Supporting Online Material for

Toward High-Resolution de Novo Structure Prediction for Small Proteins

Philip Bradley, Kira M. S. Misura, David Baker*

To whom correspondence should be addressed. E-mail: dabaker@u.washington.edu

Published 16 September 2005, *Science* **309**, 1868 (2005)
DOI: 10.1126/science.1113801

This PDF file includes:

Supplementary Online Materials for:

Towards High-Resolution De Novo Structure Prediction for Small Proteins

Philip Bradley, Kira M. S. Misura, and David Baker

Molecular Representation

In the initial, coarse-grained search of conformational space a low-resolution representation of the protein chain is used in which only backbone heavy atoms are explicitly modeled. Side chains are represented by a single "centroid" whose position depends only on the backbone coordinates. Bond lengths and angles are kept fixed at ideal values; the only degrees of freedom are the backbone torsion angles (phi, psi, and omega). In the following high-resolution stage, all atoms, including hydrogens, are represented explicitly. Bond lengths and angles are kept fixed; the degrees of freedom are the backbone and sidechain torsion angles.

Energy Function

The energy function used in the low-resolution search has been described previously(*1*). In brief, the energy of a protein conformation is a linear combination of terms modeling residue-environment and residue-residue interactions, secondary structure packing, chain density, and excluded volume. The all-atom energy function is composed of a 12-6 Lennard-Jones potential, the Lazaridis-Karplus implicit solvation model(*2*), an orientation-dependent hydrogen bonding term based on quantum chemistry calculations(*3*) and analysis of high-resolution protein structures (*4*), a relatively weak

pair interaction term representing longer range electrostatic interactions between polar atoms and pi-pi and cation pi interactions, a side-chain torsional potential derived from the Dunbrack backbone-dependent rotamer library(*5*), and a backbone torsional potential dependent on secondary structure and amino-acid type (parameters are given for most of the terms in the supporting material to Kuhlman *et al*(*6*)). Study of the radial distribution functions for aliphatic carbon atoms suggested an overcompaction due to Lennard-Jones forces between distant atom pairs not properly compensated by the (missing) interactions with solvent molecules; to compensate for this effect and speed the energy calculations, all terms were linearly ramped to zero between 5.0 and 5.5 Å. To avoid singularities in the Lennard Jones potential at very short distances, linear extrapolation using the slope of the potential at 0.60 times the sum of the radii of the two atoms was used for all distances smaller than this value; even with this damping the repulsive interactions are strong enough that very few atomic clashes are observed in models generated using the potential.

Prediction Protocol

For each protein in the test set we constructed a multiple sequence alignment (MSA) using 2 rounds of PSI-BLAST(*7*) with default parameters. This alignment was filtered to 60% non-redundancy and the 50 sequences most similar to the target were chosen for folding. For each homolog we generated 2000 models using the Rosetta de novo structure prediction protocol(*8, 9*). Beginning with a fully extended chain, conformations are generated by 36,000 Monte Carlo fragment replacement moves in which the torsion angles in a randomly selected 3 or 9 residue window of the protein chain are replaced

with the torsion angles from a fragment of a protein of known structure and similar local sequence. The score of the new conformation is calculated and the move is accepted according to the Metropolis criterion(*10*). Fragments from proteins of known structure homologous to the protein being folded were excluded from the fragment libraries (PSI-BLAST e-value < 0.05 or same SCOP superfamily). The low-resolution models were clustered based on C_α-RMSD(*11*), and all members of clusters with size 5 or greater were accepted. Clustering parameters were chosen to give approximately 500 models (25% of the initial population). The target sequence was threaded onto each model, with insertions or deletions modeled by a fast loop modeling protocol in which the lowest-scoring closed loop is chosen from 100 very short simulations consisting of fragment assembly followed by cyclic coordinate descent (*12*) chain-break closure.

The population of target-sequence models (20,000 – 30,000 structures) was refined in the Rosetta all-atom energy function described above by a streamlined Monte Carlo Minimization(*13*) based refinement protocol consisting of 120 *small* and *shear* moves(*14*). Each move consisted of three operations: a small (2°-3°) random perturbation of 5 or 10 randomly selected backbone torsion angles; one-at-a-time (non-combinatorial) sidechain rotamer optimization using a backbone-dependent rotamer library(*5*); and gradient-based minimization of the potential function in backbone and sidechain torsion space. The weight on the repulsive component of the Lennard-Jones potential is ramped up over the first 60 cycles, starting at $1/50^{th}$ of its full weight. During this portion of the simulation the sidechains are restricted to rotameric conformations; the sidechain torsion angles are only continuously minimized in the last 60 cycles. Every ten cycles a full

combinatorial optimization of the sidechains is performed(*15*). For computational efficiency, the highest-energy fifty percent of the models are eliminated after 60 cycles and again after 90 cycles. The final refined models compose the *round 1* set of models. The sequences of 10 close homologues without insertions or deletions were threaded onto the low-scoring models from round 1 and refined by the same protocol. The 10 percent lowest-scoring models for each of the 10 sequences were mapped back to the target sequence and refined again to produce the *round 2* models. The aim of this second round was to identify conformations that were compatible with the sequence variation present in close sequence homologues. All-atom refinement is the CPU-intensive step – the calculations for each protein took between 100 and 150 CPU days on 2.8 GHz processors. For each protein in the benchmark the native structure was idealized (bond angles and bond lengths were replaced with the "ideal" values used in generation of the de novo models, and compensating changes in the torsion angles are made using quasi-Newton optimization to minimize the distance matrix error) and then refined 50 times using different random number seeds with the same 120-cycle protocol to generate a population of near-native models defining the energy basin close to the native structure. These models are referred to as *refined natives* and shown as blue points in the free energy scatter plots.

In this approach, folding simulations with sequence homologs are used to generate additional sampling diversity in the low-resolution search. We rely on the all-atom force field for selection of final models. This contrasts with previous work(*16*) which used

consensus between low-resolution folding simulations for different sequence homologs as a criterion for model selection.

Construction of the Benchmark Set

We first tested the prediction protocol on six proteins from an in-house benchmark set that were members of large sequence families. These proteins are listed in Table 1 with a five character ID consisting of the Protein Data Bank (PDB) (*17*) code plus the chain identifier. A second set of 10 alpha and alpha-beta domains was chosen from the SCOP (*18*) database. These domains are listed in Table 1 with a six character ID: PDB code plus chain identifier plus domain code. The SCOP set was chosen by restricting to domains that were non-redundant at the superfamily level, between 45 and 85 residues long, lacked disulfide bonds and were less than 50% loop. For domains in this set we constructed an MSA using 2 rounds of PSI-BLAST with default parameters. We sorted the domains by MSA depth and took the top 10.

It should be emphasized that the resulting benchmark consists of proteins with deep alignments, and alignment depth often correlates with secondary structure prediction accuracy and quality of fragment libraries. Thus these proteins may not be representative of randomly selected protein domains.

Figure Legends

Figure S1: Representative free-energy landscapes for four small proteins, illustrating the sampling problem in high-resolution de novo structure prediction. (A) ribosomal protein S6 (PDB code 1ris), (B) acylphosphatase (PDB code 1aps), (C) tenascin (PDB code 1ten), (D) spectrin SH3 domain (PDB code 1bk2). Rosetta all-atom energy (y-axis) is plotted against C_α-RMSD (root mean squared deviation of alpha-carbon coordinates after optimal superposition; x-axis) for models generated by simulations starting from the native structure (blue points) or from an extended chain (black points). The free energy function includes the entropic contribution to the solvation free energy but not the configurational entropy.

Figure S2: Variation in models produced by the Rosetta de novo structure prediction method for different homologous sequences. (A-D) For four homologues of ubiquitin, C_α-RMSD to native (x-axis) is plotted against *cluster radius* (distance to the 25^{th} nearest neighbor in the population of models) for the set of low-resolution Rosetta models built by folding that sequence. Each point corresponds to a single model. Structures with smaller values for cluster radius are in densely sampled regions of the conformational space explored by the corresponding set of models. The homologue in (A) fails to generate low-RMSD models, converging instead on a non-native minimum at 10Å. The homologue in (B) converges very well on the native topology. The homologues in (C) and (D) sample the native topology in addition to non-native minima. (E-F) For 1b72, de novo models (black and green points) generated using a sequence homologue (F) but not

the native sequence (E) sample the native free energy basin defined by the refined natives (blue points).

Figure S3: Medium-resolution structure predictions. Superposition of low-energy models (blue) with experimental structures (red) showing core sidechains. The lowest-energy round 1 model (A) for the Enga protein is topologically correct (3.19Å C_α-RMSD) but does not have native-like sidechain packing or loop conformations. The same is true of the lowest-energy model (B) from round 2 for Yhhp (2.46Å C_α-RMSD). The lowest-energy round 2 model for RNA binding protein A (C) has a C_α-RMSD to native of 2.59Å, while the second-lowest energy model is very close to native (1.13Å C_α-RMSD). Examination of the corresponding panel in Fig. S5 ("1di2a_ round 2") suggests that this model might have been selected had its refinement trajectory sampled further into the energy basin defined by the refined natives (blue points).

Figure S4: Free energy landscapes after round 1 for the 16 test proteins. y axis, Rosetta all atom energy, x axis, C_α-RMSD. The blue points at the lower left of the panels represent models generated in trajectories starting with the native structure.

Figure S5: Free energy landscapes after round 2 for the 16 test proteins. y axis, Rosetta all atom energy, x axis, C_α-RMSD. The blue points at the lower left of the panels represent models generated in trajectories starting with the native structure.

1. K. T. Simons *et al.*, *Proteins* **34**, 82 (Jan 1, 1999).
2. T. Lazaridis, M. Karplus, *Proteins* **35**, 133 (May 1, 1999).

3. A. V. Morozov, T. Kortemme, K. Tsemekhman, D. Baker, *Proc Natl Acad Sci U S A* **101**, 6946 (May 4, 2004).
4. T. Kortemme, A. V. Morozov, D. Baker, *J Mol Biol* **326**, 1239 (Feb 28, 2003).
5. R. L. Dunbrack, Jr., F. E. Cohen, *Protein Sci* **6**, 1661 (Aug, 1997).
6. B. Kuhlman *et al.*, *Science* **302**, 1364 (Nov 21, 2003).
7. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (Sep 1, 1997).
8. K. T. Simons, R. Bonneau, I. Ruczinski, D. Baker, *Proteins* **Suppl 3**, 171 (1999).
9. P. Bradley *et al.*, *Proteins* **53 Suppl 6**, 457 (2003).
10. N. A. Metropolis, A. W. Rosenbluth, N. M. Rosenbluth, A. H. Teller, E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).
11. D. Shortle, K. T. Simons, D. Baker, *Proc Natl Acad Sci U S A* **95**, 11158 (Sep 15, 1998).
12. A. A. Canutescu, R. L. Dunbrack, Jr., *Protein Sci* **12**, 963 (May, 2003).
13. Z. Li, H. A. Scheraga, *Proc Natl Acad Sci U S A* **84**, 6611 (Oct, 1987).
14. C. A. Rohl, C. E. Strauss, K. M. Misura, D. Baker, *Methods Enzymol* **383**, 66 (2004).
15. B. Kuhlman, D. Baker, *Proc Natl Acad Sci U S A* **97**, 10383 (Sep 12, 2000).
16. R. Bonneau, C. E. Strauss, D. Baker, *Proteins* **43**, 1 (Apr 1, 2001).
17. H. M. Berman *et al.*, *Nucleic Acids Res* **28**, 235 (Jan 1, 2000).
18. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J Mol Biol* **247**, 536 (Apr 7, 1995).

Fig. S1

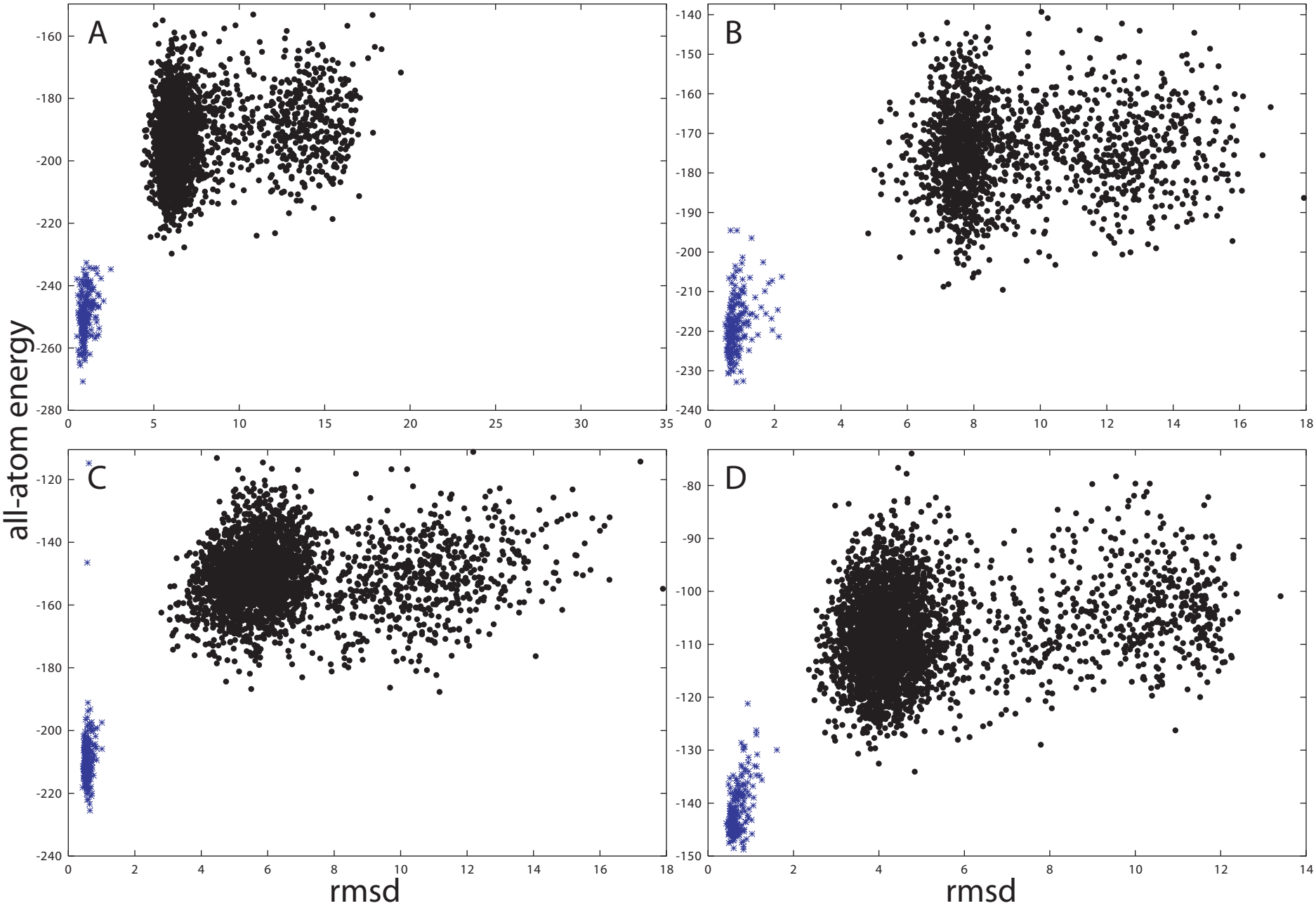


Fig. S2

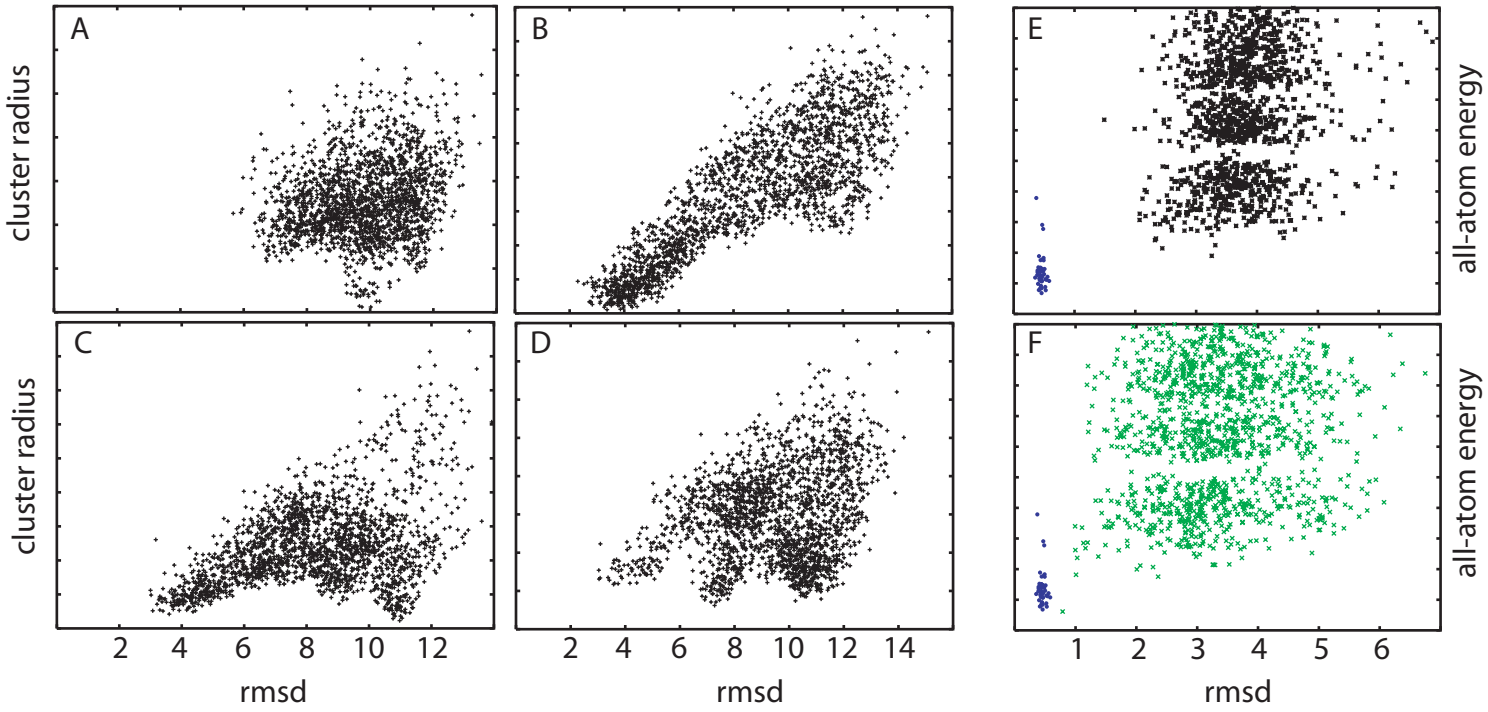

A
B
E
C
D
F
cluster radius
cluster radius
all-atom energy
all-atom energy
2 4 6 8 10 12
2 4 6 8 10 12 14
1 2 3 4 5 6
rmsd
rmsd
rmsd

Fig. S3

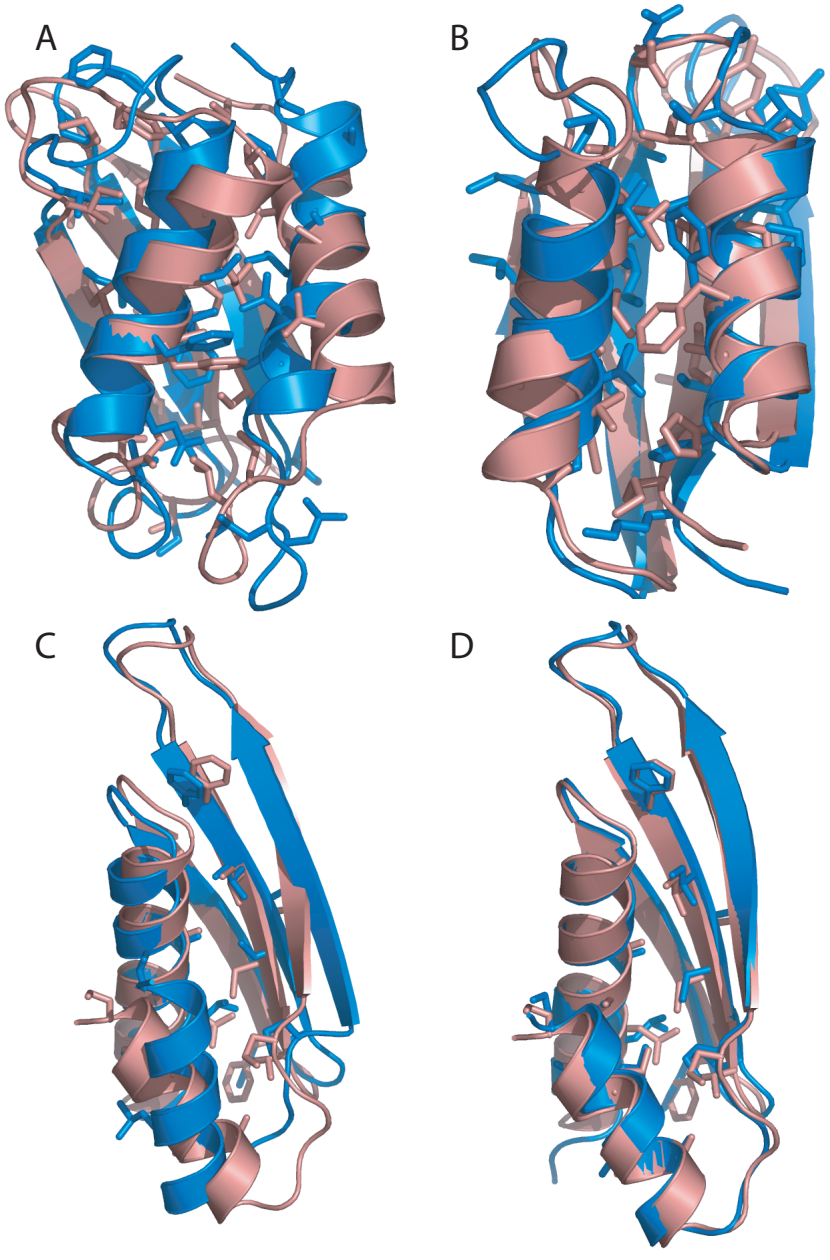


Fig. S4

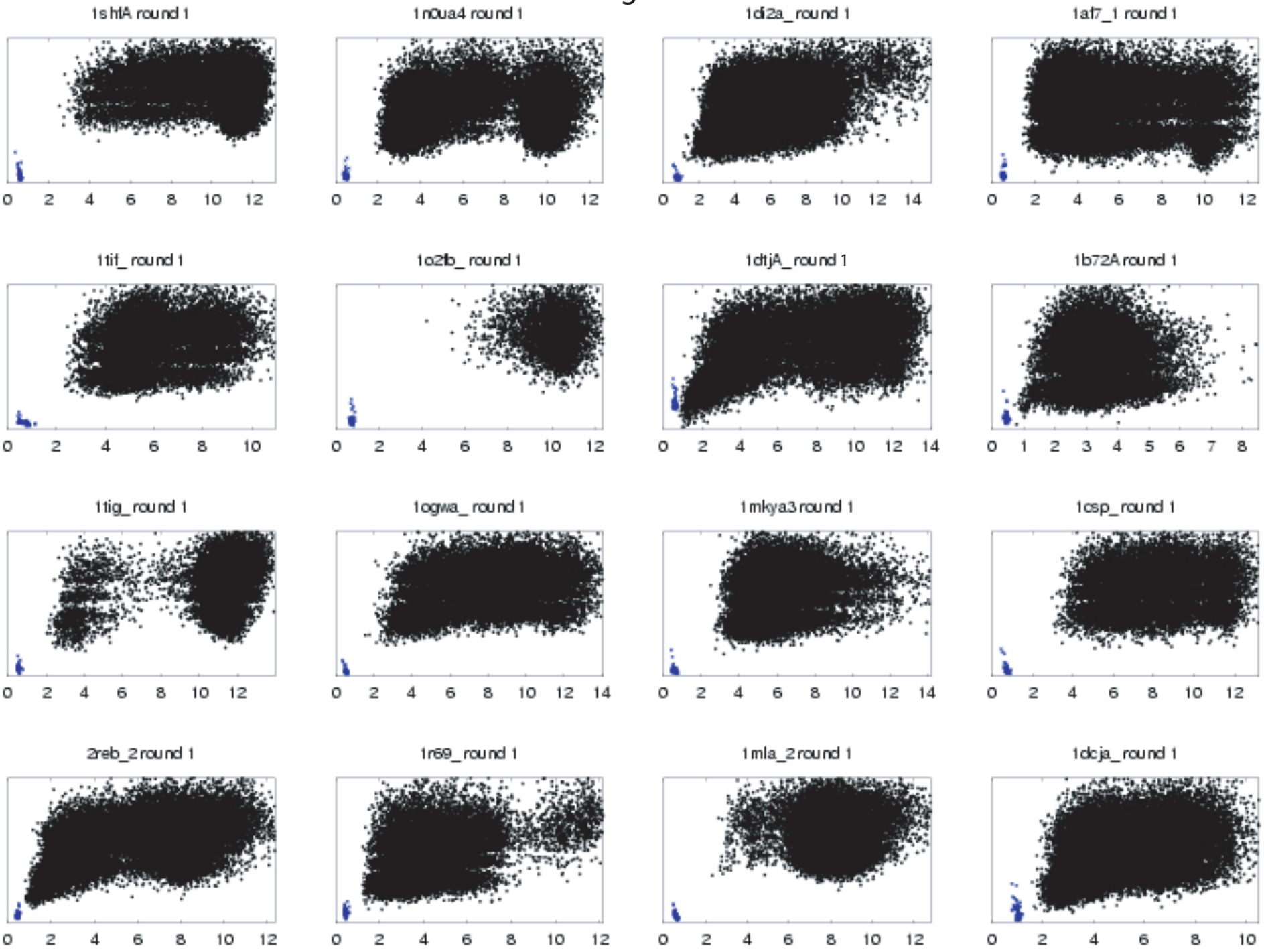
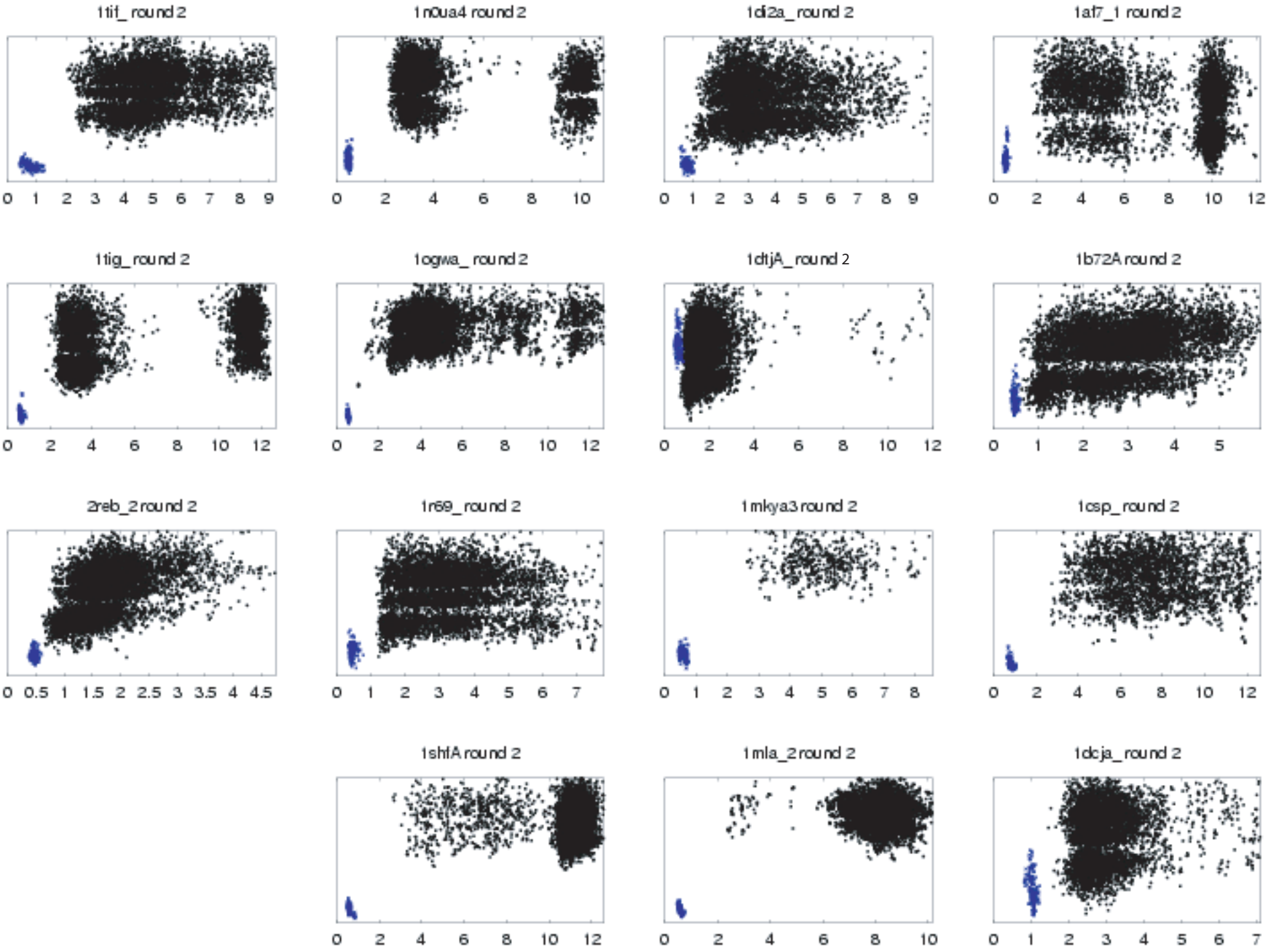

1shfA round 1
1n0ua4 round 1
1di2a_ round 1
1af7_1 round 1
1tif_ round 1
1o2fb_ round 1
1dtjA_ round 1
1b72A round 1
1tig_ round 1
1ogwa_ round 1
1mkya3 round 1
1csp_ round 1
2reb_2 round 1
1r69_ round 1
1mla_2 round 1
1dcja_ round 1

Fig S5


1tif_ round 2
1n0ua4 round 2
1di2a_ round 2
1af7_1 round 2
1tig_ round 2
1ogwa_ round 2
1dtjA_ round 2
1b72A round 2
2reb_2 round 2
1r69_ round 2
1mkya3 round 2
1csp_ round 2
1shfA round 2
1mla_2 round 2
1dcja_ round 2