
| ASP | 0.788 | **0.872** | 0.84 | 0.817 | 0.82 | 0.817 | 0.82 | 0.826 |
| GLU | 0.751 | **0.784** | 0.756 | 0.758 | 0.748 | 0.761 | 0.748 | 0.743 |
| PHE | 0.957 | **0.987** | 0.97 | 0.957 | 0.961 | 0.957 | 0.961 | 0.961 |
| HIS | 0.797 | **0.862** | 0.862 | 0.797 | 0.826 | 0.826 | 0.819 | 0.812 |
| ILE | 0.943 | **0.973** | 0.964 | 0.949 | 0.955 | 0.958 | 0.958 | 0.958 |
| LYS | 0.776 | **0.761** | 0.767 | 0.767 | 0.752 | 0.77 | 0.749 | 0.758 |
| LEU | 0.912 | **0.944** | 0.92 | 0.934 | 0.944 | 0.949 | 0.949 | 0.946 |
| MET | 0.748 | **0.871** | 0.835 | 0.799 | 0.813 | 0.835 | 0.842 | 0.835 |
| ASN | 0.796 | **0.832** | 0.845 | 0.832 | 0.841 | 0.841 | 0.841 | 0.85 |
| PRO | 0.851 | **0.851** | 0.882 | 0.895 | 0.899 | 0.895 | 0.908 | 0.904 |
| GLN | 0.796 | **0.801** | 0.768 | 0.81 | 0.82 | 0.815 | 0.81 | 0.796 |
| ARG | 0.8 | **0.844** | 0.803 | 0.797 | 0.783 | 0.797 | 0.8 | 0.776 |
| SER | 0.664 | **0.738** | 0.689 | 0.697 | 0.711 | 0.711 | 0.702 | 0.702 |
| THR | 0.89 | **0.922** | 0.88 | 0.89 | 0.899 | 0.906 | 0.906 | 0.903 |
| VAL | 0.916 | **0.916** | 0.919 | 0.892 | 0.905 | 0.914 | 0.914 | 0.916 |
| TRP | 0.913 | **0.95** | 0.95 | 0.875 | 0.9 | 0.888 | 0.9 | 0.9 |
| TYR | 0.938 | **0.953** | 0.969 | 0.943 | 0.953 | 0.953 | 0.953 | 0.958 |
| all | 0.838 | **0.872** | 0.856 | 0.849 | 0.854 | 0.859 | 0.857 | 0.855 |

The χ1 prediction accuracy has been calculated on the TST40 set of native structures.The rows Rosetta default to Rosetta ex4 correspond to results obtained using rotamer sets of increasing sizes and finer conformation resolutions.

**Table S9.** Proportion of residues with both χ1 and χ2 correctly predicted within 40°.

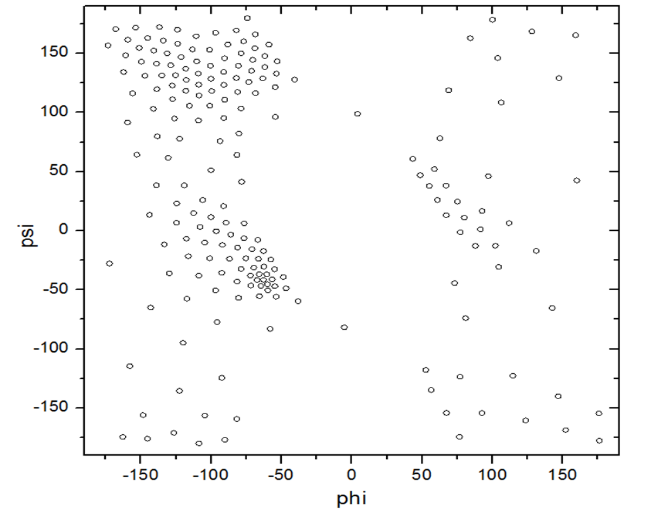
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ABACUS1 | **ABACUS2** | SCWRL4 | Rosetta default | Rosetta ex1 | Rosetta ex2 | Rosetta ex3 | Rosetta ex4 |
| ASP | 0.544 | **0.747** | 0.706 | 0.718 | 0.718 | 0.721 | 0.727 | 0.727 |
| GLU | 0.556 | **0.764** | 0.751 | 0.722 | 0.722 | 0.725 | 0.735 | 0.727 |
| PHE | 0.608 | **0.552** | 0.599 | 0.608 | 0.608 | 0.608 | 0.612 | 0.616 |
| HIS | 0.42 | **0.609** | 0.558 | 0.594 | 0.594 | 0.572 | 0.587 | 0.58 |
| ILE | 0.837 | **0.891** | 0.852 | 0.831 | 0.831 | 0.873 | 0.873 | 0.885 |
| LYS | 0.625 | **0.801** | 0.782 | 0.779 | 0.779 | 0.767 | 0.773 | 0.767 |
| LEU | 0.84 | **0.885** | 0.854 | 0.846 | 0.846 | 0.868 | 0.887 | 0.885 |
| MET | 0.698 | **0.849** | 0.763 | 0.712 | 0.712 | 0.741 | 0.791 | 0.77 |
| ASN | 0.429 | **0.668** | 0.611 | 0.664 | 0.664 | 0.673 | 0.668 | 0.668 |
| PRO | 0.789 | **0.785** | 0.816 | 0.811 | 0.811 | 0.811 | 0.811 | 0.825 |
| GLN | 0.573 | **0.725** | 0.64 | 0.654 | 0.654 | 0.682 | 0.654 | 0.692 |
| ARG | 0.688 | **0.837** | 0.82 | 0.79 | 0.79 | 0.81 | 0.814 | 0.82 |
| TRP | 0.75 | **0.875** | 0.813 | 0.838 | 0.838 | 0.838 | 0.825 | 0.863 |
| TYR | 0.594 | **0.573** | 0.583 | 0.568 | 0.568 | 0.583 | 0.589 | 0.578 |
| all | 0.655 | **0.77** | 0.744 | 0.724 | 0.74 | 0.751 | 0.757 | 0.759 |

Test set is the same as above.

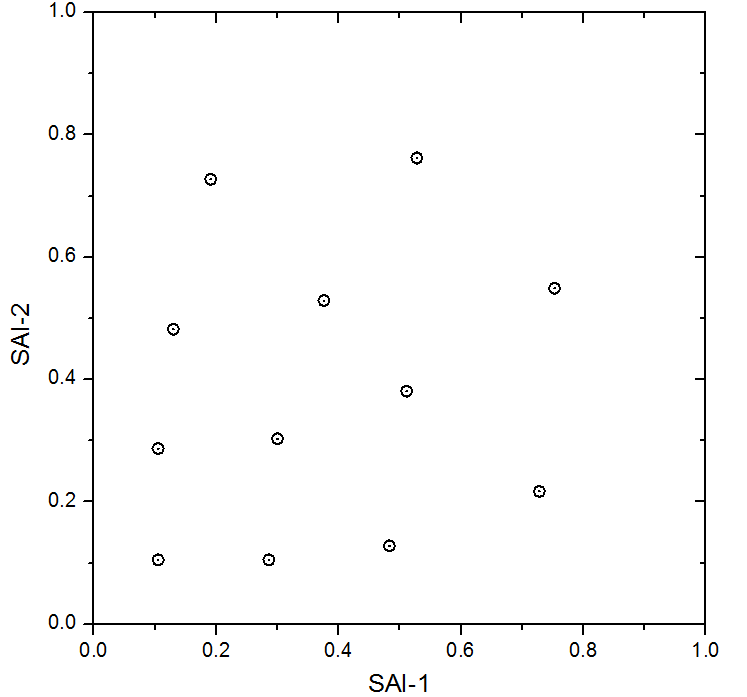
**Table S10.** Experimentally tested amino acid sequences designed with ABACUS2.

|  |  |
| --- | --- |
| 1ubq\_1 | QKIKVKTSDGKTITLTVTPDMTVKEVRELIRKKTGLPPSDLKLIYNGKVLTADMTLSDFNITKGDVLTLELVKDGG |
| 1ubq\_2 | QKIKVKTEDGKVYTLTVTPDMTVKEVRELIRKKTGLPPSDLELEYNGKVLKADMTLSDFNITAGDVLTLRLVKDGG |
| 1ubq\_3 | QKITIKTSDGKEYTLTVTPDMTVKEVRELIRKKTGIPPSDIRLIYNGKVLKADMTLSDFNITAGATITLEIIKDGG |
| 1r26\_1 | PLPPNQEPSWIKLTSVEEMKRLIALNFLVVLVFYAKNDELVEKTKEQLKELAKEYPDILIIWIDVETLPEIVKQFNITSLPAFIIMKNGKLLGKVTGPNVEKLKEILKEILAK |
| 1r26\_2 | PVPPNQEPSWIKLTSVEEMKRLIALDWLVVLVFYAKNVELAEKTKEQLKELAKEYPDVLIIWIDVETLPEIVKQFNITSIPAFILMKNGKLLGKVTGPNVEKLKELIKRYLAE |
| 1r26\_3 | PVPPNQEPSFIKLTSVEEMKRLIALDWLVILVFAAKNNELVEKTKEQLQQLAKEYPDVLLILIDVETLPEIVKKFNITSLPAFIIMKNGKLLGKVTGPNVEKLKEIIKKILAE |
| 2qsb\_1 | EEERQQLFDEIVKLLKELANLSDVPPKLREAARRALELLNDSSMSLEEALREVLELLKRMLNDPEIPPEGREIIRKIIELIRQLL |
| 2qsb\_2 | EEEREEIFKRVVELLKKLANDSDVPPEVRELARQALELLNDSSMSLEEAIREVLELLKKALNDPSVPPEGREIIRKIIALLRELL |
| 2qsb\_3 | EEERQQLFDQVVELLKKLANLSDVPPEVREAARRALELLNDSSMSLEEAIREVLELLKRVLNDPSLPPEGRELVRKIIELLRQLL |

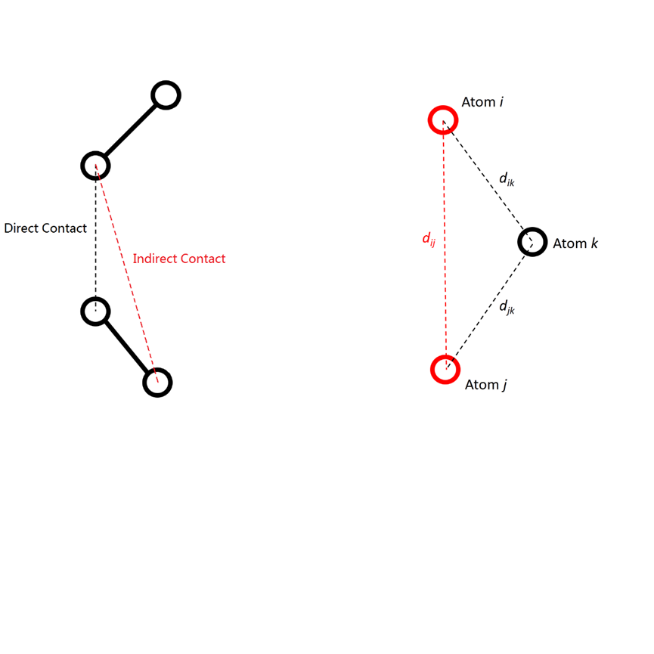
**3. Supplementary Figures**



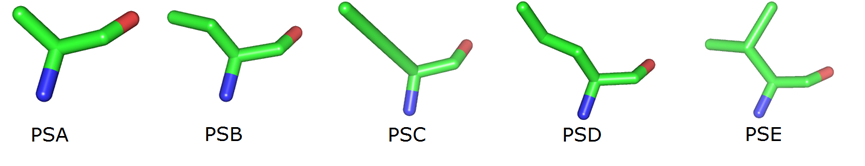
**Figure S1.** The 200 presentative points on the φ-ψ plane.



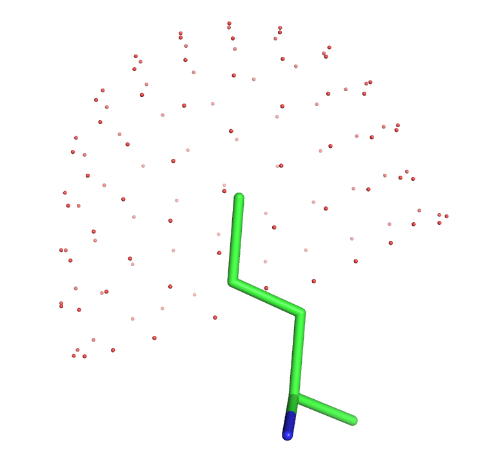
**Figure S2.** The 12 representation points for the SAI values of backbone position pairs.



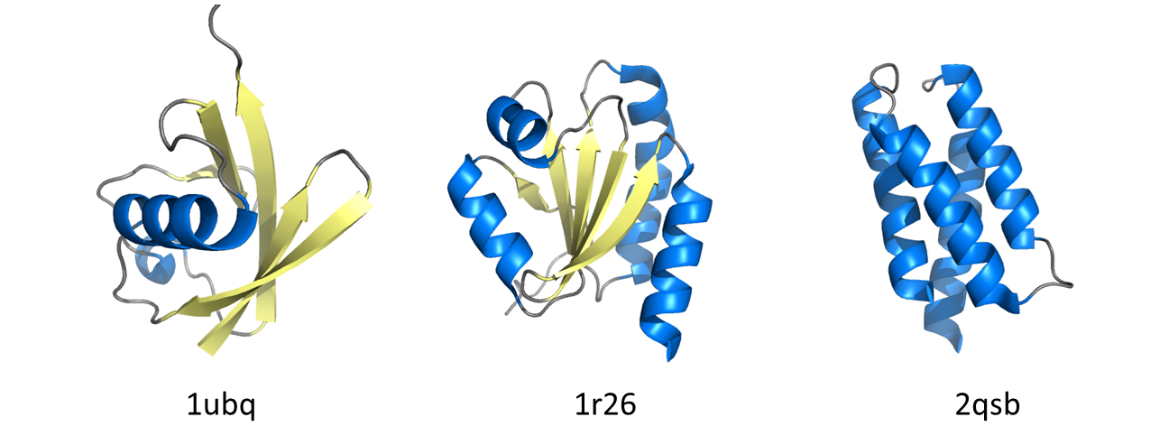
**Figure S3.** The definitions of direct and indirect atomic contacts. The circles represent atoms. The solid lines represent covalent bonds. The dashed lines represent non-bonded contacts between atoms, with those in black defined as direct contacts and those in red defined as in direct contacts. In the left side case, the indirect contact has a longer distance than another direct contact between one contacting atom and another contacting atom’s covalently bonded neighbor. On the right side case, the indirect contact corresponds to the longest edge of a triangle former by three mutually contacting atoms.



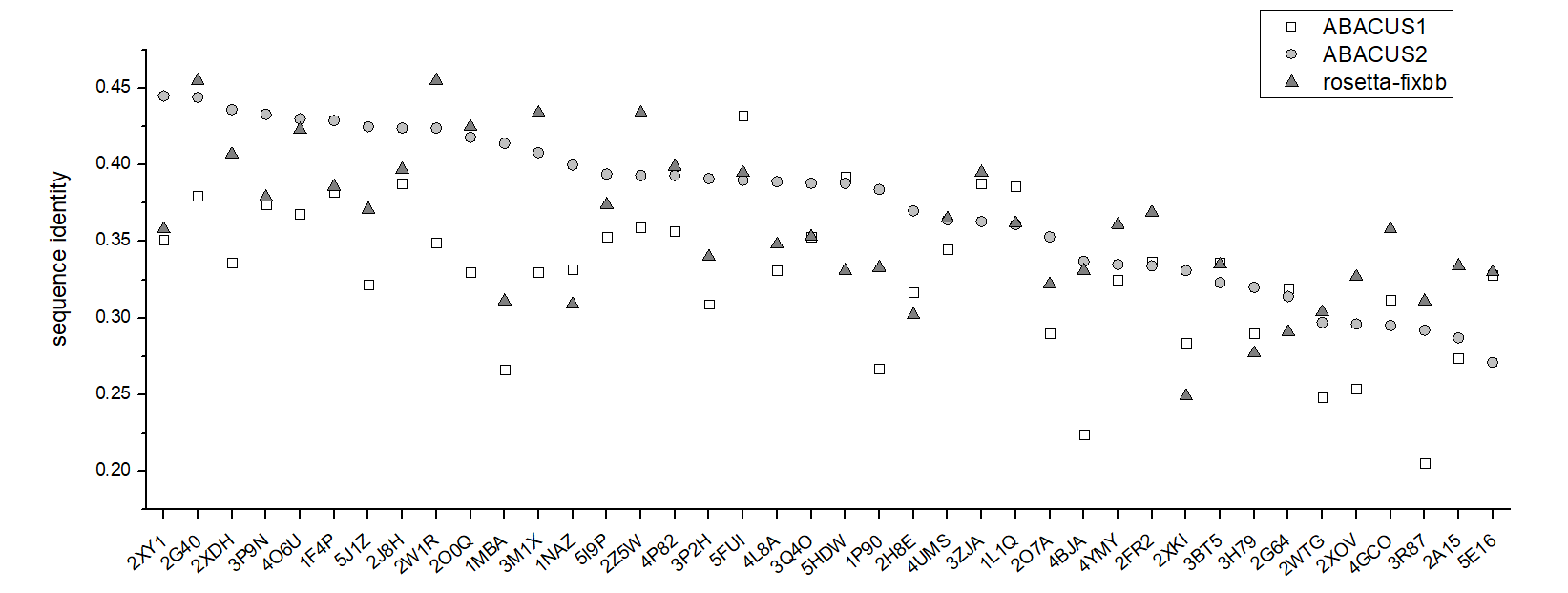
**Figure S4.** The five candidate pseudo sidechain models. The backbone nitrogen atom is shown in blue, oxygen in red. The final model is PSD. In this model, the C-C bond lengths are 1.5 Å, the bond angles are 109.5°,the dihedral angle χ1 is -90°, and χ2 180°.



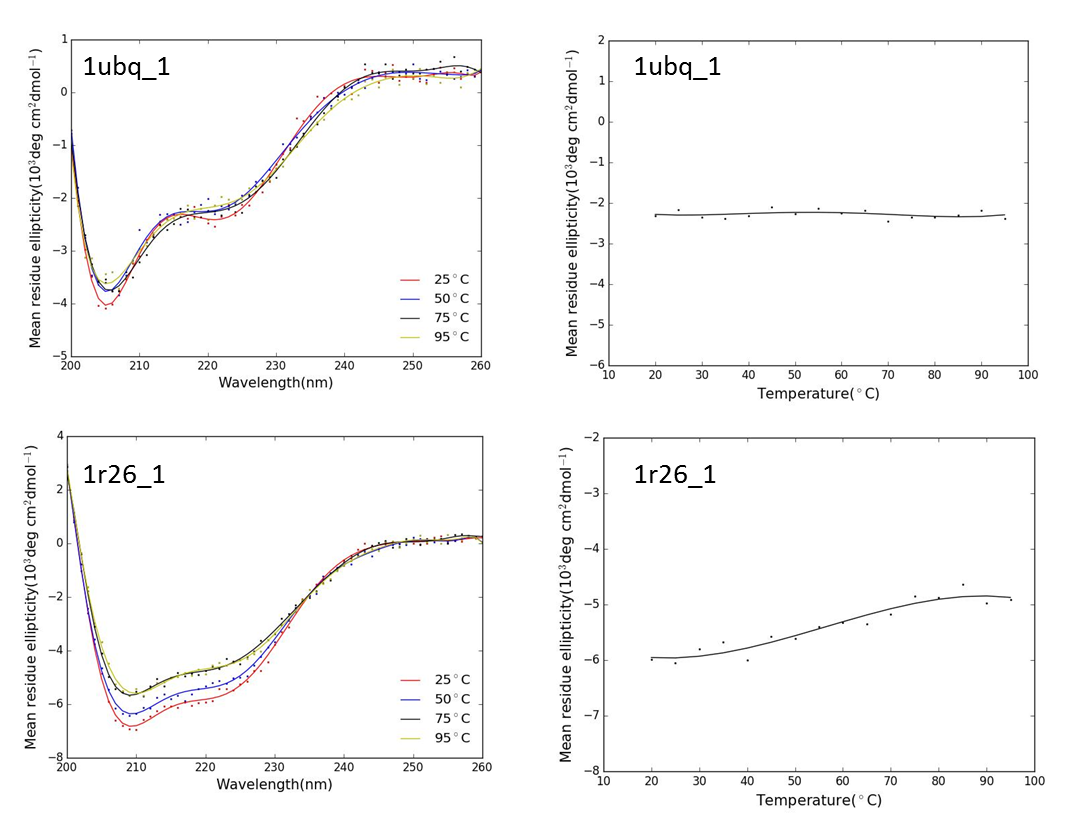
**Figure S5.** The red dots are the set of surface points of pseudo sidechain PSD that are used for SAI calculations.



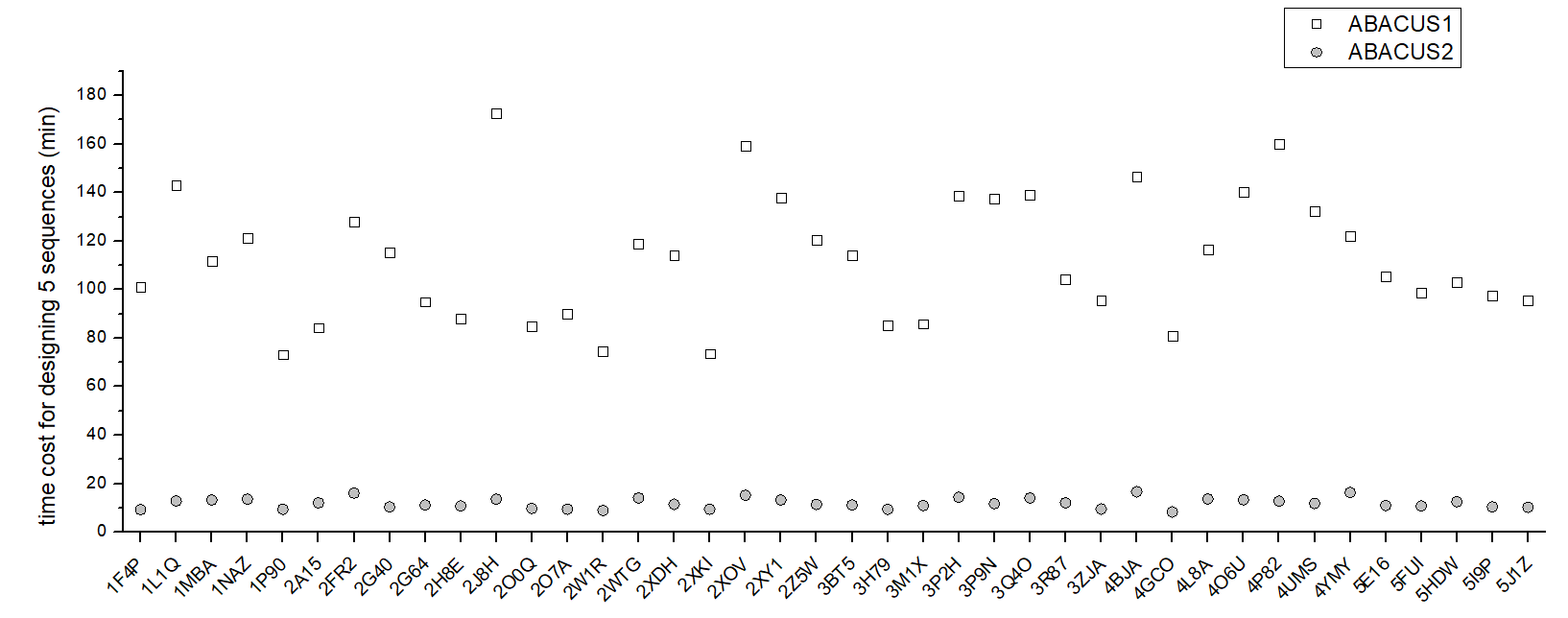
**Figure S6.** The three native backbones for which sequences designed using ABACUS2.



**Figure S7.** Native residue type recovery rate of different proteins in TST40 set



**Figure S8,** Temperature dependent circular dichroism curves for 1ubq\_1 and 1r26\_1. The right panels show signals at λ=218 nm as a function of temperature.

****

**Figure S9.** Comparison of the computational time costs (wall clock times) of ABACUS2 and ABACUS1 for the same sequence design tasks on the same computer system.

**References**

[1] Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins, 1995, 23(4): 566-79

[2] Karchin R, Cline M, Mandel-Gutfreund Y, et al. Hidden Markov models that use predicted local structure for fold recognition: Alphabets of backbone geometry. Proteins, 2003, 51(4): 504-514

[3] Wang GL, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics, 2003, 19(12): 1589-1591