Seven Amino Acid Types Suffice to Create the Core Fold of RNA Polymerase

Sota Yagi, Aditya K Padhi, Jelena Vucinic, Sophie Barbe, Thomas Schiex, Reiko Nakagawa, David Simoncini, Kam y J Zhang, Shunsuke Tagami

Seven amino acid types suffice to reconstruct the core fold of RNA polymerase

Authors: Sota Yagi[1], Aditya K. Padhi[1], Jelena Vucinic[2,3], Sophie Barbe[3], Thomas Schiex[2], Reiko Nakagawa[1], David Simoncini[4*], Kam Y. J. Zhang[1*], Shunsuke Tagami[1*]

Affiliations:

[1]RIKEN Center for Biosystems Dynamics Research, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

[2]Université Fédérale de Toulouse, ANITI, INRAE - UR 875, Toulouse, France

[3]TBI, Université Fédérale de Toulouse, CNRS, INRAE, INSA, ANITI, Toulouse, France

[4]Université Fédérale de Toulouse, ANITI, IRIT - UMR 5505, Toulouse, France.

[*]Corresponding Authors. D.S. (david.simoncini@gmail.com), K.Y.J.Z. (kamzhang@riken.jp), S.T. (shunsuke.tagami@riken.jp)

Abstract: The extant complex proteins must have evolved from ancient short and simple ancestors. Nevertheless, how such prototype proteins emerged on the primitive earth remains enigmatic. The double-psi beta-barrel (DPBB) is one of the oldest protein folds and conserved in various fundamental enzymes, such as the core domain of RNA polymerase. Here, by reverse engineering a modern DPBB domain, we reconstructed its evolutionary pathway started by "interlacing homo-dimerization" of a half-size peptide, followed by gene duplication and fusion. Furthermore, by simplifying the amino acid repertoire of the peptide, we successfully created the DPBB fold with only seven amino acid types (Ala, Asp, Glu, Gly, Lys, Arg, and Val), which can be coded by only GNN and ARR (R = A or G) codons in the modern translation system. Thus, the DPBB fold could have been materialized by the early translation system and genetic code.

Modern proteins with large and complex structures are generally thought to have evolved from small and simple ancient proteins with "prototype folds" (e.g., Rossmann fold, ferredoxin fold, and $(\beta/\alpha)_8$-barrel) (*1–10*). These prototype folds must have played essential roles in the early evolution of life, as they are often conserved in fundamental biochemical pathways such as metabolism, replication, transcription, and translation (*11–14*). However, it remains elusive how such prototype folds emerged on the ancient earth, where the primitive translation system likely performed imprecise syntheses of short peptides composed of fewer amino acids as compared to modern proteins (*15–17*). Especially, the components of the earliest genetic code are still an open question, as the 9–13 amino acid types used in previous ancestral protein reconstructions are at scattered positions in the modern codon table (*18–21*).

The double-psi beta-barrel (DPBB) has one of the most important functions and complicated structures among such prototype folds. It is conserved in several enzymes in fundamental biochemical processes (Fig. S1) (*13*, *22–24*). For example, the formate dehydrogenase in the oldest carbon fixation system, the Wood-Ljungdahl pathway, possesses a DPBB domain (*25*). The DPBB and a few related small β-barrel folds are also often conserved in essential proteins from transcription and translation systems (*26*, *27*). Most notably, the active site of RNA polymerase from all cellular life is composed of two DPBB folds, and thus the original core of RNA polymerase may have emerged by the duplication of an ancestral DPBB (*14*, *28–30*).

The DPBB fold is a six-stranded β-barrel consisting of two pseudo-symmetric ββαβ units with detectable structural and sequence homologies (Fig. 1A). All six β-strands are aligned in an interdigitated manner, giving its pseudoknot-like topology. The loop connecting β1 and β2 (β1' and β2') crosses the loop between β2' and β3' (β2 and β3), thus generating a shape similar to the Greek letter ψ. Although some other prototype folds with similar pseudo-symmetries apparently originated by oligomerization and gene duplication of shorter peptides (*31*), it remains uncertain

if the interdigitating fold of DPBB could have formed in such a simple oligomerization process (*13*, *22*, *27*, *29*). So far, no modern DPBB structure composed of a perfect sequence repeat or a dimer of shorter peptides has been reported, and thus the origin of the DPBB fold remains elusive.

In this study, we demonstrate that the DPBB fold could have emerged by oligomerization and gene-duplication of a shorter and simpler peptide, by reconstructing the DPBB domains comprising perfect sequence repeats and then self-dimerizing peptides. Surprisingly, even chemically synthesized peptides could fold into the complicated pseudoknot-like topology. We also eliminated several amino acid types from the designs, and confirmed that only seven amino acid types (Ala, Asp, Glu, Gly, Lys, Arg, and Val) are sufficient for the DPBB fold. These amino acid types can be coded by only GNN and ARR (R = A or G) codons in the modern translation system. These results reveal the plausible ancient pathway for the emergence of the complicated prototype fold and transcription machinery, which coevolved with the early translation system and genetic code.

Reconstruction of DPBB domains with perfect sequence symmetry

We first searched for extant DPBB domains with high internal sequence homologies, as the starting models to reconstruct a DPBB domain with perfect sequence repeats. The DPBB domain of a molecular chaperone, valosin-containing protein (VCP) from *Thermoplasma acidophilum* (taVCP_DPBB), reportedly has relatively high sequence identity (37%) between its N- and C-terminal halves (Table S1, Fig. S1F) (*22*). We determined its crystal structure in the isolated form, to confirm that it adopts the DPBB fold even without the other domains of VCP (Fig. 1B, Table S2). taVCP_DPBB also shares high structural homology with the core domain of RNA polymerase (RMSD = 1.6 Å), indicating their common origin (Fig. S1G).

Using the sequence of taVCP_DPBB as a query, we searched for VCP_DPBB domains with higher internal sequence homology from other organisms and found that the VCP_DPBBs from *Methanopyrus kandleri* (mkVCP_DPBB) and *Aeropyrum pernix* (apVCP_DPBB) have 42% and 45% internal sequence identities, respectively (Tables S1, S3). We determined their crystal structures (Fig. 1C and D, Table S2) and confirmed that they also exhibit more precise structural symmetries than taVCP_DPBB (Fig. S2). Thus, these two VCP_DPBBs were estimated to have similar sequences and structures to the ancestral DPBB with perfect sequence repeats. Furthermore, mkVCP_DPBB and apVCP_DPBB were from extreme thermophiles and showed high thermostability (T_m > 69 ˚C) and refolding ability (Fig. S3 and Table S4). The coexistence of the ancestral symmetric feature and the extreme thermostability in these DBPP domains is also consistent with the widely supported hypothesis suggesting that the last universal common ancestor (LUCA) was a hyper-thermophilic organism (*32*, *33*).

We then reconstructed DPBB domains with perfect internal repeats, by a method based on symmetrically-conserved positions (SC-design, Supplementary Text, Figs. S4–6, Tables S1, S2, S5). We engineered mkVCP_DPBB and apVCP_DPBB by replacing the residues in their non-symmetric positions with the symmetrically-conserved residues in VCP_DPBBs from different organisms. A few cycles of mutagenesis and structural confirmation by circular dichroism (CD), SEC, and X-ray crystallography finally resulted in two mutants, mkDPBB_sym_86 and apDPBB_sym_84, with 86% and 84% internal sequence identities, respectively (Fig. 1E and Table S1). These designs only have a limited number of non-symmetric positions clustered in two areas in their tertiary structures, defined as “cluster-1 and -2” in Fig. 1E (also see Fig. S7). By grafting the amino acid residues from each of the two clusters to the other one, we then designed four DPBBs with perfect internal sequence identities (mkDPBB_sym1, mkDPBB_sym2, apDPBB_sym1, and apDPBB_sym2, Table S1). Although one of them (apDPBB_sym2) could not

be purified due to its poor stability, the other three were readily purified. They eluted as monomeric proteins in SEC analyses, exhibited similar α/β CD spectra with the native DPBB proteins, and retained high thermostability ($T_m \geq 85^{\circ}C$, Fig. S5, Table S5). Finally, we solved the crystal structures of mkDPBB_sym1 and mkDPBB_sym2 (Fig. 1F, Table S2), which have almost perfect structural symmetries. These results strongly indicate that the DPBB fold originally arose from a perfectly symmetric ancestor.

Computational designs and possible sequence diversity of symmetric DPBBs

We also implemented two computational approaches for symmetric designs of DPBB to test if diverse strategies could result in different/similar sequences. The first computational approach was a modified "reverse engineering evolution" (*34*) (RE-design, Supplementary Text, Fig. 2A, Fig. S8). In this methodology, we constructed a phylogenetic tree with the respective aligned sequences, which were subsequently used as input to generate the ancestral sequences. The predicted ancestral sequences were mapped onto a manually constructed, perfectly symmetrical mkVCP structural backbone model, and were evaluated by their Rosetta energy scores (Fig. S9). The seven top-scoring designs were chosen for the experimental evaluation (Table S1).

The second computational approach was a multi-state computational protein design (MS-design, Methods, Fig. 2B, Fig. S10). This method describes the target protein structure as an ensemble of fixed backbone conformational states, to account for protein flexibility (*35*). The sequence that minimizes the average energy over all conformational states is computed by the AI automated reasoning prover ToulBar2 (*36*). The MS-designs were less homologous to the other designs since they were obtained by energy minimization, without any knowledge extraction from other homologous DPBB domains. From the top-scoring designs, we chose the two with the highest

sequence dissimilarities from the SC- or RE-designs for the experimental evaluation (Fig. S11, Table S1).

We tried to express the seven RE-designs and two MS-designs in *E. coli*. Although two RE-designs were not expressed well, the other designs were readily purified, exhibited similar properties to mkVCP_DPBB in SEC and CD analyses, and retained high thermostability ($T_m \geq 67$ ˚C, Figs. S12 and S13, Table S5). Furthermore, we solved the crystal structures of three RE-designs and one MS-design and confirmed that they adopt the designed structures (Fig. 2D, Fig. S6B–E, Table S2).

Interestingly, while all ten purified designs (three SC-designs and seven computational designs) have almost identical properties (structure, thermostability), their sequences are very diverse (Fig. 2C and Fig. S11, Supplementary Text). Only 34 positions (17 positions × 2 repeats) were perfectly conserved in the ~90 a.a. designs (Fig. 2D). The high sequence diversity and success ratios of these designs (Table 1) indicate that the symmetric DPBB fold can be adopted by a significant variety of sequences and had a high probability of emerging during the early evolution of life.

Homo-dimerization of halved fragments

To further investigate if the DPBB structure can be formed by the homo-dimerization of halved fragments (~46 aa) of symmetric DPBBs, we expressed the N-terminal halves of the four SC-designs (mk1h, mk2h, ap1h, and ap2h; Fig. 3A and 3B and Table S1). All four fragments were expressed as soluble peptides in *E. coli* and formed dimers with α/β structures (Fig. S14 and Table S6). Crystallographic analyses of mk2h and ap1h demonstrated that the halved fragments adopt the DPBB fold by interlacing homo-dimerization (Fig. 3C, Fig. S6F and Table S2). They also exhibited high thermostability and refoldability after heat denaturation (Fig. 3D, 3E, Fig. S14 and

Table S6). These results strongly indicate that the DPBB fold originally emerged simply by the homo-dimerization of a short peptide, in spite of its complicated interlaced topology.

We also tested the foldability of the chemically-synthesized mk2h peptide. First, the dried powder of mk2h (95.14% purity) was dissolved in 20 mM Bis-tris HCl, pH 6.0, with 150 mM NaCl, and subjected to crystallization screening. We readily obtained some crystals under various conditions (Table S7) and determined its structure to confirm that it also adopts the DPBB structure (Fig. 3C, Table S2). Thus, the chemically-synthesized peptide can fold precisely into the interlacing DPBB structure without any factors/environment in the cell, demonstrating that the amino acid sequence of mk2h encodes its homo-dimerizing and folding information.

Next, to investigate whether the peptide could fold even in the presence of contaminants with similar sequences, we analyzed the foldability of a low-purity sample of mk2h containing byproducts from the chemical synthesis (71.16% purity). In the SEC analysis, two major peaks corresponding to aggregated and dimer species appeared (Fig. 3F). The dimer fraction showed the typical CD spectra for α/β proteins, while the aggregated fraction exhibited a disordered conformation (Fig. 3G). The LC/MS analysis revealed that most of the contaminants were enriched in the aggregated fraction (e.g., the 4748.6 Da byproduct corresponding to a serine deletion, Fig. 3H and Table S8). In contrast, the full-length peptide (4835.6 Da) was enriched in the dimer fraction. This auto-purification phenomenon during protein folding was also observed with another design, mk1h (Fig. S15 and Table S8). Therefore, the homo-dimerization and folding processes of the peptides likely worked as a purification/selection system by excluding contaminated sequences, and thus might have enabled the production of the DPBB domains by an imprecise ancient translation system or prebiotic peptide synthesis.

Reduction of amino acid repertory

Interestingly, mk2h contained only 13 amino acid types, although we did not intend to enrich/exclude any specific amino acid species in the engineering process (Fig. 4A, Table S9). To examine how many and what kind of amino acid species are required to comprise a DPBB scaffold, we tried to further simplify the amino acid repertoire of mk2h (Fig. 4B). In mk2h, Ile, Leu, Met, Pro, and Ser were used only once or twice. Tyrosine was the only aromatic amino acid type and used just three times (Fig. 4A). Thus, we replaced each of these amino acid types with other amino acid residues conserved in different organisms or possessing similar chemical/structural properties (e.g., Ile to Val) to generate mutants containing 12 amino acid repertoires (Table S1). MD simulations estimated that none of the mutations would have a devastating effect on the domain structure (Fig. S16). All mutants were eluted as homo-dimers in SEC and their CD spectra were similar to that of mk2h (Fig. S17), although some of them exhibited lower T_m values than mk2h (Table S10). Furthermore, the crystal structures of mk2h_ΔP and mk2h_ΔY revealed they indeed adopt the DPBB fold (Fig. S6G and H, Table S2). Therefore, these amino acid species may have contributed to the thermostability of mk2h, but are not essential to form the DPBB fold.

Subsequently, we created a mutant in which Met, Ile, and Leu were eliminated simultaneously (mk2h_ΔMIL). The mutant was expressed as a homo-dimer and showed typical CD spectra for α/β proteins (T_m = 54.6 ºC, Fig. S18A and Table S10). The X-ray crystallographic analysis of mk2h_ΔMIL revealed it has a very primitive hydrophobic core composed of only Val and Ala, with some unfilled cavities (Fig. 4C). Recently, a *de novo* designed protein with a hydrophobic core composed mostly of Valine residues was reported, indicating that the backbone structure is the main contributor to its thermostability (*37*). GD-box, a structural motif conserved among DPBB and various protein folds, has also been suggested to stabilize the protein tertiary structures by tethering noncontiguous segments with hydrogen bonds between the main chains (*38*). Such a stable backbone of the DPBB fold would have allowed the emergence of a globular protein with

only simple and small hydrophobic amino acids during early protein evolution. Unlike hydrophilic interactions, hydrophobic interactions do not require precise angles or directions, and thus such simple hydrophobic cores might emerge relatively easily in aqueous environments without optimizing the sizes or compositions of their amino acid residues.

We also tried to exclude two amino acid types (Pro/Ser, Pro/Tyr, or Ser/Tyr) in mk2h_ΔMIL (mk2h_ΔMILPS, mk2h_ΔMILPY and mk2h_ΔMILSY; Fig. 4B and Table S1). While mk2h_ΔMILPY and mk2h_ΔMILSY were eluted as unfolded aggregates in SEC, mk2h_ΔMILPS eluted as a dimer and its CD spectra were similar to those of mk2h (Fig. S18B–D and Table S10). The crystal structures of *E. coli*-produced and chemically-synthesized mk2h_ΔMILPS were also determined (Fig. S6I and J, Table S2). Although mk2h_ΔMILPS remained thermostable (T_m = 50.5 ˚C), refoldability was not observed (Fig. S18B).

Finally, we designed an mk2h variant, mk2h_ΔMILPYS, with only seven amino acid types (Ala, Asp, Glu, Gly, Lys, Arg, and Val) by combining the mutations in mk2h_ΔMILPS and mk2h_ΔY (Fig. 4B). Although the peptides exhibited unfolded properties in the SEC and CD analyses (Fig. S18E), we obtained crystals from two conditions containing the chemically-synthesized peptide (Fig. S19). The crystal structures demonstrated that mk2h_ΔMILPYS adopts the DPBB structure through homo-dimerization (Fig. 4D, Fig. S6K, and Table S2). In the structures, the positively-charged pocket conserved in the extant VCP_DPBB is occupied by the fundamental metabolites, malonate or malic acid, contained in the crystallization conditions (Fig. 4D, Fig. S20), which might also have assisted in the folding of mk2h_ΔMILPYS. Thus, despite its limited folding propensity, the 43 a.a. peptide with only seven amino acid types can homo-dimerize and fold into the DPBB structure. Additional amino acids would have been incorporated for higher stability and foldability during protein evolution (Fig. 4B).

Discussion

In this study, we demonstrated that one of the prototypic protein folds, DPBB, can be reconstructed by perfect sequence repeats (Figs. 1, 2), as well as by dimers of the halved fragments (Fig. 3), indicating that DPBB originally emerged by the self-dimerization of ~40 a.a. peptides and then evolved via gene duplication and fusion. Chemically synthesized peptides were able to fold into the complicated interlaced topology, without any support from the modern biological machinery, and still retained thermostability and refoldability (Fig. 3). Furthermore, the success rate and diversity of our designs were surprisingly high (Fig. 2C and Fig. S11, Table 1). The fragmented peptides also exhibited auto-purification ability (Fig. 3H) and high amenability to engineering (Fig. 4B). Thus, the DPBB fold probably had a significant chance of emerging from the pool of primitive peptides synthesized by an immature/imprecise translation system on the early earth. Nature can find and utilize such a complicated protein fold as long as it is stable enough, even though in theory it appears unrealistic to the human eye.

By simplifying mk2h, we further constructed the DPBB fold with only seven amino acid types, Ala, Asp, Glu, Gly, Lys, Arg, and Val (mk2h_ΔMILPYS, Fig. 4). The previously reconstructed ancestral proteins with ~10 amino acid types also contain most of them (*19–21*, *39*), indicating that they were shared by various prototype proteins. Interestingly, these seven amino acid types can be coded on a clearly defined area in the standard codon table (GNN and ARR) (Fig. 4E). The five amino acids coded by GNN (Ala, Asp, Glu, Gly, and Val) were probably adopted into the earliest genetic code because they can be easily produced in the prebiotic environment (*16*, *17*, *40–42*). The other two amino acids (Arg and Lys) have cationic side chains and tend to interact with nucleic acid polymers. Arginine could have been abiotically synthesized by sharing some precursors to nucleotides (*42*). Recent studies have shown that peptide analogs enriched with such basic residues can be prebiotically synthesized and mutually stabilize RNA (*43*, *44*). Simple

peptides containing lysine residues could also enhance the activities of ribozymes (*45*). Considering the fact that the genetic code is realized by RNA-based machinery, arginine and lysine were probably readily recruited into the early genetic code during evolution. This idea is also supported by the fact that the core domain of RNA polymerase, the enzyme responsible for the synthesis of rRNA, mRNA, and tRNA, is composed of the DPBB fold. Like its modern descendants, the ancient symmetric DPBB fold might have interacted with nucleotide-related molecules, as the conserved positively-charged pockets on some symmetrized or simplified designs are occupied by negatively-charged ligands in their crystal structures (Fig. S20). Thus, the DPBB fold was likely established at an early evolutionary stage of the genetic code and supported the ancient RNA-based biosystem when only ~7 amino acid types were available.

References

1. R. V. Eck, M. O. Dayhoff, Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science (80-.).* **152**, 363–366 (1966).
2. J. Söding, A. N. Lupas, More than the sum of their parts: On the evolution of proteins from peptides. *BioEssays*. **25**, 837–846 (2003).
3. M. L. Romero Romero, A. Rabin, D. S. Tawfik, Functional Proteins from Short Peptides: Dayhoff's Hypothesis Turns 50. *Angew. Chemie - Int. Ed.* **55**, 15966–15971 (2016).
4. S. Setiyaputra, J. P. MacKay, W. M. Patrick, The structure of a truncated phosphoribosylanthranilate isomerase suggests a unified model for evolution of the (βα)8 barrel fold. *J. Mol. Biol.* **408**, 291–303 (2011).
5. D. Lang, R. Thoma, M. Henn-Sax, R. Sterner, M. Wilmanns, Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. *Science (80-.).* **289**, 1546–1550 (2000).
6. J. M. Thornton, C. A. Orengo, A. E. Todd, F. M. G. Pearl, Protein folds, functions and evolution. *J. Mol. Biol.* **293**, 333–342 (1999).
7. V. Alva, J. Söding, A. N. Lupas, A vocabulary of ancient peptides at the origin of folded proteins. *Elife*. **4**, 1–19 (2015).
8. B. G. Ma, L. Chen, H. F. Ji, Z. H. Chen, F. R. Yang, L. Wang, G. Qu, Y. Y. Jiang, C. Ji, H. Y. Zhang, Characters of very ancient proteins. *Biochem. Biophys. Res. Commun.* **366**, 607–611 (2008).
9. M. Henn-Sax, B. Höcker, M. Wilmanns, R. Sterner, Divergent evolution of (βα)8-barrel enzymes. *Biol. Chem.* **382**, 1315–1320 (2001).
10. B. Höcker, Design of proteins from smaller fragments-learning from evolution. *Curr. Opin. Struct. Biol.* **27**, 56–62 (2014).

11. H. Raanan, S. Poudel, D. H. Pike, V. Nanda, P. G. Falkowski, Small protein folds at the root of an ancient metabolic network. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 7193–7199 (2020).
12. A. D. Goldman, R. Samudrala, J. A. Baross, The evolution and functional repertoire of translation proteins following the origin of life. *Biol. Direct*. **5** (2010), doi:10.1186/1745-6150-5-15.
13. R. M. Castillo, K. Mizuguchi, V. Dhanaraj, A. Albert, T. L. Blundell, A. G. Murzin, A six-stranded double-psi β barrel is shared by several protein superfamilies. *Structure*. **7**, 227–236 (1999).
14. E. V. Koonin, M. Krupovic, S. Ishino, Y. Ishino, The replication machinery of LUCA: Common origin of DNA replication and transcription. *BMC Biol.* **18**, 1–8 (2020).
15. E. V. Koonin, A. S. Novozhilov, Origin and evolution of the genetic code: The universal enigma. *IUBMB Life*. **61**, 99–111 (2009).
16. K. Ikehara, Y. Omori, R. Arai, A. Hirose, A novel theory on the origin of the genetic code: A GNC-SNS hypothesis. *J. Mol. Evol.* **54**, 530–538 (2002).
17. K. Macé, R. Gillet, Origins of tmRNA: The missing link in the birth of protein synthesis? *Nucleic Acids Res.* **44**, 8041–8051 (2016).
18. K. U. Walter, K. Vamvaca, D. Hilvert, An active enzyme constructed from a 9-amino acid alphabet. *J. Biol. Chem.* **280**, 37742–37746 (2005).
19. L. M. Longo, J. Lee, M. Blaber, Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2135–2139 (2013).
20. R. Shibue, T. Sasamoto, M. Shimada, B. Zhang, A. Yamagishi, S. Akanuma, Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. *Sci. Rep.* **8**, 1–8 (2018).
21. M. Kimura, S. Akanuma, Reconstruction and Characterization of Thermally Stable and Catalytically Active Proteins Comprising an Alphabet of ~ 13 Amino Acids. *J. Mol. Evol.* **88**, 372–381 (2020).
22. M. Coles, T. Diercks, J. Liermann, A. Gröger, B. Rockel, W. Baumeister, K. K. Koretke, A. Lupas, J. Peters, H. Kessler, The solution structure of VAT-N reveals a "missing link" in the evolution of complex enzymes from a simple βαββ element. *Curr. Biol.* **9**, 1158–1168 (1999).
23. O. Erbilgin, M. Sutter, C. A. Kerfeld, The Structural Basis of Coenzyme A Recycling in a Bacterial Organelle. *PLoS Biol.* **14**, 1–20 (2016).
24. Y. Nishitani, R. Aono, A. Nakamura, T. Sato, H. Atomi, T. Imanaka, K. Miki, Structure analysis of archaeal AMP phosphorylase reveals two unique modes of dimerization. *J. Mol. Biol.* **425**, 2709–2721 (2013).
25. D. Niks, R. Hille, Molybdenum- and tungsten-containing formate dehydrogenases and formylmethanofuran dehydrogenases: Structure, mechanism, and cofactor insertion. *Protein Sci.* **28**, 111–122 (2019).
26. P. Youkharibache, S. Veretnik, Q. Li, K. A. Stanek, C. Mura, P. E. Bourne, The Small β-Barrel Domain: A Survey-Based Structural Analysis. *Structure*. **27**, 6–26 (2019).
27. V. Alva, K. K. Koretke, M. Coles, A. N. Lupas, Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. *Curr. Opin. Struct. Biol.* **18**, 358–365 (2008).
28. L. Sauguet, The Extended "Two-Barrel" Polymerases Superfamily: Structure, Function and Evolution. *J. Mol. Biol.* **431**, 4167–4183 (2019).
29. Z. F. Burton, The old and new testaments of gene regulation: Evolution of multi-subunit RNA polymerases and co-evolution of eukaryote complexity with the RNAP II CTD. *Transcription*. **5** (2014), doi:10.4161/trns.28674.
30. T. Fouqueau, F. Blombach, F. Werner, Evolutionary Origins of Two-Barrel RNA Polymerases and Site-Specific Transcription Initiation. *Annu. Rev. Microbiol.* **71**, 331–348 (2017).

31. V. Alva, A. N. Lupas, From ancestral peptides to designed proteins. *Curr. Opin. Struct. Biol.* **48**, 103–109 (2018).
32. A. Nasir, K. M. Kim, G. Caetano-Anollés, A Phylogenomic Census of Molecular Functions Identifies Modern Thermophilic Archaea as the Most Ancient Form of Cellular Life. *Archaea*. **2014** (2014), doi:10.1155/2014/706468.
33. S. Akanuma, Y. Nakajima, S. I. Yokobori, M. Kimura, N. Nemoto, T. Mase, K. I. Miyazono, M. Tanokura, A. Yamagishi, Experimental evidence for the thermophilicity of ancestral life. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11067–11072 (2013).
34. A. R. D. Voet, H. Noguchi, C. Addy, D. Simoncini, D. Terada, S. Unzai, S. Y. Park, K. Y. J. Zhang, J. R. H. Tame, Computational design of a self-assembling symmetrical β-propeller protein. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15102–15107 (2014).
35. J. Vucinic, D. Simoncini, M. Ruffini, S. Barbe, T. Schiex, Positive multistate protein design. *Bioinformatics*. **36**, 122–130 (2020).
36. M. C. Cooper, S. de Givry, M. Sanchez, T. Schiex, M. Zytnicki, T. Werner, Soft arc consistency revisited. *Artif. Intell.* **174**, 449–478 (2010).
37. R. Koga, M. Yamamoto, T. Kosugi, N. Kobayashi, T. Sugiki, T. Fujiwara, N. Koga, Robust folding of a de novo designed ideal protein even with most of the core mutated to valine. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 31149-31156 (2020).
38. V. Alva, S. Dunin-Horkawicz, M. Habeck, M. Coles, A. N. Lupas, The GD box: A widespread noncontiguous supersecondary structural element. *Protein Sci.* **18**, 1961–1966 (2009).
39. L. M. Longo, D. Despotović, O. Weil-Ktorza, M. J. Walker, J. Jabłońska, Y. Fridmann-Sirkis, G. Varani, N. Metanis, D. S. Tawfik, Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 15731–15739 (2020).
40. S. L. Miller, A production of amino acids under possible primitive earth conditions. *Science (80-.).* **117**, 528–529 (1953).
41. A. P. Johnson, H. J. Cleaves, J. P. Dworkin, D. P. Glavin, A. Lazcano, J. L. Bada, The Miller volcanic spark discharge experiment. *Science (80-.).* **322**, 404 (2008).
42. B. H. Patel, C. Percivalle, D. J. Ritson, C. D. Duffy, J. D. Sutherland, Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nat. Chem.* **7**, 301–307 (2015).
43. M. Frenkel-Pinter, J. W. Haynes, A. M. Mohyeldin, C. Martin, A. B. Sargon, A. S. Petrov, R. Krishnamurthy, N. V. Hud, L. D. Williams, L. J. Leman, Mutually stabilizing interactions between proto-peptides and RNA. *Nat. Commun.* **11**, 1–14 (2020).
44. M. Frenkel-Pinter, J. W. Haynes, C. Martin, A. S. Petrov, B. T. Burcar, R. Krishnamurthy, N. V. Hud, L. J. Leman, L. D. Williams, Selective incorporation of proteinaceous over nonproteinaceous cationic amino acids in model prebiotic oligomerization reactions. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 16338–16346 (2019).
45. S. Tagami, J. Attwater, P. Holliger, Simple peptides derived from the ribosomal core potentiate RNA polymerase ribozyme function. *Nat. Chem.* **9**, 325–332 (2017).

Acknowledgements

This work is based on experiments performed at KEK (project number: 20G056) and SPring-8. The authors are grateful to the beamline staff scientists at KEK, SPring-8, and SLS. We thank Hideaki Niwa, Toshiaki Hosaka, and Kentaro Ihara for helping with the X-ray diffraction experiments. We also thank Shigehiro Kuraku for assistance in the LC-MS analysis. We are deeply

grateful to Ryutaro Furukawa for providing an informative in-house protein database to search for natural DPBB proteins. We acknowledge RIKEN ACCC for the supercomputing resources at the Hokusai BigWaterfall supercomputer used in this study. A.K.P. acknowledges the Japan Society for the Promotion of Science (JSPS), Govt. of Japan, for the research fellowship. S.Y., K.Y.J.Z. and S.T. were supported by JSPS (20K15854, 18H02395 and 18H01328). This work was also supported by the French ANR through an ANR-19-PI3A-0004 grant. We thank the CALMIP HPC center for computational resources. We thank Satoshi Akanuma, Hiroshi Sasaki, Loren D. Williams, and Claudia Alvarez-Carreno for fruitful discussions.

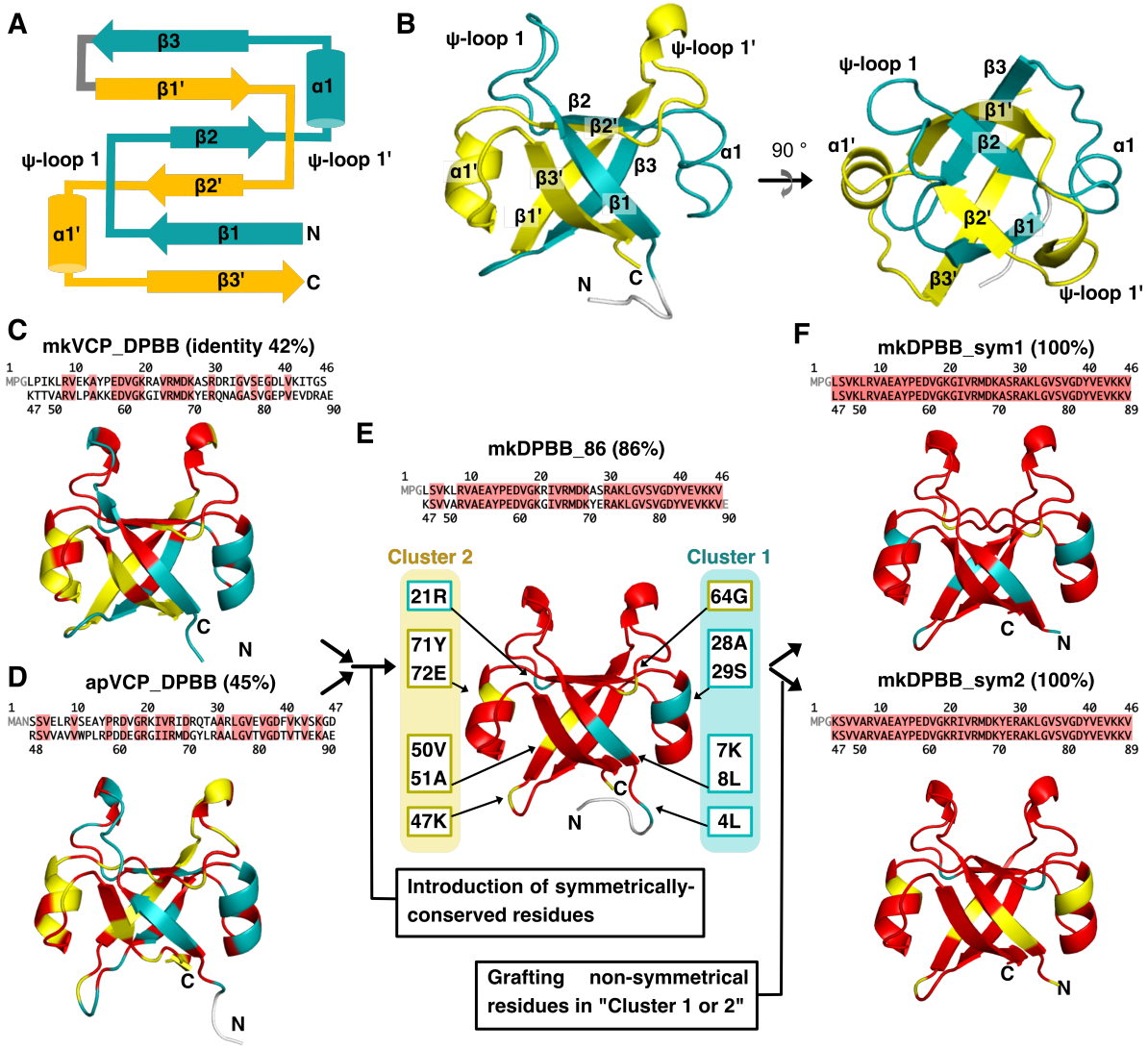

Figure 1. Structures and symmetrical engineering of the DPBB domains. (A) Topology diagram of the DPBB fold. The N- and C-terminal ββαβ units are colored cyan and yellow. **(B)** The crystal structure of taVCP_DPBB. Both halves have the same color as in the topology diagram of panel A. The right view shows the left view rotated by 90 degrees around the horizontal axis. **(C–F)** Symmetric-conservation design (SC-design). The amino acid sequence and crystal structure of (C) mkVCP_DPBB, (D) apVCP_DPBB, (E) mkDPBB_86, and (F) mkDPBB_sym1 and 2 are shown in each panel. The conserved residues in both halves are highlighted in red. (E)

mkDPBB_86 has only a limited number of non-symmetric positions, which are grouped in two areas (clusters 1 and 2).

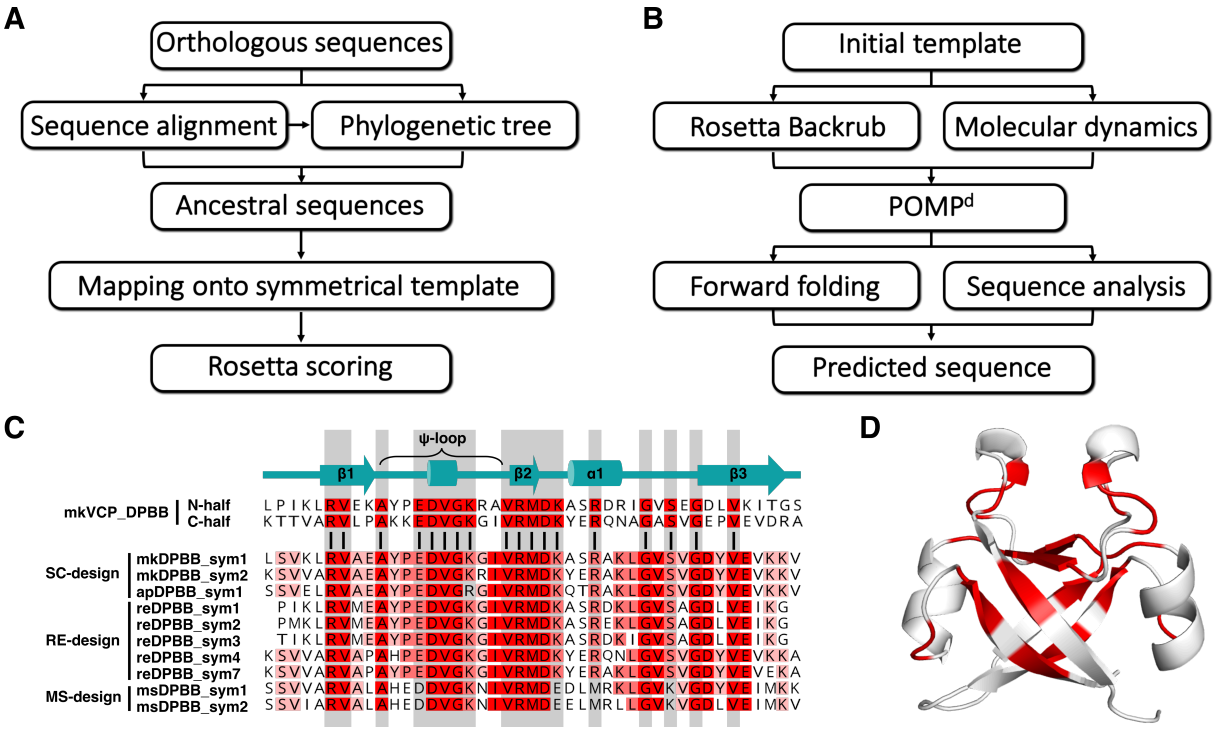

Figure 2. Computational design of symmetric DPBB and diversity of the symmetric DPBBs. **(A)** Flow-chart of the RE-design scheme. **(B)** Flow-chart of the MS-design scheme. **(C)** Multiple sequence alignment of one repeat unit in the symmetric designs along with the mkVCP_DPBB, the starting template DPBB. The columns corresponding to the consensus residues in mkVCP_DPBB are highlighted in gray. The perfectly identical columns in the symmetric designs are colored red, and the columns with sequence identities over 60% are colored pink. **(D)** The perfectly conserved residues among all experimentally confirmed symmetric designs are mapped on the crystal structure of reDPBB_sym1.

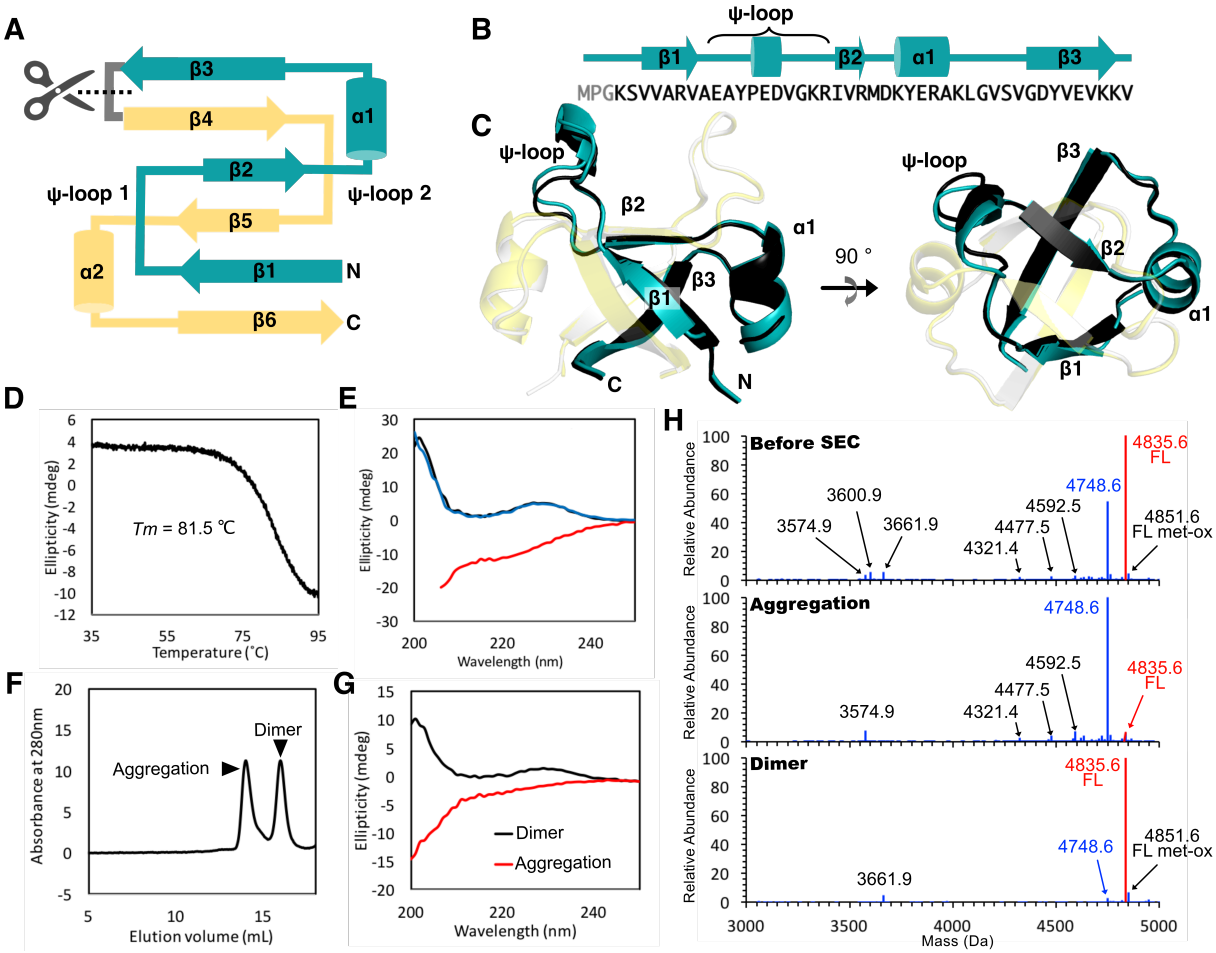

Figure 3. The DPBB fold formed by halved fragments. (A) Topology diagram of the halved fragments. **(B)** The sequence and secondary structure of mk2h. **(C)** The crystal structures of homo-dimeric mk2h. The *E. coli*-produced peptides are colored cyan and yellow. The chemically synthesized peptides are colored black and white. **(D and E)** Thermostability and refolding ability of mk2h. (D) Denaturation was monitored by measuring the CD ellipticity at 222 nm. (E) CD spectra at 35 °C (black line), after heating to 95 °C (red line), and upon cooling back to 35 °C (blue line). **(F–H)** Folding of the low-purity mk2h sample (71.16%). (F) SEC analysis showing that the dissolved mk2h peptide adopted aggregated and dimeric states. (G) CD spectra indicating that the aggregated and dimeric species, separated in Fig. 3F, adopt random-coil and α/β structures, respectively. (H) The peptide species in the sample before and after SEC purification were

analyzed by LC/MS. The deconvoluted mass spectra are shown. The labels for the full-length mk2h peptide (4835.6 Da) and the major contaminant peptide (4748.6 Da) are highlighted in red and blue, respectively.

5

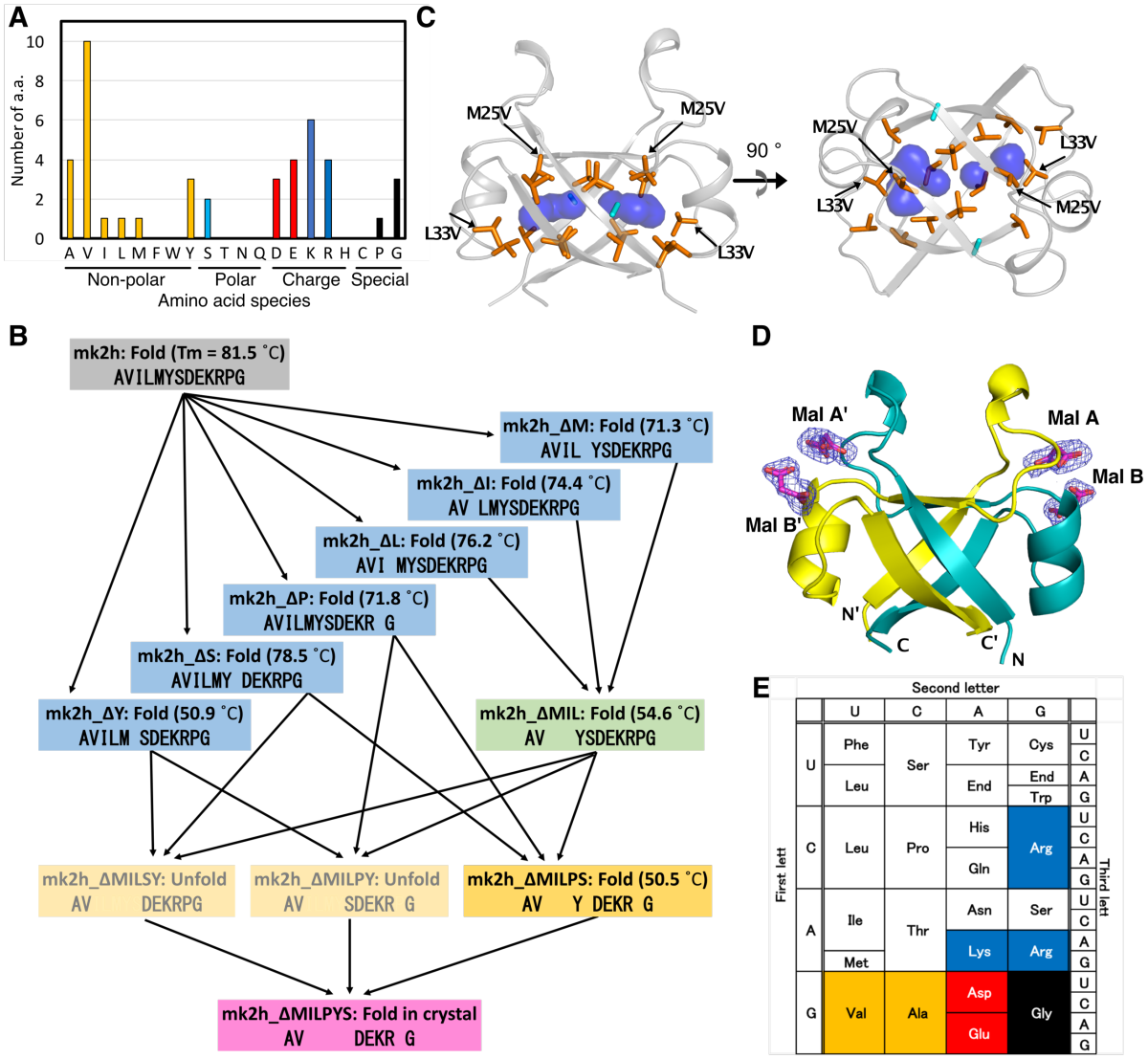

Figure 4. Reduction of amino acid repertory in mk2h. (A) The amino acid usage in mk2h. **(B)** Scheme of the design process of the simplified mk2h variants. **(C)** The hydrophobic core of mk2h_ΔMIL, composed of only Val (orange) and Ala (cyan). The substituted residues (M25V and L33V) from mk2h are indicated. The cavities in the core are shown as blue surface models. **(D)** The crystal structure of mk2h_ΔMILPYS. Four malonate ions, Mal A (A') and B (B') are bound to two symmetrical positions. The malonates are shown by stick models, and their Fo-Fc electron density omit maps are presented as a blue mesh (contoured at 1.0σ). **(E)** The seven amino acid species used in mk2h_ΔMILPYS are highlighted on the standard codon table.

Table 1. Summary of experimental tests for three different design strategies.

	Tested designs	Express [a]	Soluble [a]	α/β structure	Monomeric [b]	Crystal structure
SC-design	4	4	3	3	3	2
RE-design	7	5	5	5	5	3
MS-design	2	2	2	2	2	1

[a] Expressions and solubilities were examined by SDS-PAGE.
[b] Oligomeric states were examined by size exclusion chromatography.

SUPPLEMENTARY MATERIALS

Materials and Methods

Identification of modern DPBB sequences with high internal symmetry

Sequences of the molecular chaperone VCP were identified from the in-house protein database with reduced taxonomic bias (*1*) by searching with the BlastP program (*2*) using the taVCP_DPBB sequence as a query, and only the DPBB domains were extracted from the full-length VCP proteins. The 249 sequences thus obtained were aligned with the Muscle program. We fragmented the aligned sequences at the loop between the N-half and C-half ββαβ elements, by referring to the structural information of taVCP_DPBB, and re-aligned them together. Based on the obtained multiple alignments, the internal sequence identities in each organism's DPBB were evaluated and ranked, as shown in Table S3.

SC-design

Perfectly symmetric DPBBs were constructed by introducing mutations in a stepwise manner, using mkVCP_DPBB and apVCP_DPBB as templates. Initially, mkDPBB_sym_65 (internal sequence identity: 65%) and apDPBB_sym_63 (63%) were constructed by introducing the symmetrically-conserved residues in the N- and C-halves of mkVCP_DPBB and apVCP_DPBB into each other. The conserved residues in the other eight DPBBs with high internal sequence identities were then introduced into mkDPBB_sym_65 and apDPBB_sym_63 to construct mkDPBB_sym_79 (79%) and apDPBB_sym_79 (79%). The engineering step was repeated one more time using the sequence information from ten additional organisms, to construct mkDPBB_sym_86 (86%) and apDPBB_sym_84 (84%). mkDPBB_sym_86 and apDPBB_sym_84 have only a limited number of non-symmetrized positions, which are clustered in two areas in their tertiary structures. By adopting the amino acid residues from each of the two

non-symmetrized areas, we then designed four DPBB proteins with perfect internal sequence symmetries (mkDPBB_sym1, mkDPBB_sym2, apDPBB_sym1, and apDPBB_sym2).

RE-design

The design of completely symmetrical DPBB sequences was carried out using the "reverse engineering evolution" computational protein design approach (*3*). This method starts from the pseudo-symmetric sequences of each subunit, then constructs a phylogenetic tree, and subsequently generates putative ancestral sequences for further evaluation. This has been successfully applied to the design of symmetric proteins with 3-8 subunits (*4*, *5*). However, DPBB is an extreme case with only 2 subunits, which makes the prediction of ancestral sequences based on the phylogenetic tree constructed from only 2 sequences challenging. To overcome this challenge, we have introduced the use of orthologous sequences instead of pseudo-symmetric sequences for the construction of phylogenetic trees. We first collected 498 VCP sequences from different organisms and then removed the redundant sequences having >75% sequence identity. The resulting sequences were aligned to generate a phylogenetic tree using a maximum-likelihood method (Jones-Taylor-Thornton (JTT) model) with 50-100 bootstrapping (*6*, *7*). The aligned sequences and constructed phylogenetic tree were used together to generate putative ancestral consensus sequences by the FastML server, using a joint reconstruction substitution model (*8*). Approximately 28,000 ancestral sequences were then mapped onto a manually constructed, symmetrical mkVCP backbone structural model, and their energies were calculated using an in-house "sequence_mapping.py" program that utilizes the PyRosetta of Rosetta protein modeling suite (*9–11*). This program uses an input list of plausible sequences and maps them onto a template protein backbone structure to output a PDB model for each sequence and the associated score. The top-scored designs were then analyzed for Rosetta total score, and root mean square deviations

(RMSDs) from the symmetrical mkVCP backbone model structure. Finally, shortlisted designs were selected for experimental validation based on the Rosetta scores, RMSD from the design template, predicted solubility and visual inspection.

MS-design

1) POMP[d] (POsitive Multistate Protein design)

POMP[d] was used to compute minimum energy protein sequences from an ensemble of conformational states (*12*). Let us assume a rigid backbone and a pairwise decomposable energy function taking the form:

$$E(s) = \sum_{i=1}^{n} E_i(s_i) + \sum_{i<j} E_{ij}(s_i, s_j)$$

with $E(s)$ the total energy of protein sequence s of length n, $E_i(s_i)$ a unary energy term for residue s_i and $E_{ij}(s_i, s_j)$ a binary term representing energy interactions between residue pairs (s_i, s_j). In this context, POMP[d] looks for the sequence that minimizes the energy of an ensemble of conformational states described as the sum of the energies on each state. POMP[d] models this problem as a cost function network and solves it exactly, using the constraint programming prover toulbar2, by returning the global minimum of the energy function and proving its optimality. This deterministic approach provides optimality guarantees: given an ensemble of conformational states and an energy function, the sequence returned is the global minimum of the energy function.

2) Conformational state ensemble preparation

mkVCP_sym2 crystal structure was used as the initial template. Two strategies were used in order to generate conformational states. In the first one, 100 protein models were generated using Rosetta Backrub protocol, with harmonic restraints on initial atoms coordinates. In the second one,

a 100-ns molecular dynamics (MD) simulation at 300 K was performed starting from the initial template. Conformational states were extracted every 0.5 ns, generating 200 protein models. For each strategy, the resulting protein models were clustered using Durandal (*13*). The clustering radius was set to 0.15 Å for the first strategy and 0.5 Å for the second. In each case, the cluster centers of the 4 biggest clusters were selected as the ensemble of conformational states. From these two strategies, we obtained two ensembles of 4 conformational states: the backrub ensemble and the MD ensemble.

3) Protein sequence predictions

POMP[d] was used to compute optimal protein sequences from backrub and MD ensembles using two different setups, which allowed the prediction of a total of four protein sequences. For both setups, we used the ability of the toulbar2 prover to accept hard constraints in energy minimization to constrain sequences to be identical in each DPBB symmetrical subunit. In the first setup, all amino acid types were allowed at each position in a DPBB subunit. In the second setup, additional hard constraints were used in order to prevent the formation of solvent exposed hydrophobic patches.

While environmental factors such as pH, ionic strength, temperature, and the presence of various solvent additives may influence protein solubility, internal factors are defined by the amino acids present at the protein surface (*14*). Protein solubility is determined by the amount of exposed hydrophobic surface area in the protein folded state (*15*, *16*). Furthermore, the rate of aggregation of proteins and peptides increases as the amount of exposed hydrophobic surface area increases (*17*). Therefore, computational protein design tools must take surface hydrophobicity into account when designing new sequences. In order to do so, we limit the formation of exposed hydrophobic surface area by adding constraints in the form of new energy function terms. This functionality has

been implemented in POMP[d] an additional feature called hpatch. The hpatch procedure is described in Supplementary Algorithm S1.

Forward folding experiments were performed on the four predicted sequences, using the protein structure prediction software EdaRose (*18*). Forward folding aims at assessing the quality of a protein design by predicting whether it will fold into the target structure or not. For each sequence, 30,000 structural models were predicted with EdaRose (Fig. S10). The number of iterations of EdaRose was set to 6, and the beta_nov16 scoring function from the Rosetta modeling software was used.

After examination of sequences and forward folding results, two sequences were selected for experimental characterization. The first one (msDPBB_sym1), from the backrub ensemble and using hpatch, was selected based on its forward folding profile. The second one (msDPBB_sym2), from the backrub ensemble and without using hpatch, was selected due to its sequence dissimilarity with other designs.

Molecular dynamics simulations

All simulations were performed using the Amber ff14SB force field (*19*) implemented in the AMBER 16 package (*20*). To obtain a neutral charge of the simulated systems, several counter-ions were included. Each protein with the counter-ions was solvated with TIP3P water molecules, using an octahedral box with a minimum distance of 12 Å between the solute and the simulation box edges. All systems were first subjected to 7 iterations of 1,000 minimization steps consisting of 500 steps of steepest descent minimization followed by 500 steps of conjugated gradient. A decreasing harmonic restraining potential was applied to the solute heavy atoms during the 6 first minimization iterations using a force constants of 100, 50, 20, 10, 5, and 1 $kcal.mol^{-1}.Å^{-2}$ respectively. Heating of each system (NVT simulation) up to 300 K was carried out during 100 ps

under periodic boundary conditions, with positional restraints applied to the solute heavy atoms using a force constant of 25 kcal.mol^{-1}.Å^{-2}. NVT simulation at the target temperature (300 K) was further conducted during 300 ps in the same conditions. A 200 ps simulation at constant pressure and temperature (NPT) was later performed to equilibrate the pressure of each system around the target value of 1 bar. During this step, a weak positional restraint was applied on the solute heavy atoms, using a force constant of 5 kcal.mol^{-1}.Å^{-2}. An unrestrained MD simulation of 100 ns was finally performed under the same conditions. The Berendsen algorithm (*21*) was used to keep both temperature and pressure constant during simulations. A cutoff of 9 Å was used to define long-range electrostatic interactions, which were calculated using the Particle Mesh Ewald algorithm (*22*). Bonds involving hydrogen atoms were constrained using the SHAKE algorithm (*23*) to enable the use of a 2 fs time step. All trajectory analyses were carried out with the CPPTRAJ module (*24*).

Construction of expression vectors

The synthetic genes encoding the proteins used in this study were purchased (Thermo Fisher Scientific, MA, USA). After amplifying the genes by PCR, each product DNA fragment was cloned using In-Fusion™ (TAKARA Bio, Japan) into the pET47b vector, to add a cleavable N-terminal His_6-tag to the sequences. When sub-cloning the genes for a halved fragment, "XXX_half" and "Cloning_upstream" primers were used in the PCR amplification. The DNA sequences used in this study are listed in Supplementary Table S11.

Protein expression and purification

To produce the proteins, competent *E. coli* BL21-Gold(DE3) cells (Agilent Technologies, CA) were transformed with the respective expression vectors. The transformants were cultured at 37˚C

overnight in 20 mL of Luria Broth medium supplemented with 20 μg/mL kanamycin. The cells were then inoculated into 2 L of Luria Broth medium and cultured at 37˚C for 2 hours. For induction, 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to the media, and the desired proteins were expressed for 4 hours under the same conditions. After harvesting the cells, the pellets were resuspended in 60 mL of 50 mM potassium phosphate buffer, pH 6.5, 150 mM NaCl, and sonicated. The bacterial lysate was fractionated into supernatant and precipitant by centrifugation at 8,000 rpm and 4 ˚C, for 20 min. To precipitate the contaminating *E. coli* proteins, the supernatants were incubated at 70 ˚C for 20 min, and then each soluble fraction was isolated by centrifugation at 8,000 rpm and 4 ˚C for 20 min. In the cases of mk2h_ΔMILPS, mk2h_ΔMILYS, mk2h_ΔMILPY, and mk2h_ΔMILPYS, the above heat treatment process was omitted in order to preserve their native structures. The soluble proteins were purified by HisTrap HP nickel affinity chromatography (GE Healthcare, IL). The N-terminal His_6-tags were cleaved with HRV-3c protease (Funakoshi, Japan) at 4˚C for 1–2 days. To remove the cleaved His_6-tag and residual uncleaved proteins, the treated solutions were loaded onto HisTrap columns, and each flow-through fraction was recovered. The protein solutions were additionally loaded onto a HiLoad 16/600 Superdex75 (GE Healthcare, IL) size exclusion chromatography column, equilibrated with 50 mM potassium phosphate buffer, pH 6.0, 150 mM NaCl, and the peak fractions were collected. The purity of each protein was verified by SDS-PAGE, and the protein concentrations were determined using their A_{280} values, measured with an ultraviolet spectrophotometer (NanoDrop, Thermo Scientific).

For the preparation of seleno-methionine (Se-Met) substituted taVCP_DPBB, *E. coli* BL21-Gold(DE3) cells were grown in 2 L of M9 minimal medium containing 20 μg/L of kanamycin at 37 ˚C until they reached an absorbance at 600 nm (A_{600}) of 0.4. An amino acid mixture (50 mg/L isoleucine, leucine and valine and 100 mg/l phenylalanine, threonine and lysine) and seleno-L-

methionine (60 mg/mL) were then added to the culture, and the cells were grown at 37˚C. After reaching an A_{600} of 0.8, the protein expression was induced with 0.5 mM IPTG, and the cells were grown further at 37˚C for 5 hours.

Biophysical characterization

For the gel filtration analysis, the concentrations of full-length proteins and halved fragments were adjusted to 20 μM and 40 μM, respectively. A 100 μl aliquot of each purified protein was applied to a Superdex 75 Increase 10/300 size exclusion chromatography column, equilibrated with 50 mM potassium phosphate buffer, pH 6.0, 150 mM NaCl, and run on an AKTA FPLC (Amersham Biosciences) at a flow rate of 0.75 mL/min.

CD spectra were collected on a JASCO J820 circular dichroism spectrometer (JASCO, Japan). Samples containing 20 μM full-length proteins or 40 μM halved fragments, in 50 mM potassium phosphate, pH 6.0, 150 mM NaCl, were loaded into a 1 mm pathlength quartz cuvette. Spectra were recorded in the wavelength range from 200 to 250 nm at 1 nm intervals at 25˚C, and each spectrum was the average of 10 scans. Spectra for mk2h mutants containing 7, 8, or 10 amino acid repertoires were recorded at 10 ˚C.

The melting curves were collected on a JASCO J820 CD spectrometer monitored at 222 nm, in 50 mM potassium phosphate buffer, pH 6.0, 150 mM NaCl. The temperature was increased at a rate of 1.0 °C/min. Data points were collected at 0.1 °C increments from 35 °C to 95 °C, or from 10 °C to 95 °C.

Crystallography

For crystallization, all purified protein solutions were dialyzed against 20 mM Bis-tris HCl, pH 6.0, 150 mM NaCl and then concentrated to 30–70 mg/ml. Crystallization screenings were carried

out in 96-well sitting-drop vapor-diffusion plates. Sample solutions (200 nL) were mixed with an equal amount of reservoir solutions and incubated at 20°C. Almost all crystals were obtained after a few hours to a few days. The crystals were cryo-cooled using the reservoir solution including 13–30% glycerol as a cryo-protectant (Table S12).

Data were collected at the Photon Factory (Tsukuba, Japan)(*25*, *26*), SPring 8 (Harima, Japan)(*27–32*), or Swiss Light Source (Villigen, Swiss). The beam lines and detectors are listed in Table S9. The X-ray diffraction data were processed with XDS (*33*). All structures were solved and refined with the program PHENIX (*34*, *35*). The structures of taVCP_DPBB, mkVCP_DPBB, and apVCP_DPBB were solved by the SAD phasing method, using phenix.autosol. The initial structure models for other mutants were determined by the MR phasing method, using phenix.phaser-MR and the crystal structures of DPBB determined in this study as the search models. The model structures were updated manually using Coot (*36*) and iteratively refined with Phenix.refine. Statistics for diffraction data collection and refinement are summarized in Table S2.

Characterization of chemically-synthesized peptides

All chemically-synthesized peptides tested in this report were obtained from Japan Bio Services. The synthesized mk2h (95.14% purity) and mk2h_ΔMILPS (95.56%) peptides in powdered forms were dissolved in 20 mM Bis-tris HCl, pH 6.0, 150 mM NaCl. The concentrations were determined by the absorbance at 280 nm, using an ultraviolet spectrophotometer (NanoDrop One, Thermo Scientific). For the crystallization screenings, the mk2h_ΔMILPYS peptide (95.85%) was dissolved in 20 mM Tris-HCl, pH 8.5, 200 mM lithium sulfate. The dissolved peptide sample was separated into supernatant and precipitate fractions, and then the precipitate was resuspended in the same buffer. Both the supernatant and undissolved suspension were used for the crystallization screenings, which were carried out in 96-well sitting-drop vapor-diffusion

plates. The obtained crystals were processed in the same way as the bacteria-produced proteins described above.

To examine the foldability in the presence of the contaminated sequences, the low purity peptides of mk1h (75.06%) and mk2h (71.16%) were dissolved in 20 mM Bis-tris HCl, pH 6.0, 150 mM NaCl. The aggregated and dimer species were separated on a size exclusion chromatography column (Superdex 75 Increase 10/300) equilibrated with 20 mM Bis-tris HCl, pH 6.0, 150 mM NaCl. The peptide concentrations of three samples, 1) before SEC purification, 2) aggregated fraction, and 3) dimer fraction, were adjusted to 20 μM and 10 μM for CD spectrum measurements and liquid chromatography mass spectrometry, respectively. Using a JASCO J820 circular dichroism spectrometer (JASCO, Japan), CD spectra were recorded at wavelengths from 200 to 250 nm at 1 nm intervals at 25 ˚C, and each spectrum was the average of 10 scans.

LC-MS analysis

The peptide solutions were desalted with in-house made C18 stage-tips, dried under a vacuum, and dissolved in 2% acetonitrile and 0.1% formic acid. The 10 pmol peptide mixtures were fractionated by C18 reverse-phase chromatography (1.8 μm, ID 0.075 mm x 250 mm, Aurora UHPLC Column; IonOpticks, ADVANCE UHPLC; AMR Inc.) and applied directly into a hybrid linear ion trap mass spectrometer (LTQ Orbitrap Velos Pro; Thermo Fisher Scientific). The peptides were eluted at a flow rate of 200 nL/min with a linear gradient of 5–35% solvent B over 20 min. The compositions of Solvent A and B were 0.1% TFA in water and 100% acetonitrile, respectively.

The Orbitrap mass spectrometer was programmed to carry out 4 successive scans, with the first consisting of a full MS scan from 350–2,000 m/z at a resolution of 60,000, and the second to fourth consisting of data‐dependent scans of the top three most abundant ions obtained in the first scan,

at a resolution of 7,500. Automatic MS/MS spectra were obtained from the highest peak in each scan, by setting the relative collision energy to 35% and the exclusion time to 90 s for molecules in the same m/z value range. Calculations of peptide masses and intensities in the time range from 15.0–35.0 min were performed with the Xtract deconvolution algorithm in FreeStyle, version 1.5 (Thermo Fisher Scientific).

Supplementary Text

SC-design

Most of the amino acid residues conserved between the N- and C-halves in each VCP_DPBB have probably remained unchanged from their perfectly symmetric ancestor, as the chance of having the same amino acid residues in the symmetric positions by random mutation is low. This idea is supported by the observation that the symmetrically-conserved residues in the DPBB domain from each archaeon are often shared with other species (Fig. S4). Studies have also shown that *M. kandleri*, the organism possessing one of the highest symmetric extant DPBBs, is close to the phylogenetic root of archaea (*37*–*39*).

We used mkVCP_DPBB and apVCP_DPBB, with the highest symmetrical sequence identities among the DPBB domains in our database, as the starting templates to create perfectly symmetrical DPBBs. Initially, the symmetrically-conserved residues in mkVCP_DPBB or apVCP_DPBB were introduced into each other to construct the chimeric DPBBs, mkDPBB_sym_67 (internal sequence identity: 67%) and apDPBB_sym_63 (63%) (Table S1). We confirmed that both proteins were monomers with an α/β structure (Fig. S5A and C). We then introduced the symmetrically-conserved residues in the other eight VCP_DPBB sequences top-ranked by internal sequence identities (Fig. S4 and Table S3) into mkDPBB_sym_67 and apDPBB_sym_63, resulting in mkDPBB_sym_81 (81%) and apDPBB_sym_79 (79%). Additionally, the symmetrically-conserved residues in ten more VCP_DPBB sequences were introduced, to construct mkDPBB_sym_86 (86%) and apDPBB_sym_84 (84%). Three proteins, except for mkDPBB_sym_81, were verified to be folded (Fig. S5B, D, and E). X-ray-crystallography confirmed that mkDPBB_sym_86 and apDPBB_sym_79 adopt the DPBB fold (Figs. 1E, S6A). The mkDPBB_sym_86 and apDPBB_sym_84 proteins have only a limited number of non-

symmetrized positions, which are clustered in two areas. In mkDPBB_sym_86, cluster-1 comprises 4L, 7K, 8L, 28A, 29S, and 64G around α-helix 1 and β-strand 1', and cluster-2 comprises 21R, 47K, 50V, 51A, 71Y, and 72E around α-helix 1' and β-strand 1 (Fig. 1E). Looking at the overall structure of the original full-length VCP containing the other domains, these symmetrical faces in the DPBB domain are in different environments: the residues at cluster-1 are exposed to the solvent, while the residues at cluster-2 contact another domain (Fig. S7). This difference in the molecular environments probably led to a breakdown of the symmetry in this area during the evolutionary process.

Subsequently, we designed four DPBB domains with perfect internal sequence symmetries (mkDPBB_sym1, mkDPBB_sym2, apDPBB_sym1, and apDPBB_sym2) by adopting the amino acid residues from either cluster-1 or -2. While apDPBB_sym2 could not be purified due to its poor stability, mkDPBB_sym1, mkDPBB_sym2, and apDPBB_sym1 were purified as stable proteins (Fig. S5F–H). We determined the crystal structures of mkDPBB_sym1 and mkDPBB_sym2 to confirm that they adopt the DPBB fold as designed (Fig. 1F, Table S2). Therefore, we succeeded in reconstructing the perfectly symmetrical DPBB structures, using the sequence information of VCP-like chaperones from only twenty archaeal species. This result led us to anticipate that the archaeal common ancestor possessed a nearly perfect symmetric DPBB sequence in the VCP-like chaperone gene. Furthermore, all of the perfectly symmetrical DPBB and intermediate mutants constructed by the SC-design exhibited high stability. Except for mkDPBB_sym_67 (T_m=69.2 °C), they did not completely unfold even at 95 °C (Fig. S5). These results support the hypothesis that the common ancestor of archaea is a thermophile (*40*).

RE-design

We utilized a modified "reverse engineering evolution" protein design approach to design the symmetrical DPBB sequences. In this design methodology, orthologous sequences of the target protein are used instead of pseudo-symmetric sequences for the construction of phylogenetic trees. Specifically, we used VCP_DPBB sequences from different organisms to construct a phylogenetic tree with respective aligned sequences, which were subsequently used as input to generate ancestral sequences (Figure S8A and B). The predicted ancestral sequences were mapped onto a manually constructed, perfectly symmetrical mkVCP_DPBB structural backbone model using an in-house program that utilizes PyRosetta, and each sequence was ranked by the Rosetta score (Figure S8C and D). The top-scored designs were analyzed for Rosetta total score, RMSD from the manually generated, symmetrical mkVCP_DPBB model and through visual inspection. First, our analysis of the Rosetta total score revealed that many output models showed significantly lower energy and converged well (Figure S9A). When the backbone RMSDs of the designs were computed and plotted against the Rosetta total score, a broad spread of total score/RMSD scores was obtained. However, the majority of the top-scored output models exhibited RMSDs <1 Å, suggesting even with diverse sequences, they did not deviate much from the starting structure (Figure S9B). Interestingly, several of the top-scored models exhibited RMSDs <0.7 Å. The reDPBB_sym1 and reDPBB_sym2 exhibited 0.65 Å and 0.64 Å RMSD from the manually generated symmetrical mkVCP_DPBB model, respectively. Next, an analysis of Rosetta total score versus percentage sequence identity revealed that the top-scored structures tend to have higher percentage sequence identity (Figure S9C). However, the shortlisted reDPBB_sym1 and reDPBB_sym2 designs shared a moderate 78% and 74% sequence identity with the manually generated, symmetrical mkVCP_DPBB model respectively.

We compared the Rosetta generated structural models to that of the crystal structures of reDPBB_sym1 and reDPBB_sym2. First, we found that in both reDPBB_sym1 and

reDPBB_sym2 crystal structures, each half of the proteins exhibit 0.41 Å RMSD with each other (Figures S9D–G). Moreover, while the crystal structure of reDPBB_sym1 and the Rosetta generated model of reDPBB_sym1 share 0.66 Å RMSD, the reDPBB_sym2 crystal structure and Rosetta generated model of reDPBB_sym2 share 0.58 Å RMSD (Figures S9H and S9I). This indicated that the Rosetta- generated structural models are in close agreement with the crystal structures. This result demonstrated that our computational symmetric design approach can be successfully applied to the design of DPBB fold as well.

Our designed reDPBB_sym1 and reDPBB_sym2 sequences share 67% and 65% sequence identity with mkVCP_DPBB, and 44% identity with taVCP_DPBB respectively. This further confirms that diverse sequences can fold into DPBB structures, as predicted from our computational symmetric protein design approach. Our computational strategy, of using orthologous sequences to construct phylogenetic trees for the design of symmetric proteins, has advantages over that of using pseudo-symmetric sequences from each subunit of the target protein. It circumvents the challenges when the number of subunits in the target protein is very small, such as the two subunits in DPBB. The successful computational design of symmetric DPBB proteins has verified the applicability of our design strategy under this circumstance.

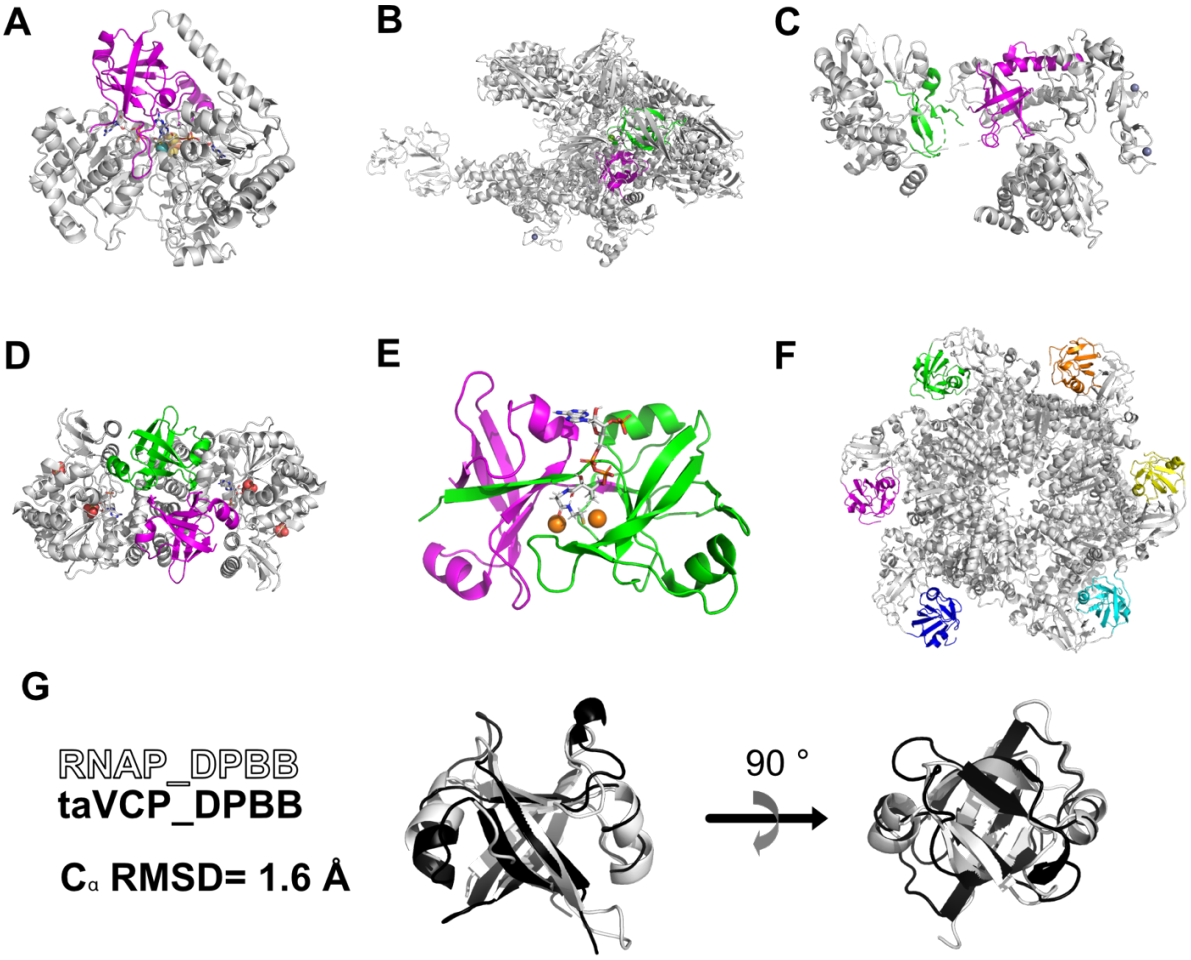

Figure S1. DPBB domains in natural proteins. Structures of (A) formate dehydrogenase (PDB ID 1FDO), (B) RNA polymerase (PDB ID 6ASG), (C) D-type DNA polymerase (PDB ID 5IJL), (D) AMP phosphorylase (PDB ID 4GA6), (E) phosphate propanoyltransferase (PDB ID 5CUO), and (F) molecular chaperone VCP (PDB ID 5G4F) are shown. Each DPBB domain is colored differently. (G) The superimposed structures of the DPBB domain-2 from RNA polymerase (colored pink in Fig. S1B) and the isolated DPBB domain of molecular chaperone VCP (Fig. 1B). The Cα RMSD was calculated by CRICK (*41*, *42*) (http://cospi.iiserpune.ac.in/click/).

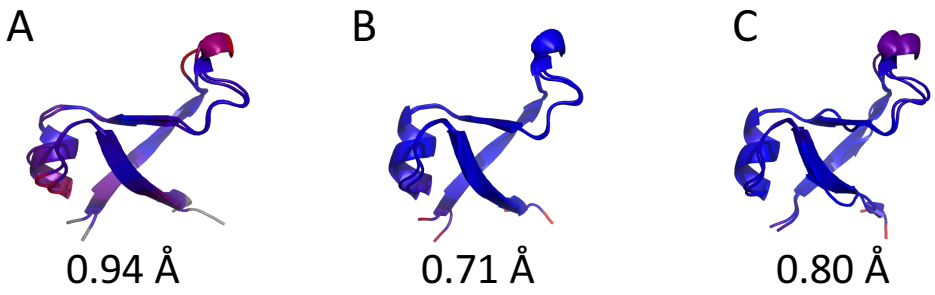

Figure S2. Superimposed structures of the N- and C-terminal halves of extant VCP_DPBBs. Superimposition of the N- and C-terminal halves of the crystal structures of (A) taVCP_DPBB, (B) mkVCP_DPBB, and (C) apVCP_DPBB. Darker blue indicates better alignments in the structures. The values are the Cα RMSDs calculated by CRICK (*41*, *42*) (http://cospi.iiserpune.ac.in/click/).

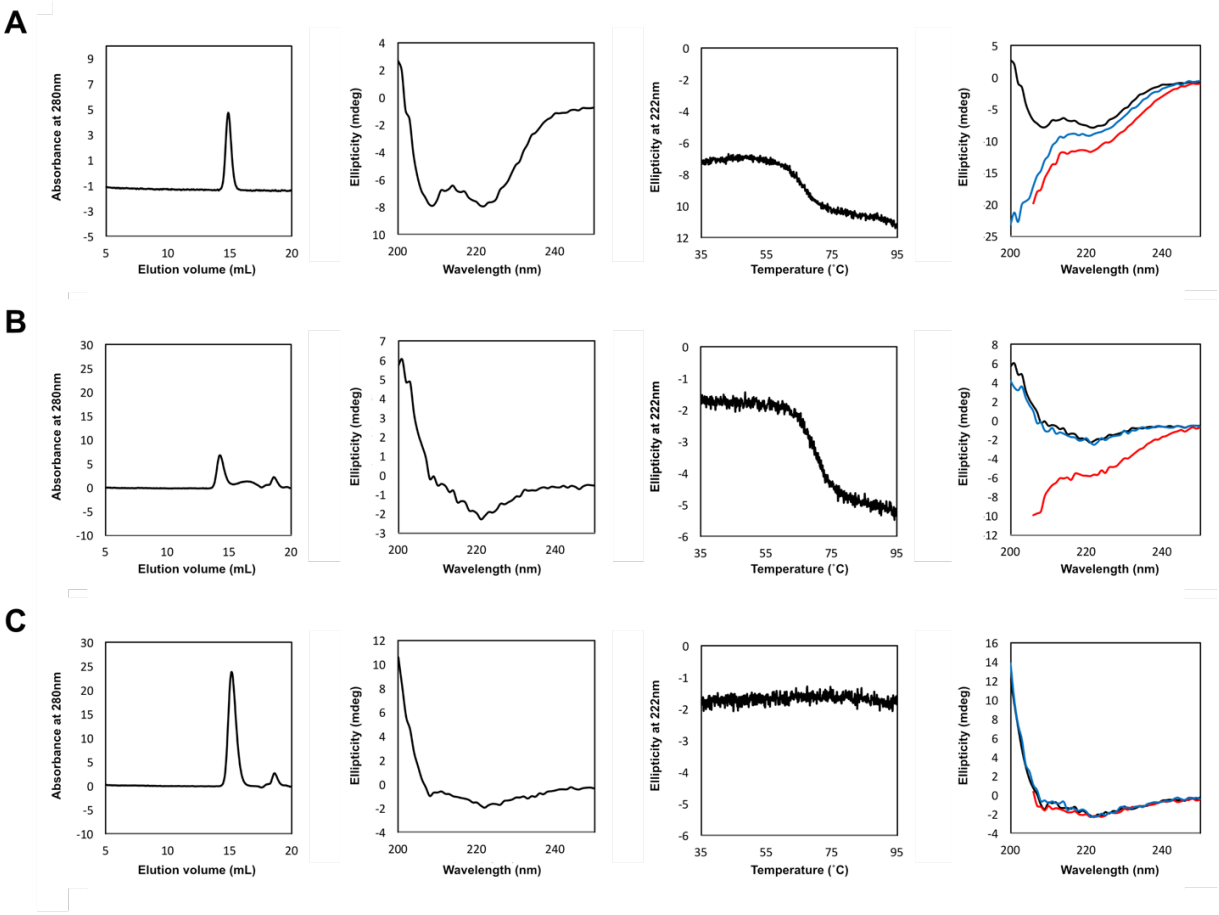

Figure S3. Experimental characterization of extant VCP_DPBBs. Size exclusion chromatography, CD spectra, denaturation curves, and comparisons of CD spectra at different temperatures (black: 35ºC; red: 95ºC; blue (refolding): 95ºC → 35ºC) for (A) taVCP_DPBB, (B) mkVCP_DPBB, and (C) apVCP_DPBB are shown in the panels from left to right.

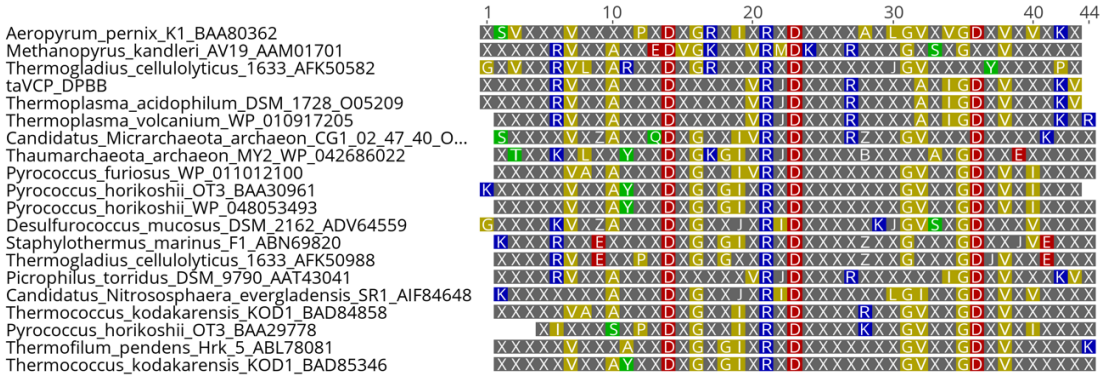

Figure S4. The symmetrically-conserved residues in natural VCP_DPBB domains. The symmetrically-conserved residues between the N- and C-halves in each 20 top-ranked VCP_DPBB with high internal sequence identity are highlighted. The non-symmetrically-conserved residues are represented as X (gray). The sequences are ordered by the internal sequence identity shown in Table S9. To create mkDPBB_sym_81 and apDPBB_sym_79, the symmetrically-conserved residues in the top ten sequences were considered. To create mkDPBB_sym_86 and apDPBB_sym_84, the symmetrically-conserved residues of all sequences were considered.

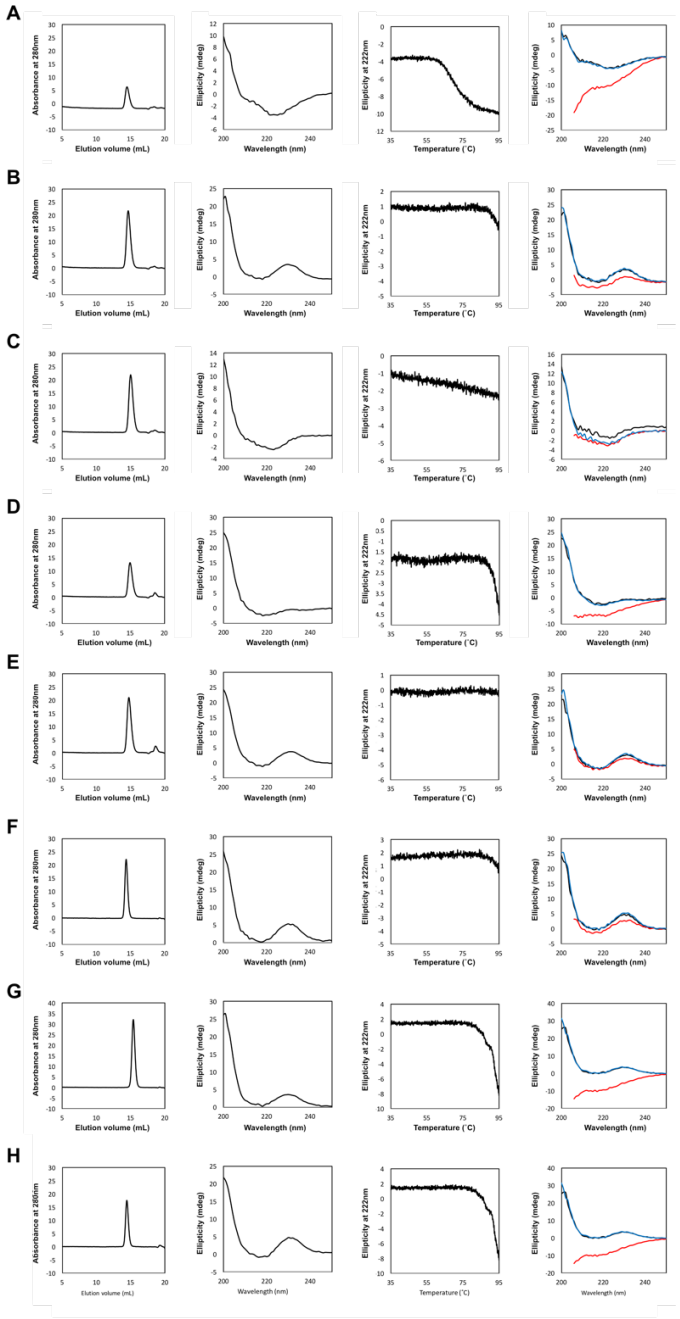

Figure S5. Experimental characterization of symmetrical DPBBs constructed by the SC-design method. Size exclusion chromatography, CD spectra, denaturation curves, and comparisons of CD spectra at different temperatures (black: 35ºC; red: 95ºC; blue (refolding): 95ºC → 35ºC) for (A) mkDPBB_sym_67, (B) mkDPBB_sym_86, (C) apDPBB_sym_63, (D)

apDPBB_sym79, (E) apDPBB_sym_84, (F) mkDPBB_sym1, (G) mkDPBB_sym2, and (H) apDPBB_sym1 are shown in the panels from left to right.

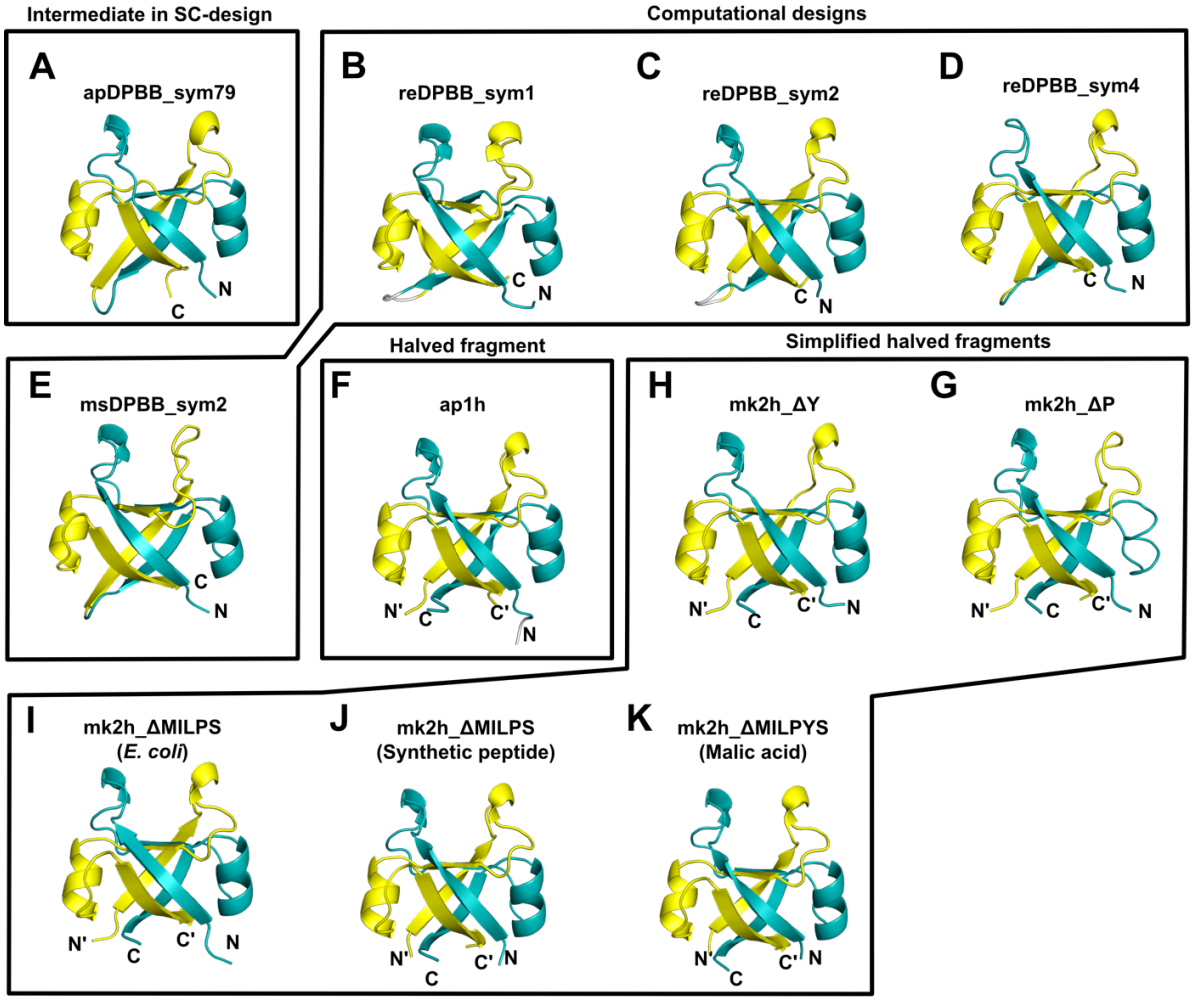

Figure S6. Crystal structures of the designed DPBB domains. (A) apDPBB_sym_79 (PDB ID 7DI0), (B) reDPBB_sym1 (7DVC), (C) reDPBB_sym2 (7DVF), (D), reDPBB_sym4 (7DVH), (E) msDPBB_sym2 (7DWW), (F) ap1h (7DXS), (G) mk2h_ΔP (7DXU), (H) mk2h_ΔY (7DXV), (I) *E. coli*-produced mk2h_ΔMILPS (7DXX), (J) chemically-synthesized mk2h_ΔMILPS (7DXY), and (K) chemically-synthesized mk2h_ΔMILPYS in the presence of DL-malic acid (7DYC).

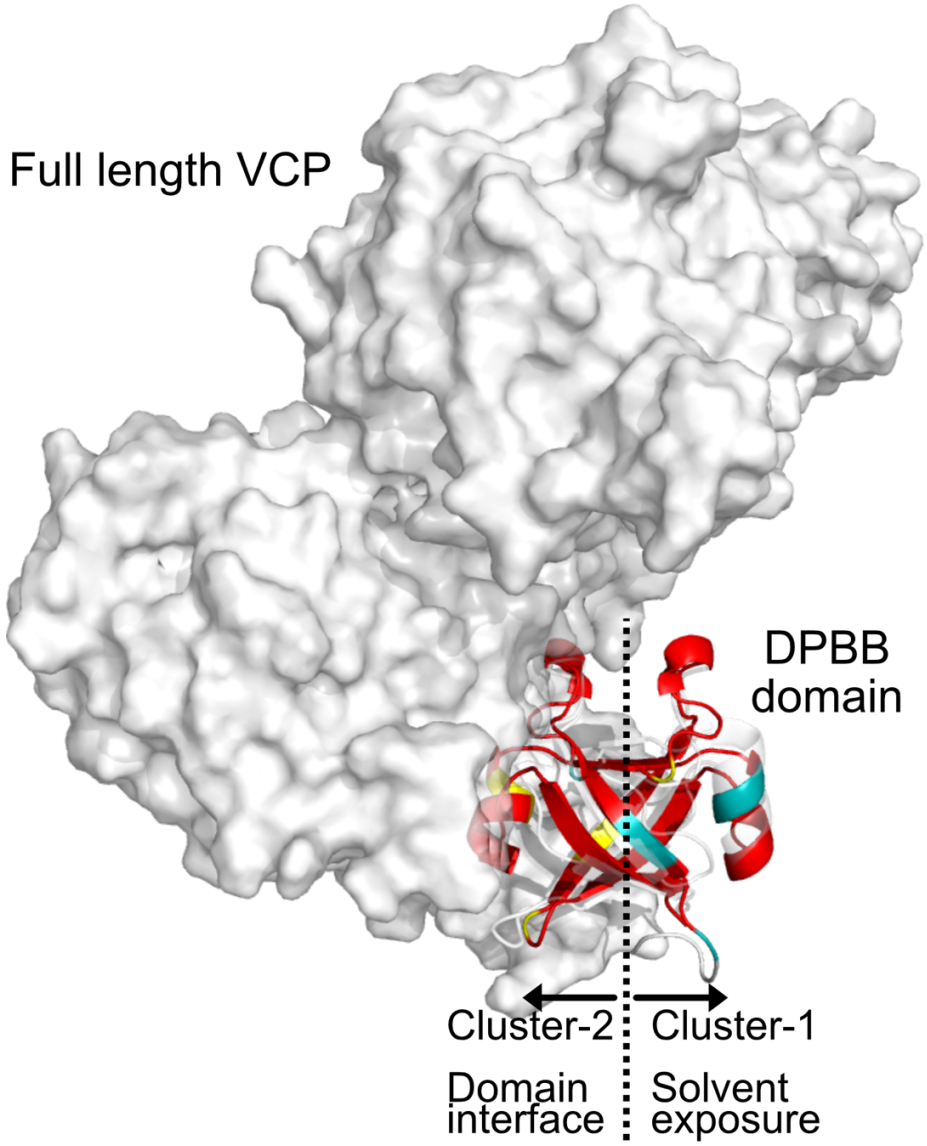

Figure S7. Different environments of symmetrical faces in the full-length VCP protein. The DPBB domain and other domains in the full-length VCP from *Thermoplasma acidophilum* (PDBID 5G4F) are represented by the white cartoon model and surface model, respectively. The crystal structure of mkDPBB_84 is superimposed with the DPBB domain and colored as in Fig. 1E. Cluster-1 is exposed to the solvent and cluster-2 is an interface to the other domain.

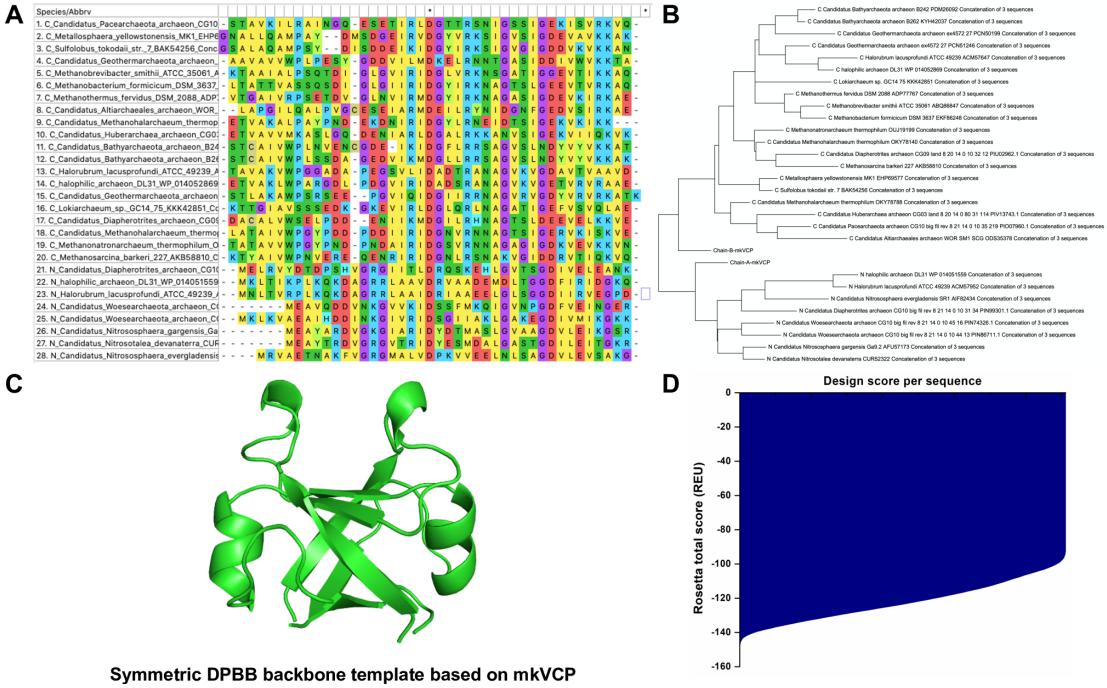

Figure S8. Reverse engineering evolution strategy to design a fully symmetric DPBB domain. (A) DPBB sequences from different organisms were aligned (B) to produce a phylogenetic tree, for use as input to generate possible ancestral sequences. (C) The ancestral sequences were mapped onto the symmetric DPBB backbone template based on mkVCP_DPBB and (D) scored using pyRosetta.

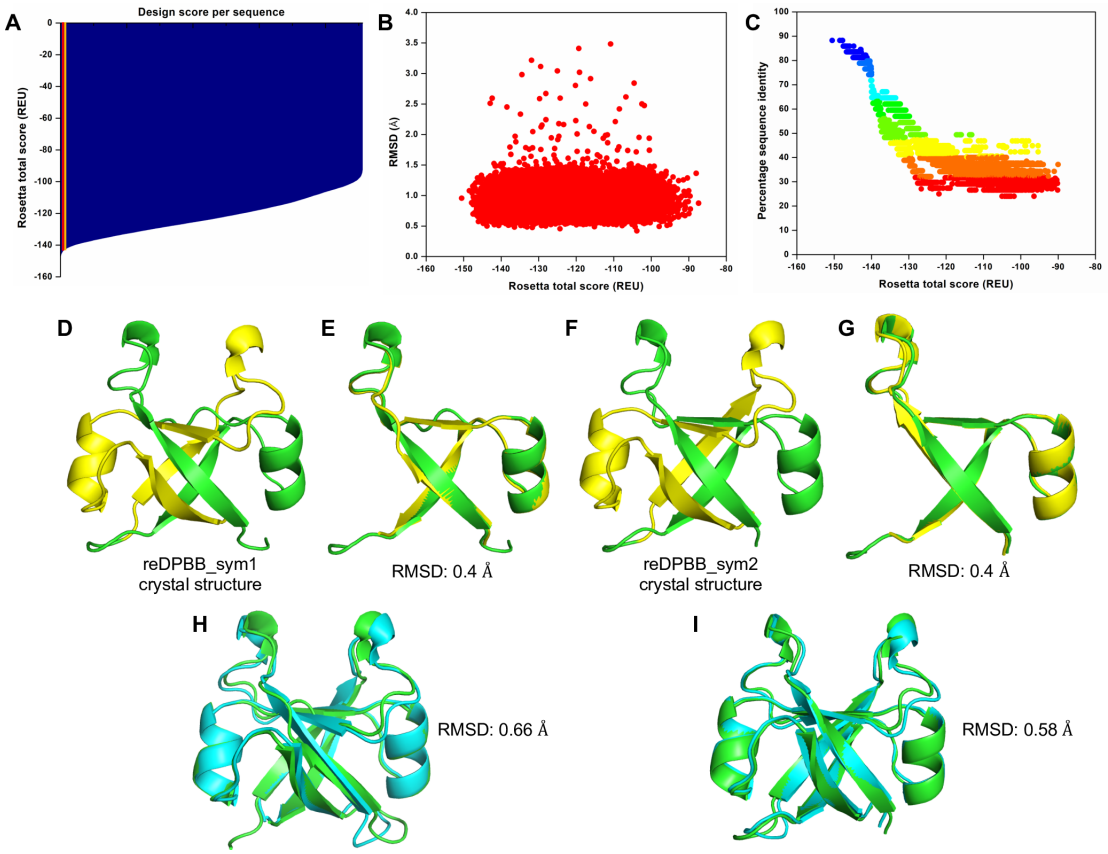

Figure S9. Structural and physicochemical parameters from Rosetta-design calculations, and structural differences between crystal structures and computationally modeled structures of symmetrical DPBBs. (A) Rosetta computed design score per sequence after the ancestral sequences were mapped onto the fully symmetrical DPBB structural model, where the red and yellow lines represent the top-scored reDPBB_sym1 and reDPBB_sym2 designs, respectively. (B) Rosetta scores versus the RMSD of the designs showing well-spread RMSDs from the template. (C) Rosetta total score versus the percentage sequence identity of the designs, which had sequence identities ranging from 25-90%. (D) Crystal structure of reDPBB_sym1. (E) Superimposition and computed RMSD between each half-barrel. (F) Crystal structure of reDPBB_sym2. (G) Superimposition and computed RMSD of each half-barrel, shown in green and blue colors, respectively. Structural superimpositions of (H) the reDPBB_sym1 crystal

structure (green) with the Rosetta generated model (cyan) and (I) the reDPBB_sym2 crystal structure (green) with the Rosetta generated model (cyan), with RMSDs of 0.66 Å and 0.58 Å, respectively.

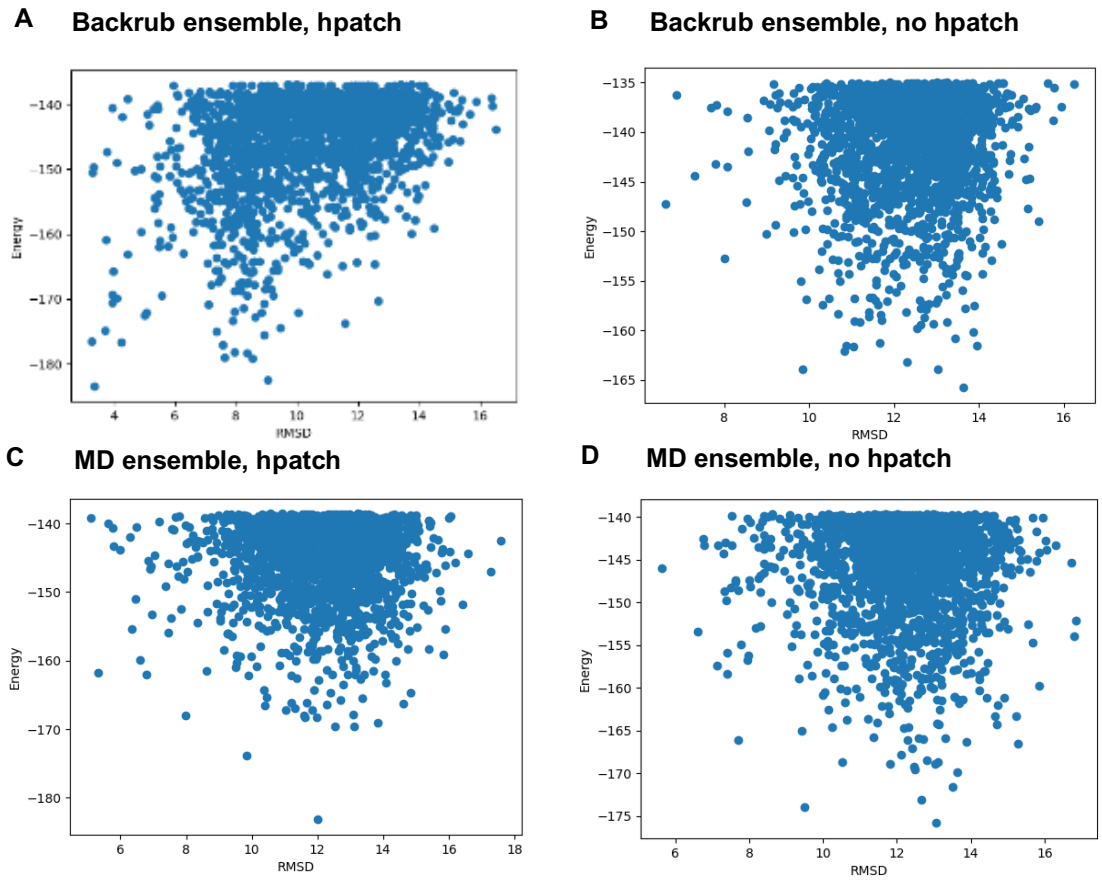

Figure S10. Forward folding profiles of MS designs. For each design, 30,000 protein models were predicted. The RMSD to the template structure was computed for the 1,000 top scoring models. The presence of dots in the bottom left corner of the Backrub ensemble with hpatch plot (A) indicates that EdaRose predicted low energy models within 4 Å from the target structure. For comparison, the Backrub ensemble without hpatch (B) and the MD ensembles with (C) and without (D) hpatch are shown.

A

	mkDPBB_sym1	mkDPBB_sym2	apDPBB_sym1	reDPBB_sym1	reDPBB_sym2	reDPBB_sym3	reDPBB_sym4	reDPBB_sym7	msDPBB_sym1	msDPBB_sym2
mkDPBB_sym1		86	88	80	80	78	77	81	60	58
mkDPBB_sym2	86		84	68	68	66	86	91	65	60
apDPBB_sym1	88	84		71	71	68	74	79	60	58
reDPBB_sym1	80	68	71		95	95	63	66	51	51
reDPBB_sym2	80	68	71	95		90	63	66	51	51
reDPBB_sym3	78	66	68	95	90		61	63	49	49
reDPBB_sym4	77	86	74	63	63	61		91	65	60
reDPBB_sym7	81	91	79	66	66	63	91		65	58
msDPBB_sym1	60	65	60	51	51	49	65	65		88
msDPBB_sym2	58	60	58	51	51	49	60	58	88	

B

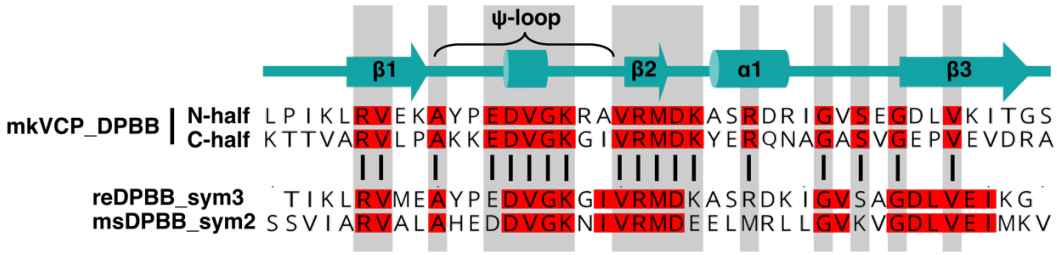

Fig. S11. Sequence diversity in symmetrically designed DPBBs. (A) Sequence identities (%) between each pair of repeat units in the symmetrically designed DPBBs. msDPBB_sym1 and 2 share only 49% sequence identity with reDPBB_sym3. (B) Pairwise sequence alignment of the repeat units in reDPBB_sym3 and msDPBB_sym2 along with the starting template, mkVCP_DPBB. The symmetrically-conserved positions in the mkVCP_DPBB are highlighted in gray. Considering that both designs were created from the same starting model and should be biased toward being highly homologous to the original sequence, the potential diversity of the possible DPBB sequences could still be underestimated. If we compare the residues at the non-

symmetric positions in the starting model, then reDPBB_sym3 and msDPBB_sym2 share only 26% sequence identity.

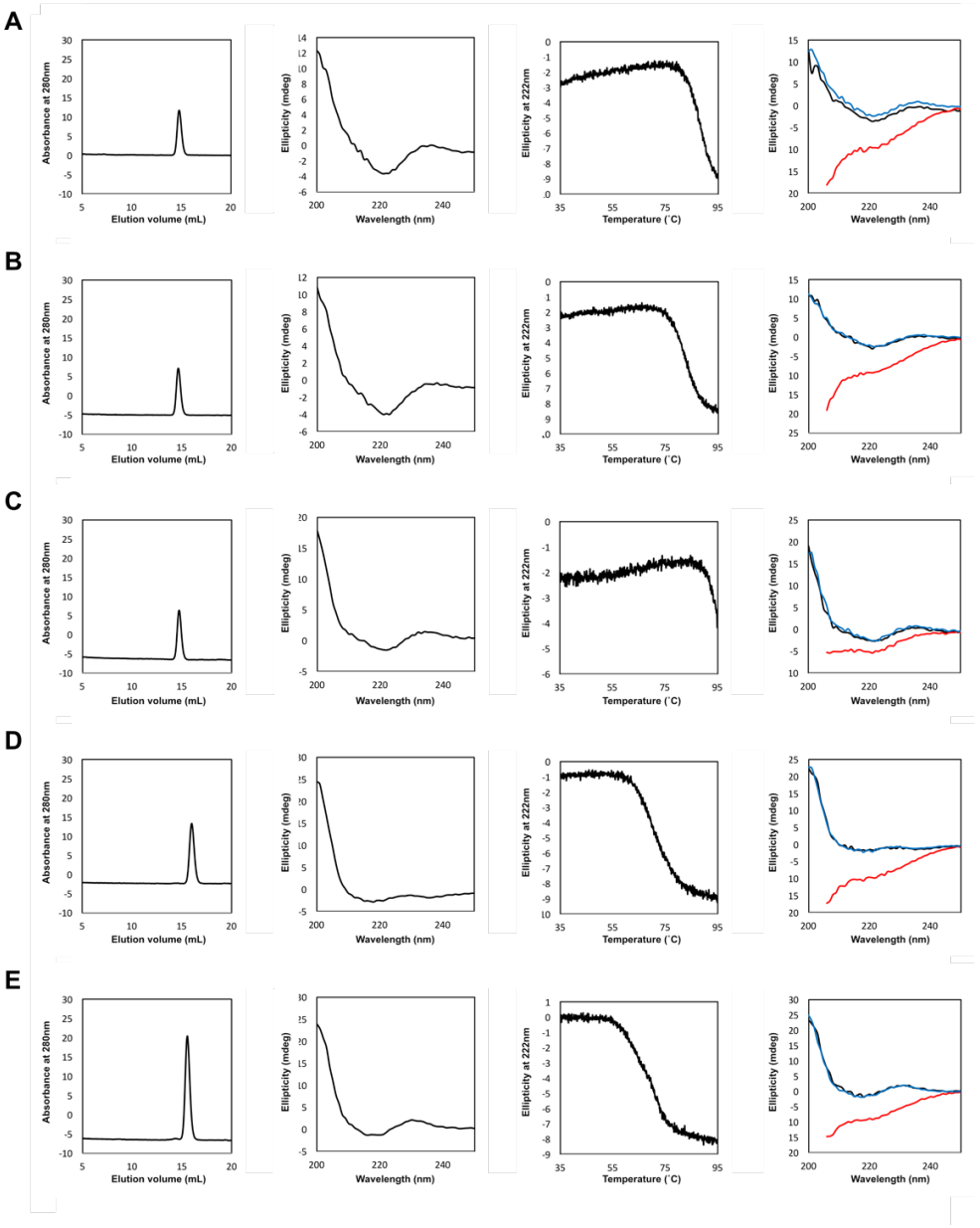

Figure S12. Experimental characterization of symmetrical DPBBs designed by the RE-design method. Size exclusion chromatography, CD spectra, denaturation curves, and comparisons of CD spectra at different temperatures (black: 35ºC; red: 95ºC; blue (refolding): 95ºC → 35ºC) for (A) reDPBB_sym1, (B) reDPBB_sym2, (C) reDPBB_sym3, (D) reDPBB_sym4, and (E) reDPBB_sym7 are shown in the panels from left to right.

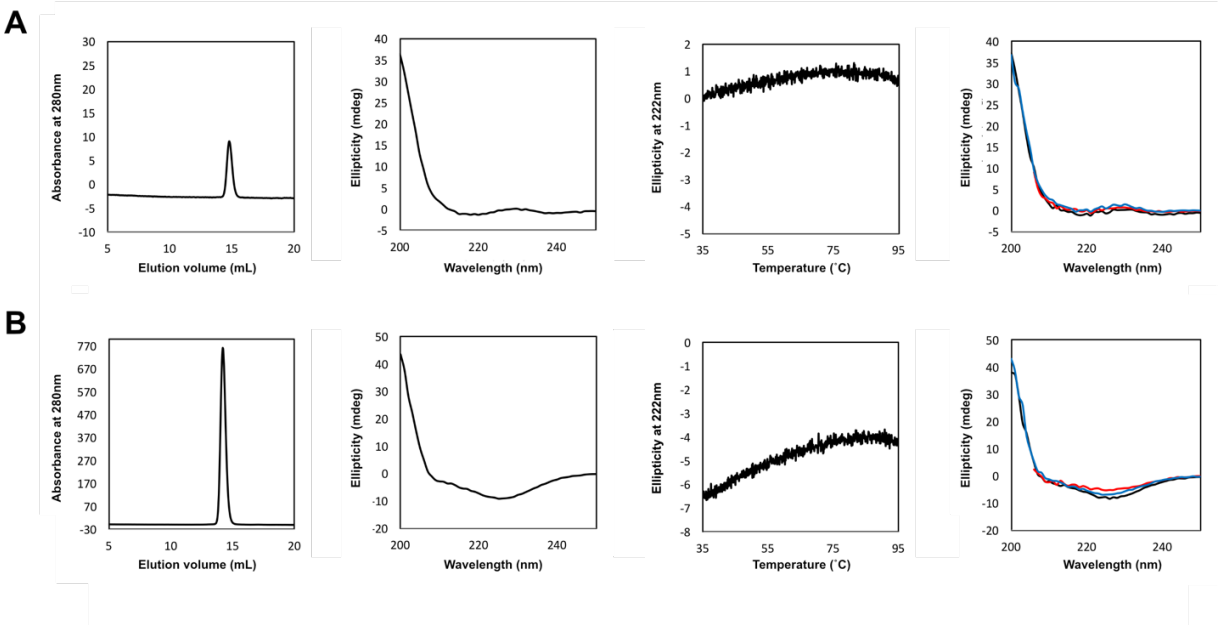

Figure S13. Experimental characterization of symmetrical DPBBs designed by the MS-design method. Size exclusion chromatography, CD spectra, denaturation curves, and comparisons of CD spectra at different temperatures (black: 35ºC; red: 95ºC; blue (refolding): 95ºC → 35ºC) for (A) msDPBB_sym1 and (B) msDPBB_sym2 are shown in the panels from left to right.

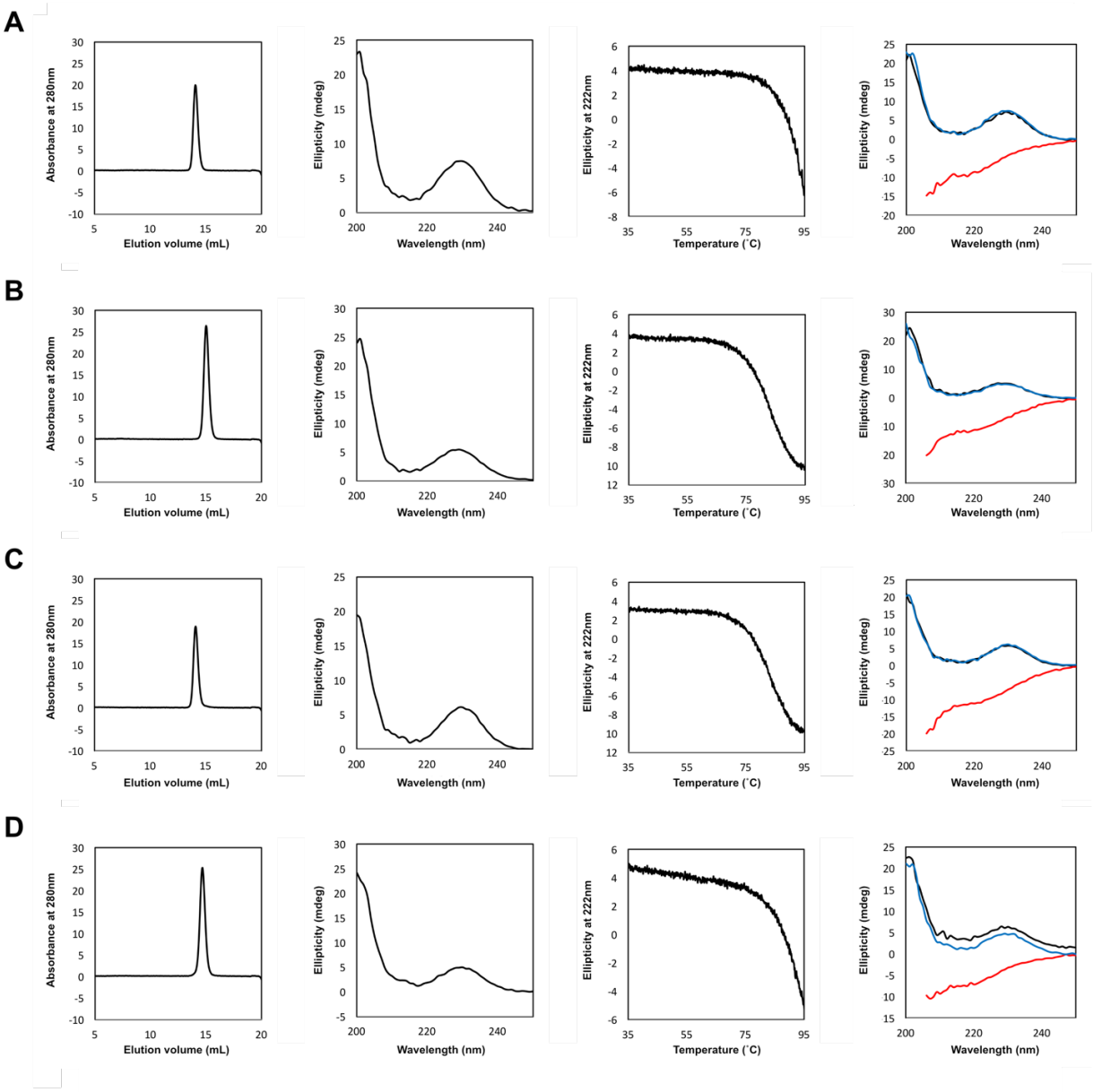

Figure S14. Experimental characterization of the halved fragments. Size exclusion chromatography, CD spectra, denaturation curves, and comparisons of CD spectra at different temperatures (black: 35ºC; red: 95ºC; blue (refolding): 95ºC → 35ºC) for (A) mk1h, (B) mk2h, (C) ap1h, and (D) ap2h are shown in the panels from left to right.

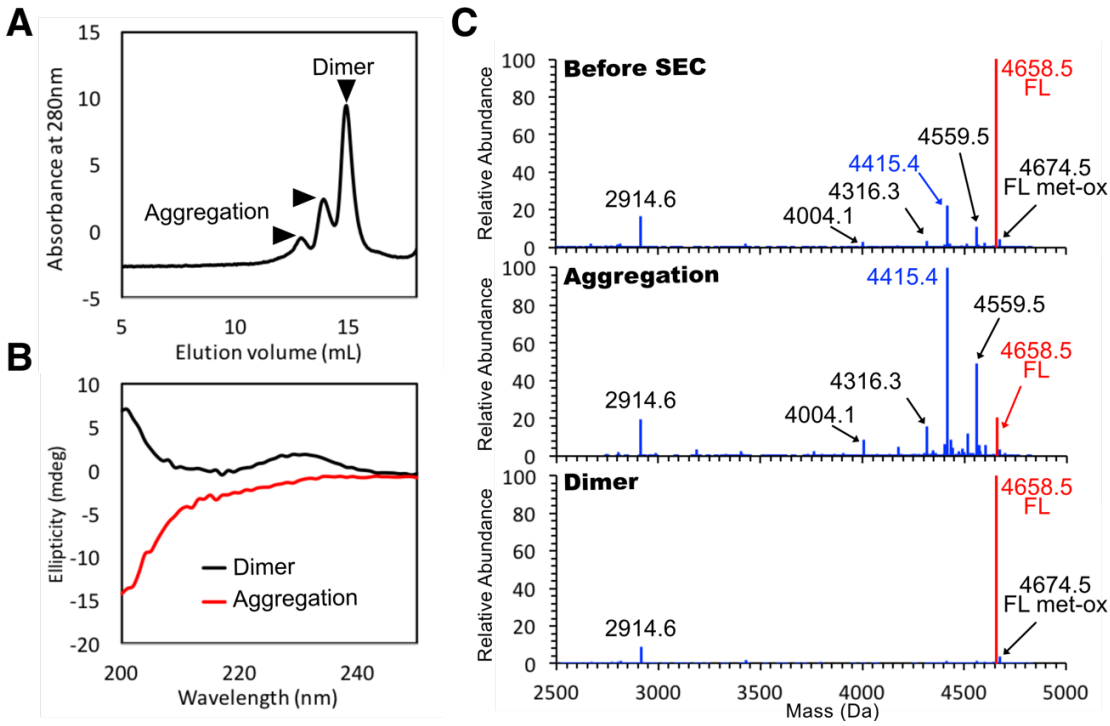

Figure S15. Characterization of the low-purity chemically-synthesized mk1h peptide (75.06%). (A) SEC analysis showing the aggregated and dimeric states of the dissolved mk1h peptide. (B) CD spectra indicating that the aggregated and dimer species separated in Fig. S12B, adopt random-coil and α/β structures, respectively. (C) The peptide species in the sample before and after the SEC purification were analyzed by LC/MS, and the deconvoluted mass spectra are shown. The labels for the full-length mk1h peptide (4658.5 Da) and the major contaminant peptide (4415.4 Da) are highlighted in red and blue, respectively. While most of the contaminant peptides were enriched in the aggregation fraction, the full-length mk1h was enriched in the dimer fraction.

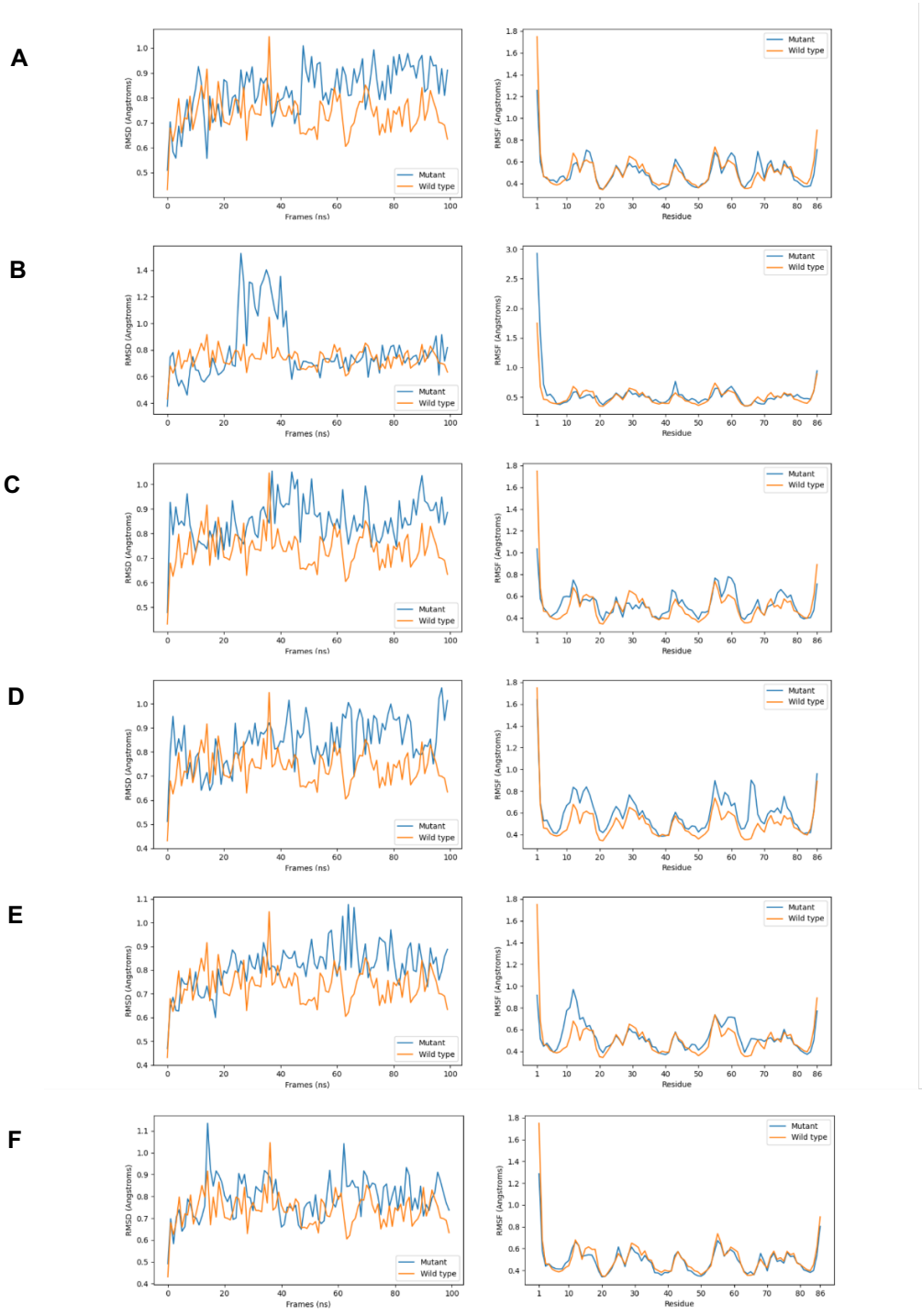

Figure S16. Molecular dynamics simulations of the mk2h_ΔI (A), mk2h_ΔL (B), mk2h_ΔM (C), mk2h_ΔP (D), mk2h_ΔS (E), and mk2h_ΔY (F) mutants. The simulations were performed starting

from structural models in the form of the linked repeats with 86 amino acid residues. The backbone RMSD to the reference structure along 100 ns of simulation is plotted on the left, and the backbone RMSF by residue averaged over 100 ns of simulation is plotted on the right. No large conformational change, except around the N-terminus, was observed in both the original (mk2h) and mutant designs.

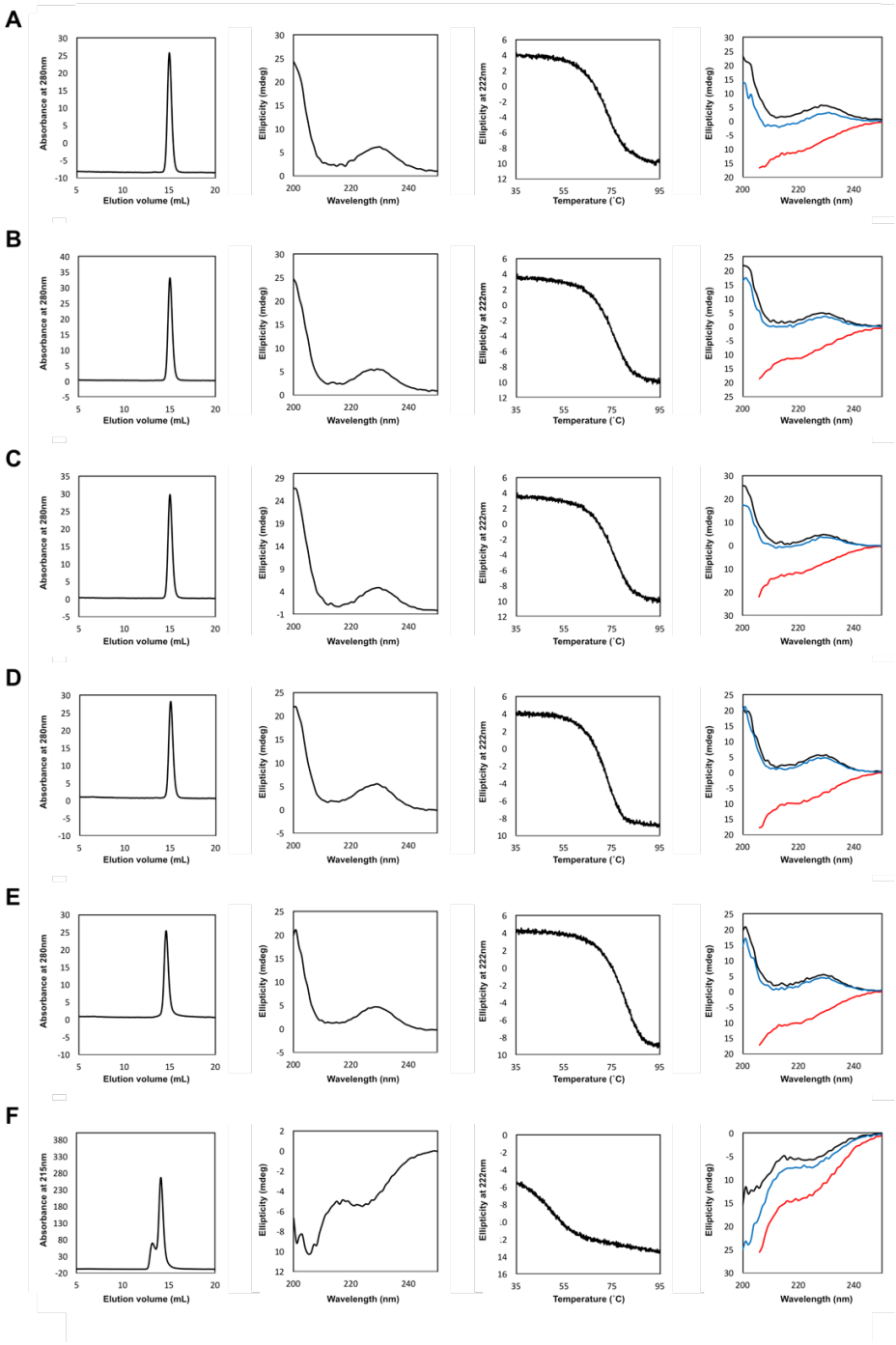

Figure S17. Experimental characterization of mk2h mutants containing 12 amino acid repertoires. Size exclusion chromatography, CD spectra, denaturation curves, and comparisons of CD spectra at different temperatures (black: 35ºC; red: 95ºC; blue (refolding): 95ºC → 35ºC)

for (A) mk2h_ΔM, (B) mk2h_ΔI, (C) mk2h_ΔL, (D) mk2h_ΔP, (E) mk2h_ΔS, and (F) mk2h_ΔY are shown in the panels from left to right.

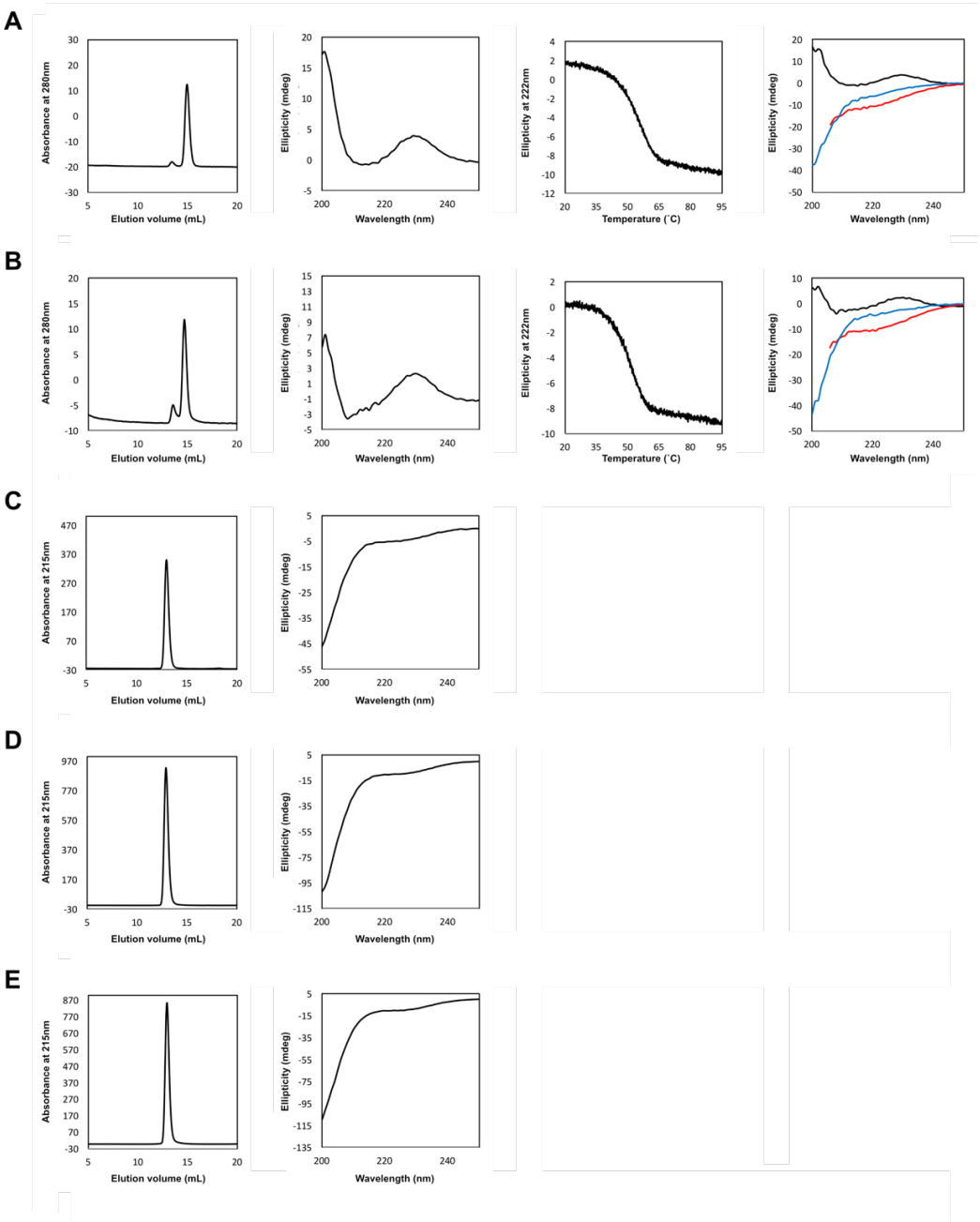

Figure S18. Experimental characterization of mk2h mutants containing 7, 8, or 10 amino acid repertoires. Size exclusion chromatography, CD spectra, denaturation curves, and comparisons of CD spectra at different temperatures (black: 10ºC; red: 95ºC; blue (refolding): 95ºC → 10ºC) for (A) mk2h_ΔMIL, (B) mk2h_ΔMILPS, (C) mk2h_ΔMILPY, (D) mk2h_ΔMILSY, and (E) mk2h_ΔMILPYS are shown in the panels from left to right.

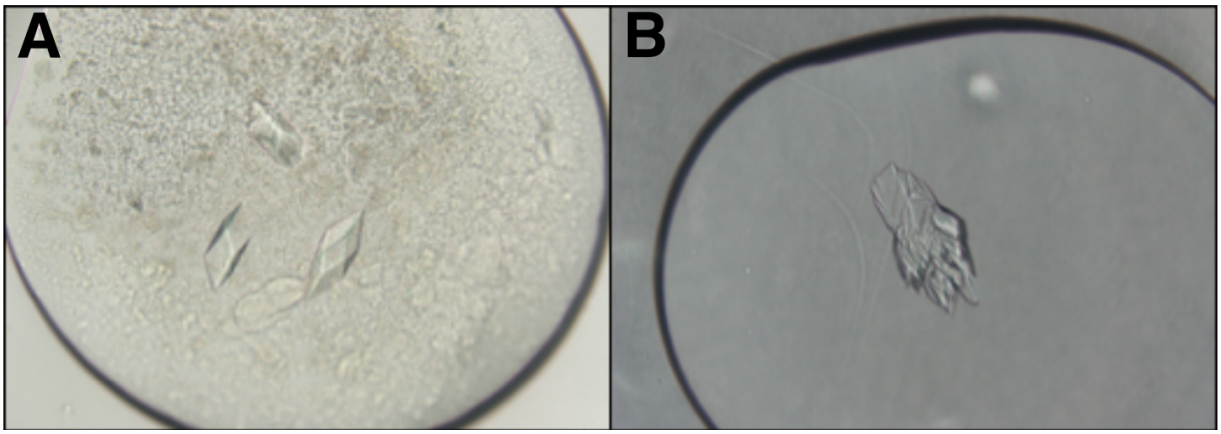

Fig. S19. Crystals of mk2h_ΔMILPYS. The crystals were obtained under two different conditions containing (A) an undissolved suspension of the chemically-synthesized peptide in 3 M sodium malonate and (B) the dissolved peptide in 2.1 M DL-malic acid, pH 7.0. The determined crystal structures are shown in Fig. 4D and Fig. S6K, respectively.

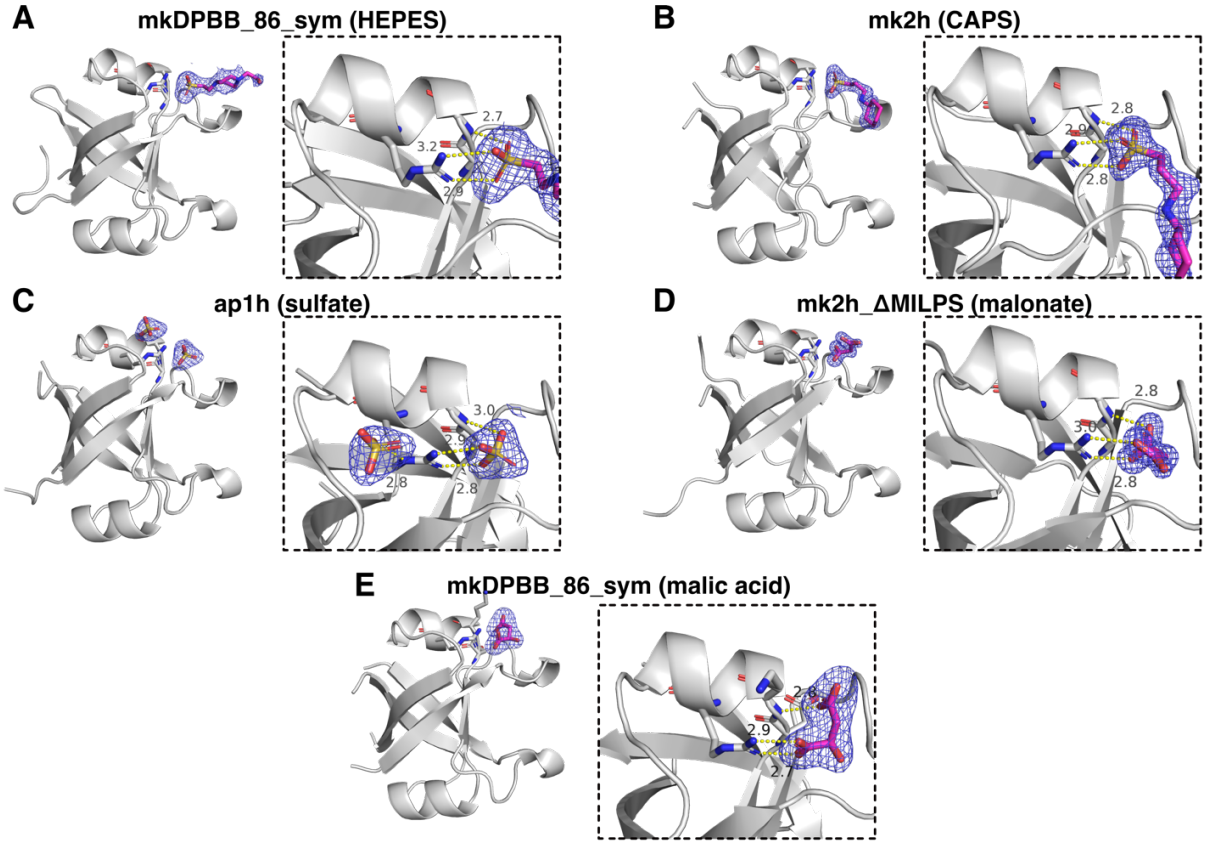

Figure S20. Positively-charged pockets in designed DPBBs occupied by negatively-charged ligands. (A–E) Crystal structures and close-up views of the conserved positively-charged pockets of designed DPBB domains. (A) mkDPBB_sym_84, (B) mk2h, (C) ap1h, (D) mk2h_ΔMILPS, and (E) mk2h_ΔMILPYS interactions with HEPES, CAPS, sulfate ion, malonate ion, and malic acid, respectively, at the positively charged pocket around their α-helices. The N-H group at the N-terminal peptide bond of the α1 helix and the arginine residue positioned in the middle of the α1 helix (Arg30 in mk2h_ΔMILPYS) form salt-bridges with the negatively-charged ligands. This observation supports the idea that the ancestral DPBB proteins composed of the repeated sequence or halved fragments could have functioned as cofactor- or nucleic acid-binding proteins, like their extant descendants (*43*, *44*).

Algorithm: Hpatch in POMP^d.

```
Inputs: C : protein conformation, efunc : energy function network
exp_residues = ∅
exp_neighbors = ∅
C = mutate_all_residues(C,"Leu")
for res in residues(C) do
   if is_exposed(res) then
      exp_residues = exp_residues ⋃ res
   end if
end for
for res in residues(C) do
   current_res_exp_neighbors = ∅
   for all neighbor = neighbors(res) do
      if neighbor ∈ exp_residues then
         current_res_exp_neighbors = current_res_exp_neighbors ⋃ neighbor
      end if
   end for
   exp_neighbors = exp_neighbors ⋃ current_res_exp_neighbors
end for
for res in residues(C) do
   if res ∈ exp_residues then
      for all neighbor ∈ exp_neighbors[res] do
         add_terms(efunc,res,neighbor)
      end for
   end if
end for
```

Algorithm S1. Hpatch procedure. Exposed surface residues are calculated with pyRosetta, after mutating the whole protein to LEU in order to ensure the equal treatment of all residue positions. For each residue, the list of all its exposed neighbors is computed. Finally, the neighboring hydrophobic pairs at the surface are forbidden with new energy terms.

Table S1. Sequences and internal identities of the natural and designed DPBB domains.

Protein name	aa	pI	Mw (kDa)	Sequence	Sequence identity
Isolated DPBBs from VCP proteins					
taVCP_DPBB	91	9.6	10.2	MESNNGIILRVAEANSTDPGMSRVRLDESSRRLLDAEIGDVVEIEKVR KTVGRVYRARPEDENKGIVRIDSVMRNNCGASIGDKVKVRKVR	16/43 (37%)
mkVCP_DPBB	90	9.5	9.8	MPGLPIKLRVEKAYPEDVGKRAVRMDKASRDRIGVSEGDLVKITGS KTTVARVLPAKKEDVGKGIVRMDKYERQNAGASVGEPVEVDRAE	18/43 (42%)
apVCP_DPBB	91	8.1	9.9	MANSSVELRVSEAYPRDVGRKIVRIDRQTAARLGVEVGDFVKVSKGD RSVVAVVWPLRPDDEGRGIIRMDGYLRAALGVTVGDTVTVEKAE	20/44 (45%)
Symmetric DPBBs designed using Symmetric-conservation method (SC-design)					
apDPBB_sym_63	91	9.8	10.0	MANSSVELRVSEAYPEDVGRKIVRMDKQTRARLGVSVGDFVKVSKGD RSVVARVWPARPEDVGRGIVRMDKYLRAALGVSVGDTVTVEKAE	28/44 (63%)
apDPBB_sym_79	91	9.8	10.1	MANSSVELRVLEARPEDVGRKIVRMDKQTRARLGVSVGDYVEVKKVD RSVVARVLPARPEDVGRGIVRMDKYLRAALGVSVGDYVEVKKVE	35/44 (79%)
apDPBB_sym_84	91	9.2	10.1	MANSSVELRVAEAYPEDVGRKIVRMDKQTRAKLGVSVGDYVEVKKVD RSVVARVAEAYPEDVGRGIVRMDKYLRAKLGVSVGDYVEVKKVE	37/44 (84%)
mkDPBB_sym_67	90	9.8	9.7	MPGLSVKLRVEKAYPEDVGKRIVRMDKASRARLGVSVGDLVKVTKS KSVVARVLPAKPEDVGKGIVRMDKYERANLGVSVGDPVEVDKAE	29/43 (67%)
mkDPBB_sym_81	90	10.1	9.9	MPGLSVKLRVLKARPEDVGKRIVRMDKASRARLGVSVGDYVEVKKV KSVVARVLPARPEDVGKGIVRMDKYERANLGVSVGDYVEVKKVE	35/43 (81%)
mkDPBB_sym_86	90	9.5	9.9	MPGLSVKLRVAEAYPEDVGKRIVRMDKASRAKLGVSVGDYVEVKKV KSVVARVAEAYPEDVGKGIVRMDKYERAKLGVSVGDYVEVKKVE	37/43 (86%)
mkDPBB_sym1	89	9.7	9.6	MPGLSVKLRVAEAYPEDVGKGIVRMDKASRAKLGVSVGDYVEVKKV LSVKLRVAEAYPEDVGKGIVRMDKASRAKLGVSVGDYVEVKKV	43/43 (100%)
mkDPBB_sym2	89	9.6	9.9	MPGKSVVARVAEAYPEDVGKRIVRMDKYERAKLGVSVGDYVEVKKV KSVVARVAEAYPEDVGKRIVRMDKYERAKLGVSVGDYVEVKKV	43/43 (100%)
apDPBB_sym1	91	6.3	10.0	MANSSVELRVAEAYPEDVGRGIVRMDKQTRAKLGVSVGDYVEVKKVD SSVELRVAEAYPEDVGRGIVRMDKQTRAKLGVSVGDYVEVKKVD	44/44 (100%)
apDPBB_sym2	91	9.7	10.3	MANRSVVARVAEAYPEDVGRKIVRMDKYLRAKLGVSVGDYVEVKKVE RSVVARVAEAYPEDVGRKIVRMDKYLRAKLGVSVGDYVEVKKVE	44/44 (100%)
Symmetric DPBBs designed using reverse evolution engineering method (RE-design)					
reDPBB_sym1	85	9.3	9.2	PIKLRVMEAYPEDVGKGIVRMDKASRDKLGVSAGDLVEIKGSKT PIKLRVMEAYPEDVGKGIVRMDKASRDKLGVSAGDLVEIKG	41/41 (100%)
reDPBB_sym2	85	9.3	9.2	PMKLRVMEAYPEDVGKGIVRMDKASREKLGVSAGDLVEIKGSKT PMKLRVMEAYPEDVGKGIVRMDKASREKLGVSAGDLVEIKG	41/41 (100%)
reDPBB_sym3	85	9.3	9.2	TIKLRVMEAYPEDVGKGIVRMDKASRDKIGVSAGDLVEIKGSKT TIKLRVMEAYPEDVGKGIVRMDKASRDKIGVSAGDLVEIKG	41/41 (100%)
reDPBB_sym4	89	9.4	9.7	MPGKSVVARVAPAHPEDVGKGIVRMDKYERQNLGVSVGDYVEVKKA KSVVARVAPAHPEDVGKGIVRMDKYERQNLGVSVGDYVEVKKA	43/43 (100%)
reDPBB_sym5	89	6.2	9.6	MPGKSVVARVAPAYPEDVGKGIVRMDKYERANLGVSVGDYVEVDKA KSVVARVAPAYPEDVGKGIVRMDKYERANLGVSVGDYVEVDKA	43/43 (100%)
reDPBB_sym6	89	8.8	9.7	MPGKTVVARVLPAYPEDVGKGIVRMDKYERAKLGVSVGDYVEVEKA KTVVARVLPAYPEDVGKGIVRMDKYERAKLGVSVGDYVEVEKA	43/43 (100%)
reDPBB_sym7	89	8.8	9.6	MPGKSVVARVAPAYPEDVGKGIVRMDKYERAKLGVSVGDYVEVEKA KSVVARVAPAYPEDVGKGIVRMDKYERAKLGVSVGDYVEVEKA	43/43 (100%)
Symmetric DPBBs designed using multi-state design method (MS-design)					
msDPBB_sym1	86	6.7	9.6	SSVVARVALAHEDDVGKNIVRMDEDLMRKLGVKVGDYVEIMKK SSVVARVALAHEDDVGKNIVRMDEDLMRKLGVKVGDYVEIMKK	43/43 (100%)
msDPBB_sym2	86	5.1	9.5	SSVIARVALAHEDDVGKNIVRMDEELMRLLGVKVGDLVEIMKV SSVIARVALAHEDDVGKNIVRMDEELMRLLGVKVGDLVEIMKV	43/43 (100%)

Half fragmented DPBBs					
mk1h	46	9.5	4.9	MPGLSVKLRVAEAYPEDVGKGIVRMDKASRAKLGVSVGDYVEVKKV	-
mk2h	46	9.5	5.1	MPGKSVVARVAEAYPEDVGKRIVRMDKYERAKLGVSVGDYVEVKKV	-
ap1h	47	6.1	5.2	MANSSVELRVAEAYPEDVGRGIVRMDKQTRAKLGVSVGDYVEVKKVD	-
ap2h	47	9.5	5.3	MANRSVVARVAEAYPEDVGRKIVRMDKYLRAKLGVSVGDYVEVKKVE	-
Simplified mk2h					
mk2h_ΔM	46	9.5	5.1	MPGKSVVARVAEAYPEDVGKRIVRVDKYERAKLGVSVGDYVEVKKV	-
mk2h_ΔI	46	9.5	5.1	MPGKSVVARVAEAYPEDVGKRVVRMDKYERAKLGVSVGDYVEVKKV	-
mk2h_ΔL	46	9.5	5.1	MPGKSVVARVAEAYPEDVGKRIVRMDKYERAKVGVSVGDYVEVKKV	-
mk2h_ΔP	46	9.5	5.1	MPGKSVVARVAEAYAEDVGKRIVRMDKYERAKLGVSVGDYVEVKKV	-
mk2h_ΔS	46	9.8	5.2	MPGKKVVARVAEAYPEDVGKRIVRMDKYERAKLGVKVGDYVEVKKV	-
mk2h_ΔY	46	10.0	5.0	MPGKSVVARVAEARPEDVGKRIVRMDKAERAKLGVSVGDVVEVKKV	-
mk2h_ΔMIL	46	9.5	5.1	MPGKSVVARVAEAYPEDVGKRVVRVDKYERAKVGVSVGDYVEVKKV	-
mk2h_ΔMILPS	46	9.8	5.1	MPGKKVVARVAEAYAEDVGKRVVRVDKYERAKVGVKVGDYVEVKKV	-
mk2h_ΔMILPY	46	10.0	4.9	MPGKSVVARVAEARAEDVGKRVVRVDKAERAKVGVSVGDVVEVKKV	-
mk2h_ΔMILYS	46	10.2	5.0	MPGKKVVARVAEARPEDVGKRVVRVDKAERAKVGVKVGDVVEVKKV	-
mk2h_ΔMILPYS	46	10.2	5.0	MPGKKVVARVAEARAEDVGKRVVRVDKAERAKVGVKVGDVVEVKKV	-

Table S2. Data collection and refinement statistics.

Proteins	taVCP_DPBB	mkVCP_DPBB	apVCP_DPBB	apDPBB_sym79	mkDPBB_sym86
PDBID	7DB0	7DG7	7DG9	7DI0	7DI1
Data collection					
Space group	R 3 2	P2(1)	P2(1)2(1)2(1)	P2(1)2(1)2(1)	P2(1)
Unit cell parameters					
a, b, c (Å)	119.0, 119.0, 68.5	32.0, 29.7, 42.1	26.0, 31.0, 94.0	30.3, 77.9, 108.9	31.3, 85.4, 32.7
α, β, γ (°)	90, 90, 120	90, 106, 90	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)	50.00-1.90 (2.02-1.90)	50.00-1.60 (1.70-1.60)	50-1.60 (1.70-1.60)	50-1.60 (1.70-1.60)	50-2.1 (2.22-2.10)
Unique reflections	14642 (2292)	10094 (1569)	19200 (3136)	65486 (10554)	8524 (1471)
Redundancy	22.1 (22.2)	13.20 (13.47)	7.21 (7.22)	3.84 (3.77)	7.67 (7.66)
Average I/sigma(I)	16.63 (2.69)	22.97 (5.92)	33.41 (20.39)	10.08 (1.53)	12.8 (2.83)
CC half	99.5 (61.5)	99.9 (97.9)	99.9 (99.7)	99.7 (56.0)	99.7 (83.3)
*R*symm (%)	11.4 (133)	8.0 (43.3)	4.4 (10.9)	10.0 (87.0)	13.1 (79.3)
Completeness (%)	99.7 (98.1)	98.7 (96.2)	98.8 (99.9)	99.7 (99.4)	84.8 (90.6)
Refinement					
Resolution (Å)	41.07-1.90	40.44-1.60	47.078-1.60	44.63-1.60	30.455-2.097
*R*work	0.2168	0.2065	0.1693	0.179	0.205
*R*free	0.2642	0.2273	0.2048	0.2043	0.2641
Number of atoms	1524	750	857	2424	1533
Protein atoms	1407	664	709	2087	1410
Ligands/ion	45	18	5	50	30
Water	72	69	143	287	93
R.m.s. deviations					
Bond lengths (Å)	0.009	0.01	0.008	0.006	0.008
Bond angles (°)	0.987	1.242	1.021	0.81	1.078
Ramachandran plot (%)					
Most favorable	97.69	100	100	98.84	97.78
Allowed	2.31	0	0	0	1.67
Disallowed	0	0	0	1.16	0.56

mkDPBB_sym1	mkDPBB_sym2	reDPBB_sym1	reDPBB_sym2	reDPBB_sym4	msDPBB_sym2
7DU7	7DU6	7DVC	7DVF	7DVH	7DWW
P31	R3	P4(3)2(1)2	P2(1)2(1)2(1)	P2	P3(1)21
31.2, 31.2, 132.0	70.8, 70.8, 64.0	98.1, 98.1, 177.3	26.0, 33.2, 78.5	64.9, 30.5, 93.4	44.7, 44.7, 147.5
90, 90, 120	90, 90, 120	90, 90, 90	90, 90, 90	90, 101, 90	90, 90, 120
50-1.20 (1.27-1.20)	50-1.60 (1.70-1.60)	50-1.70 (1.81-1.70)	50-1.20 (1.28-1.21)	50-1.70 (1.80-1.70)	50-1.80 (1.91-1.80)
44946 (7142)	15809 (2522)	94555 (14747)	21779 (3340)	40180 (6345)	16713 (2581)
4.88 (4.72)	10.00 (10.12)	28.0 (28.64)	12.1 (7.12)	6.59 (6.42)	9.45 (9.58)
16.14 (2.47)	32.34 (2.77)	33.00 (5.26)	29.03 (4.25)	20.25 (5.53)	12.76 (1.60)
99.9 (91.5)	100.0 (88.3)	100.0 (96.0)	99.9 (99.5)	99.0 (93.7)	99.9 (76.3)
4.0 (50.7)	3.7 (72.3)	7.6 (70.1)	5.0 (37.1)	29.3 (58.9)	8.9 (99.8)
99.7 (98.9)	99.9 (99.8)	99.6 (97.6)	98.6 (99.2)	99.6 (98.5)	99.9 (99.4)
27.071 -1.300	44.319-1.600	47.298-1.705	21.896-1.209	47.905-1.698	38.853-1.802
0.1993	0.1952	0.1619	0.137	0.178	0.2242
0.2126	0.21	0.1732	0.1804	0.2005	0.2399
1464	763	4849	781	3040	1436
1342	678	4104	648	2713	1316
0	0	29	8	0	0
128	85	716	126	327	120
0.009	0.008	0.016	0.013	0.013	0.013
1.047	1.05	1.523	1.527	1.272	1.64
98.81	98.81	98.04	98.8	98.85	95.24
1.19	1.19	1.96	1.2	1.15	3.57
0	0	0	0	0	1.19

ap1h	mk2h	mk2h (Synthetic peptide)	mk2h_ΔP	mk2h_ΔY	mk2h_ΔMIL
7DXS	7DXR	7DXT	7DXU	7DXV	7DXW
P2(1)	P2(1)2(1)2(1)	R32	P2(1)2(1)2(1)	R3	R32
42.6, 40.3, 56.2	55.1, 56.1, 56.6	70.1, 70.1, 65.2	54.1, 55.5, 56.0	72.7, 72.7, 48.9	69.1, 69.1, 65.3
90, 91, 90	90, 90, 90	90, 90, 120	90, 90, 90	90, 90, 120	90, 90, 120
50-2.10 (2.23-2.10)	50-1.60 (1.69-1.60)	50-1.80 (1.91-1.80)	50-2.30 (2.45-2.30)	50-2.21 (2.34-2.21)	50-1.50 (1.60-1.50)
11344 (1776)	23802 (3762)	5895 (938)	7930 (1264)	4833 (705)	9804 (1557)
6.60 (6.21)	11.11 (11.17)	19.72 (19.83)	13.42 (14.27)	10.87 (8.86)	21.62 (21.60)
17.76 (4.17)	25.31 (2.73)	33.39 (4.15)	54.21 (34.38)	29.72 (13.59)	23.33 (4.33)
99.9 (93.8)	100 (89.9)	100 (93.4)	100 (99.9)	99.9 (99.3)	99.8 (94.5)
8.0 (37.8)	5.1 (89.2)	5.9 (73.5)	3.8 (6.5)	5.4 (17.4)	9.3 (88.3)
99.7 (98.1)	99.9 (99.3)	100 (99.8)	99.4 (99.6)	98.0 (88.1)	99.9 (99.6)
33.663-2.102	39.80-1.60	35.075-1.798	38.988-2.310	36.551-2.301	34.798-1.509
0.1886	0.2137	0.1814	0.2569	0.2016	0.1816
0.225	0.2488	0.2061	0.2646	0.2483	0.196
1561	1563	405	1457	699	445
1403	1405	359	1352	665	378
45	39	0	30	0	5
114	119	46	75	34	62
0.015	0.006	0.016	0.009	0.009	0.007
1.165	0.829	1.346	1	1.018	0.988
99.43	100	100	97.56	100	100
0.57	0	0	2.44	0	0
0	0	0	0	0	0

mk2h_ΔMILPS	mk2h_DMILPS (Synthetic peptide)	mk2h_DMILPYS (Synthetic peptide, malonate)	mk2h_DMILPYS (Synthetic peptide, D/L-malic acid)
7DXX	7DXY	7DXZ	7DYC
P4(3)2(1)2	P2(1)2(1)2(1)	P3(1)21	P3(1)21
33.6, 33.6, 147.7	30.5, 39.6, 64.0	52.1, 52.1, 146.4	52.2, 52.2, 146.7
90, 90, 90	90, 90, 90	90, 90, 120	90, 90, 120
50-1.40 (1.49-1.40)	50-1.35 (1.44-1.35)	50-1.90 (2.01-1.90)	50-2.3 (2.44-2.30)
17646 (2652)	17566 (2759)	18940 (2985)	10931 (1638)
11.73 (8.14)	6.26 (6.12)	21.37 (21.55)	21.19 (21.49)
29.38 (5.56)	18.85 (1.73)	31.49 (6.00)	17.17 (2.48)
99.9 (94.0)	100 (83.2)	100 (95.7)	99.8 (85.7)
5.4 (38.8)	3.4 (78.9)	8.6 (72.7)	23.0 (149.9)
99.3 (95.9)	99.2 (98.4)	99.8 (99.5)	99.7 (98.7)
37.053-1.403	24.263-1.400	18.415-1.900	33.225-2.300
0.1838	0.211	0.1948	0.2255
0.2001	0.2296	0.2429	0.2617
874	784	1463	1396
726	688	1327	1283
7	0	14	18
141	96	122	95
0.007	0.008	0.009	0.012
1.039	1.036	1.172	1.347
100	100	100	97.5
0	0	0	2.5
0	0	0	0

Table S3. Ranking of VCP_DPBBs based on internal sequence identity.

Rank	Internal sequence identity (%)	Species
1	45.455	Aeropyrum_pernix_K1_BAA80362
2	41.86	Methanopyrus_kandleri_AV19_AAM01701
3	38.636	Thermogladius_cellulolyticus_1633_AFK50582
4	37.209	taVCP_DPBB
4	37.209	Thermoplasma_acidophilum_DSM_1728_005209
6	34.884	Thermoplasma_volcanium_WP_010917205
6	34.884	Candidatus_Micrarchaeota_archaeon_CG1_02_47_40_01020729.1
6	34.884	Thaumarchaeota_archaeon_MY2_WP_042686022
6	34.884	Pyrococcus_furiosus_WP_011012100
6	34.884	Pyrococcus_horikoshii_OT3_BAA30961
6	34.884	Pyrococcus_horikoshii_WP_048053493
12	34.091	Desulfurococcus_mucosus_DSM_2162_ADV64559
13	32.558	Staphylothermus_marinus_F1_ABN69820
13	32.558	Thermogladius_cellulolyticus_1633_AFK50988
13	32.558	Picrophilus_torridus_DSM_9790_AAT43041
13	32.558	Candidatus_Nitrososphaera_evergladensis_SR1_AIF84648
13	32.558	Thermococcus_kodakarensis_KOD1_BAD84858
13	32.558	Pyrococcus_horikoshii_OT3_BAA29778
13	32.558	Thermofilum_pendens_Hrk_5_ABL78081
13	32.558	Thermococcus_kodakarensis_KOD1_BAD85346
13	32.558	Thermococcus_kodakarensis_KOD1_CAT68952
22	32.432	Candidatus_Nitrososphaera_gargensis_Ga9.2_AFU57173
23	31.818	Candidatus_Micrarchaeota_archaeon_CG_4_10_14_0_2_um_filter_60_11_PIZ91015.1
23	31.818	Candidatus_Nitrososphaera_evergladensis_SR1_AIF82607
25	30.952	Natrialba_magadii_ATCC_43099_ADD03993
26	30.233	Archaeoglobus_veneficus_SNP6_AEA48091
26	30.233	Nitrososphaera_viennensis_EN76_AIC15426
26	30.233	Candidatus_Micrarchaeota_archaeon_CG10_big_fil_rev_8_21_14_0_10_45_29_PIT84804.1
26	30.233	Candidatus_Nitrososphaera_gargensis_Ga9.2_AFU59974
26	30.233	Candidatus_Nitrosopelagicus_brevis_AJA92264
26	30.233	Candidatus_Bathyarchaeota_archaeon_B24-2_PDM26593
26	30.233	Archaeoglobus_veneficus_SNP6_AEA46762
26	30.233	Methanococcus_maripaludis_S2_CAF29732
34	29.545	Nitrososphaera_viennensis_EN76_AIC16229
35	28.571	Halobacterium_salinarum_R1_CAP15469
36	27.907	Candidatus_Pacearchaeota_archaeon_CG10_big_fil_rev_8_21_14_0_10_32_42_PJE81802.1
36	27.907	Methanonatronarchaeum_thermophilum_OUJ19303
36	27.907	Ferroglobus_placidus_DSM_10642_ADC66306
36	27.907	Ferroplasma_acidarmanus_fer1_AGO60899
36	27.907	Candidatus_Parvarchaeum_acidophilus_ARMAN-5_EFD92611
36	27.907	Caldisphaera_lagunensis_DSM_15908_AFZ71117
36	27.907	Pyrolobus_fumarii_1A_AEM37983
36	27.907	Candidatus_Altiarchaeales_archaeon_ex4484_2_0YT54626
36	27.907	Thaumarchaeota_archaeon_MY2_WP_042683811
36	27.907	Cenarchaeum_symbiosum_A_ABK78575
36	27.907	Nitrosopumilus_maritimus_SCM1_ABX12001
36	27.907	Candidatus_Nitrosoarchaeum_limnia_BG20_EPA04818
36	27.907	Thaumarchaeota_archaeon_N4_CDI06512
36	27.907	Thaumarchaeota_archaeon_N4_WP_048197167
36	27.907	Candidatus_Geothermarchaeota_archaeon_ex4572_27_PCN50067

36	27.907	Candidatus_Odinarchaeota_archaeon_LCB_4_OLS17411
36	27.907	Candidatus_Bathyarchaeota_archaeon_B24_KYH37740
36	27.907	Pyrococcus_furiosus_WP_011013021
54	27.273	Candidatus_Heimdallarchaeota_archaeon_LC_3_OLS20813
54	27.273	Candidatus_Heimdallarchaeota_archaeon_LC_2_OLS27801
54	27.273	Ignisphaera_aggregans_DSM_17230_ADM27979
54	27.273	Ignicoccus_hospitalis_KIN4/I_ABU82607
54	27.273	Ignicoccus_hospitalis_WP_052570589
54	27.273	Staphylothermus_marinus_F1_ABN70219
54	27.273	Methanocaldococcus_jannaschii_DSM_2661_AAB99153
61	26.19	Halorubrum_lacusprofundi_ATCC_49239_ACM57952
62	26.087	Candidatus_Nanosalinarum_sp._J07AB56_EGQ40547
62	26.087	Nanohaloarchaea_archaeon_SG9_AOV94540
64	25.581	halophilic_archaeon_DL31_WP_014052869
64	25.581	Candidatus_Diapherotrites_archaeon_CG08_land_8_20_14_0_20_34_12_PIU21023.1
64	25.581	Theionarchaea_archaeon_DG-70_KYK34434
64	25.581	Candidatus_Methanohalarchaeum_thermophilum_OKY77569
64	25.581	Methanohalophilus_mahii_DSM_5219_ADE35783
64	25.581	Methanosarcina_barkeri_227_AKB57915
64	25.581	Candidatus_Heimdallarchaeota_archaeon_LC_2_OLS28009
64	25.581	Candidatus_Parvarchaeum_acidiphilum_ARMAN-4_EEZ93171
64	25.581	Ignicoccus_hospitalis_KIN4/I_ABU81875
64	25.581	Aciduliprofundum_boonei_T469_EDY34590
64	25.581	Nitrososphaera_viennensis_EN76_AIC15017
64	25.581	Candidatus_Nitrososphaera_gargensis_Ga9.2_AFU59997
64	25.581	Candidatus_Nitrososphaera_evergladensis_SR1_AIF85237
64	25.581	Candidatus_Nitrosotalea_devanaterra_CUR52664
64	25.581	Candidatus_Bathyarchaeota_archaeon_B24-2_PDM25984
79	25	Pyrolobus_fumarii_1A_AEM37966
79	25	Marine_Group_III_euryarchaeote_CG-Bathy1_OIR19087
79	25	Candidatus_Caldiarchaeum_subterraneum_BAJ47789
79	25	Methanothermus_fervidus_DSM_2088_ADP77767
79	25	Pyrolobus_fumarii_1A_AEM38259
84	24.324	Candidatus_Nitrosotalea_devanaterra_CUR52322
85	23.913	Candidatus_Nanosalina_sp._J07AB43_EGQ43838
86	23.81	Candidatus_Micrarchaeota_archaeon_Mia14_WP_088820610.1
87	23.256	Thaumarchaeota_archaeon_N4_CDI06696
87	23.256	Thaumarchaeota_archaeon_MY2_WP_042684250
87	23.256	Candidatus_Pacearchaeota_archaeon_CG10_big_fil_rev_8_21_14_0_10_34_76_PIN89331.1
87	23.256	Candidatus_Pacearchaeota_archaeon_CG_4_9_14_0_2_um_filter_39_13_PJC44446.1
87	23.256	Theionarchaea_archaeon_DG-70-1_KYK31403
87	23.256	Candidatus_Odinarchaeota_archaeon_LCB_4_OLS17850
87	23.256	Marine_Group_II_euryarchaeote_MED-G33_PDH24769.1
87	23.256	Marine_Group_II_euryarchaeote_MED-G34_PDH26101.1
87	23.256	uncultured_Candidatus_Thalassoarchaea_euryarchaeote_AKQ06073
87	23.256	Archaeoglobus_fulgidus_DSM_4304_AAB89157
87	23.256	Methanocella_paludicola_SANAE_BAI62891
87	23.256	halophilic_archaeon_DL31_WP_014051352
87	23.256	Halorubrum_lacusprofundi_ATCC_49239_ACM57299
87	23.256	Candidatus_Woesearchaeota_archaeon_CG10_big_fil_rev_8_21_14_0_10_44_13_PIN86640.1

87	23.256	Aeropyrum_pernix_K1_BAA81490
87	23.256	Acidilobus_saccharovorans_345-15_ADL18813
87	23.256	Fervidicoccus_fontis_Kam940_AFH42351
87	23.256	Candidatus_Heimdallarchaeota_archaeon_AB_125_OLS33307
87	23.256	Candidatus_Altiarchaeales_archaeon_IMC4_ODS42491
87	23.256	Methanohalophilus_mahii_DSM_5219_ADE36585
87	23.256	Methanohalophilus_mahii_WP_048902269
87	23.256	Methanosarcina_barkeri_227_AKB57217
87	23.256	Candidatus_Bathyarchaeota_archaeon_BA2_KPV62335
87	23.256	Candidatus_Bathyarchaeota_archaeon_B26-2_KYH41545
87	23.256	Ferroglobus_placidus_WP_048086329
87	23.256	Archaeoglobus_fulgidus_DSM_4304_AAB89948
87	23.256	Aciduliprofundum_boonei_T469_EDY35166
114	22.727	Candidatus_Micrarchaeum_acidiphilum_ARMAN-2_EET90444
114	22.727	Candidatus_Micrarchaeota_archaeon_Mia14_ASI13900.1
114	22.727	Candidatus_Nitrososphaera_evergladensis_SR1_AIF85315
114	22.727	Thermosphaera_aggregans_DSM_11486_ADG91457
114	22.727	Methanobacterium_formicicum_DSM_3637_EKF86248
119	21.951	Caldisphaera_lagunensis_DSM_15908_AFZ70901
120	21.739	Candidatus_Haloredivivus_sp._G17_EHK01882
121	21.429	Candidatus_Methanohalarchaeum_thermophilum_OKY78788
121	21.429	Halobacterium_salinarum_R1_CAP14060
123	20.93	Candidatus_Thorarchaeota_archaeon_AB_25_OLS24014
123	20.93	Halorubrum_lacusprofundi_ATCC_49239_ACM57647
123	20.93	Halorubrum_lacusprofundi_WP_049933528
123	20.93	Desulfurococcus_mucosus_DSM_2162_ADV65387
123	20.93	Thermosphaera_aggregans_DSM_11486_ADG91259
123	20.93	Methanocella_paludicola_SANAE_BAI60129
123	20.93	Candidatus_Pacearchaeota_archaeon_CG10_big_fil_rev_8_21_14_0_10_35_219_PIO08257.1
123	20.93	uncultured_Candidatus_Thalassoarchaea_euryarchaeote_ANV81027
123	20.93	Marine_Group_II_euryarchaeote_MED-G38_PDH22657.1
123	20.93	Halobacterium_salinarum_R1_CAP14207
123	20.93	Natrialba_magadii_ATCC_43099_ADD03768
123	20.93	Candidatus_Methanoperedens_nitroreducens_KCZ71217
123	20.93	Candidatus_Nitrosopelagicus_brevis_AJA91966
123	20.93	Caldisphaera_lagunensis_WP_048816895
123	20.93	Nanoarchaeum_equitans_Kin4-M_AAR39317
123	20.93	Thermofilum_pendens_Hrk_5_ABL77714
123	20.93	Thermofilum_pendens_WP_052885007
123	20.93	Candidatus_Methanoperedens_nitroreducens_KCZ71917
123	20.93	Pyrodictium_delaneyi_ALL01915
123	20.93	Candidatus_Thorarchaeota_archaeon_SMTZ1-83_KXH77531
123	20.93	Candidatus_Thorarchaeota_archaeon_SMTZ1-45_KXH73371
123	20.93	Candidatus_Thorarchaeota_archaeon_AB_25_OLS31443
123	20.93	Methanobacterium_formicicum_DSM_3637_EKF85426
123	20.93	Candidatus_Bathyarchaeota_archaeon_CG07_land_8_20_14_0_80_47_9_PIU59706
123	20.93	Pyrobaculum_aerophilum_str._IM2_AAL64724
123	20.93	Thermoproteus_tenax_Kra_1_CCC81686
123	20.93	Thermoproteus_tenax_WP_052883121
150	20.455	Candidatus_Nitrososphaera_gargensis_Ga9.2_AFU59866

150	20.455	Candidatus_Diapherotrites_archaeon_CG08_land_8_20_14_0_20_30_16_PIU22450.1
150	20.455	Natrialba_magadii_ATCC_43099_ADD06942
150	20.455	Candidatus_Huberarchaea_archaeon_CG03_land_8_20_14_0_80_31_114_PIV13743.1
150	20.455	Candidatus_Bathyarchaeota_archaeon_BA2_KPV63518
150	20.455	Candidatus_Methanoperedens_nitroreducens_KCZ72076
150	20.455	Fervidicoccus_fontis_Kam940_AFH43154
150	20.455	Caldivirga_maquilingensis_IC-167_ABW01805
158	20	Metallosphaera_yellowstonensis_MK1_EHP69309
158	20	Hyperthermus_butylicus_DSM_5456_ABM80172
158	20	Hyperthermus_butylicus_WP_048061379
161	19.565	Acidilobus_saccharovorans_345-15_ADL19616
162	19.048	Candidatus_Diapherotrites_archaeon_CG10_big_fil_rev_8_21_14_0_10_31_34_PIN99301.1
163	18.605	Candidatus_Diapherotrites_archaeon_CG09_land_8_20_14_0_10_32_12_PIU02624.1
163	18.605	Cenarchaeum_symbiosum_A_ABK78458
163	18.605	Candidatus_Thorarchaeota_archaeon_SMTZ1-83_KXH72124
163	18.605	Candidatus_Thorarchaeota_archaeon_SMTZ1-45_KXH74223
163	18.605	Candidatus_Nitrososphaera_evergladensis_SR1_AIF83909
163	18.605	Candidatus_Pacearchaeota_archaeon_CG10_big_fil_rev_8_21_14_0_10_32_42_PJE81799.1
163	18.605	Candidatus_Geothermarchaeota_archaeon_ex4572_27_PCN50199
163	18.605	Candidatus_Caldiarchaeum_subterraneum_BAJ47117
163	18.605	Theionarchaea_archaeon_DG-70_KYK31941
163	18.605	Lokiarchaeum_sp._GC14_75_KKK40880
163	18.605	Candidatus_Woesearchaeota_archaeon_CG10_big_fil_rev_8_21_14_0_10_36_11_PIN75767.1
163	18.605	Candidatus_Nitrososphaera_gargensis_Ga9.2_AFU57477
163	18.605	Nanoarchaeota_archaeon_7A_AMD29691
163	18.605	nanoarchaeote_Nst1_EOD42766
163	18.605	Candidatus_Methanoperedens_nitroreducens_KCZ73677
178	18.182	Candidatus_Nitrosotalea_devanaterra_CUR51378
178	18.182	Candidatus_Methanomethylophilus_alvus_Mx1201_AGI86363
178	18.182	Candidatus_Korarchaeum_cryptofilum_OPF8_ACB06954
178	18.182	Caldivirga_maquilingensis_IC-167_ABW01385
178	18.182	Candidatus_Aenigmarchaeota_archaeon_ex4484_224_OYT42612.1
178	18.182	Candidatus_Aenigmarchaeota_archaeon_CG_4_10_14_3_um_filter_37_21_PIY34793.1
178	18.182	Candidatus_Heimdallarchaeota_archaeon_LC_2_OLS28431
178	18.182	Candidatus_Bathyarchaeota_archaeon_B24_KYH36593
178	18.182	Methanobrevibacter_smithii_ATCC_35061_ABQ86847
178	18.182	Methanosarcina_barkeri_227_AKB58810
178	18.182	Methanocella_paludicola_SANAE_BAI62887
178	18.182	Metallosphaera_yellowstonensis_MK1_EHP69577
178	18.182	Vulcanisaeta_moutnovskia_768-28_ADY00631
191	17.778	Pyrodictium_delaneyi_ALL01313
192	16.279	Candidatus_Nitrosopumilus_salaria_WP_008298811.1
192	16.279	Candidatus_Nitrosopumilus_salaria_WP_008297191.1
192	16.279	Nitrosopumilus_maritimus_SCM1_ABX11911
192	16.279	Candidatus_Nitrosotalea_devanaterra_CUR52859
192	16.279	Candidatus_Nitrosotenuis_cloacae_AJZ75973
192	16.279	Lokiarchaeum_sp._GC14_75_KKK44519
192	16.279	Candidatus_Pacearchaeota_archaeon_CG10_big_fil_rev_8_21_14_0_10_35_219_PIO07960.1
192	16.279	Marine_Group_II_euryarchaeote_MED-G37_PDH23088.1
192	16.279	Candidatus_Bathyarchaeota_archaeon_B24-2_PDM26092

192	16.279	Candidatus_Bathyarchaeota_archaeon_B26-2_KYH42037
192	16.279	archaeon_GW2011_AR15_AJF61321
192	16.279	Hyperthermus_butylicus_DSM_5456_ABM81185
192	16.279	Hyperthermus_butylicus_WP_048061869
192	16.279	Candidatus_Lokiarchaeota_archaeon_CR_4_OLS12095
192	16.279	Candidatus_Methanohalarchaeum_thermophilum_OKY78140
192	16.279	Methanonatronarchaeum_thermophilum_OUJ19199
192	16.279	Methanosarcina_barkeri_227_AKB58911
192	16.279	Candidatus_Odinarchaeota_archaeon_LCB_4_OLS17239
210	15.909	archaeon_GW2011_AR5_KHO47409
210	15.909	Candidatus_Korarchaeum_cryptofilum_OPF8_ACB06989
212	14.286	Natrialba_magadii_ATCC_43099_ELY26084
212	14.286	halophilic_archaeon_DL31_WP_014051559
214	13.953	Candidatus_Heimdallarchaeota_archaeon_LC_2_OLS24790
214	13.953	Candidatus_Diapherotrites_archaeon_CG09_land_8_20_14_0_10_32_12_PIU02962.1
214	13.953	Candidatus_Nitrosotalea_devanaterra_CUR51973
214	13.953	Cenarchaeum_symbiosum_A_ABK77156
218	13.636	Natrialba_magadii_ATCC_43099_ADD04359
218	13.636	Candidatus_Heimdallarchaeota_archaeon_LC_3_OLS27535
218	13.636	Candidatus_Heimdallarchaeota_archaeon_LC_3_OLS19891
218	13.636	Pyrobaculum_aerophilum_str._IM2_AAL62959
218	13.636	Vulcanisaeta_moutnovskia_768-28_ADY01148
223	13.514	Candidatus_Woesearchaeota_archaeon_CG10_big_fil_rev_8_21_14_0_10_45_16_PIN74326.1
224	13.333	Candidatus_Aenigmarchaeota_archaeon_ex4484_56_OYT43213.1
225	12.766	Sulfolobus_tokodaii_str._7_BAK54193
225	12.766	Sulfolobus_tokodaii_WP_052846861
227	11.905	Candidatus_Woesearchaeota_archaeon_CG10_big_fil_rev_8_21_14_0_10_44_13_PIN86711.1
228	11.628	Candidatus_Nitrosoarchaeum_koreensis_MY1_EGP94659
228	11.628	Candidatus_Nitrosoarchaeum_limnia_WP_010191855
228	11.628	Candidatus_Altiarchaeum_sp._CG2_30_32_3053_OIQ05114
228	11.628	Thaumarchaeota_archaeon_N4_CDI05953
228	11.628	Candidatus_Nitrosotenuis_cloacae_AJZ76520
228	11.628	Nitrosopumilus_maritimus_SCM1_ABX13119
228	11.628	Candidatus_Nitrosoarchaeum_limnia_WP_010191195
228	11.628	Candidatus_Nitrososphaera_evergladensis_SR1_AIF83719
228	11.628	Candidatus_Nitrososphaera_gargensis_Ga9.2_AFU58810
228	11.628	archaeon_GW2011_AR20_AJF62478
238	11.364	Candidatus_Geothermarchaeota_archaeon_ex4572_27_PCN51246
238	11.364	Sulfolobus_tokodaii_str._7_BAK54256
240	9.524	Candidatus_Altiarchaeales_archaeon_WOR_SM1_SCG_ODS35378
241	9.302	Thaumarchaeota_archaeon_MY2_WP_042687141
241	9.302	Candidatus_Nitrosoarchaeum_koreensis_WP_048109923
241	9.302	Nitrososphaera_viennensis_EN76_AIC14530
244	9.091	Lokiarchaeum_sp._GC14_75_KKK42851
244	9.091	Thermoproteus_tenax_Kra_1_CCC81122
244	9.091	Hadesarchaea_archaeon_YNP_N21_KUO42916
247	7.5	Candidatus_Nitrososphaera_evergladensis_SR1_AIF82434
248	4.651	Candidatus_Nitrosopumilus_salaria_WP_008297438.1
249	4.545	Nitrososphaera_viennensis_EN76_AIC16426

Table S4. Experimental characterization of native DPBBs.

Protein name	Mw (kDa)	Experimentally characterized properties							
		Expression	Solubility	SEC analysis		CD analysis			Crystal structure (PDBID)
				Apparent Mw (kDa)	Oligomeric state	2nd structure	Refolding abiity	T_m	
taVCP_DPBB	10.2	✔	✔	10.1	1.0	α/β	✔	64.2	✔ (7DB0)
mkVCP_DPBB	9.8	✔	✔	12.6	1.3	α/β	✔	69.5	✔ (7DG7)
apVCP_DPBB	9.9	✔	✔	9.1	0.9	α/β	✔	>85	✔ (7DG9)

Table S5. Experimental characterization of designed symmetric DPBBs.

Design class	Protein name	Mw (kDa)	Experimentally characterized properties							
			Expression	Solubility	SEC analysis		CD analysis			Crystal structure
					Apparent Mw (kDa)	Oligomeric state	2nd structure	Refolding ability	T_m	
Intermediate in SC-design	apDPBB_sym_63	10.0	✔	✔	9.7	1.0	α/β	✔	>85	
	apDPBB_sym_79	10.1	✔	✔	9.7	1.0	α/β	✔	>85	✔ (7DI0)
	apDPBB_sym_84	10.1	✔	✔	10.6	1.0	α/β	✔	>85	
	mkDPBB_sym_67	9.7	✔	✔	11.6	1.2	α/β	✔	69.2	
	mkDPBB_sym_81[a]	9.9	✔	✔	–	–	–	–	–	
	mkDPBB_sym_86	9.9	✔	✔	10.8	1.1	α/β	✔	>85	✔ (7DI1)
SC designs	mkDPBB_sym1	9.6	✔	✔	12.0	1.3	α/β	✔	>85	✔ (7DU7)
	mkDPBB_sym2	9.9	✔	✔	8.2	0.8	α/β	✔	>85	✔ (7DU6)
	apDPBB_sym1	10.0	✔	✔	11.7	1.2	α/β	✔	>85	
	apDPBB_sym2[a]	10.3	✔	✔	–	–	–	–	–	
RE desigs	reDPBB_sym1	9.2	✔	✔	10.4	1.1	α/β	✔	>85	✔ (7DVC)
	reDPBB_sym2	9.2	✔	✔	10.9	1.2	α/β	✔	81.8	✔ (7DVF)
	reDPBB_sym3	9.2	✔	✔	10.7	1.2	α/β	✔	>85	
	reDPBB_sym4	9.7	✔	✔	6.6	0.7	α/β	✔	70.0	✔ (7DVH)
	reDPBB_sym5	9.6	–	–	–	–	–	–	–	
	reDPBB_sym6	9.7	–	–	–	–	–	–	–	
	reDPBB_sym7	9.6	✔	✔	8.0	0.8	α/β	✔	67.3	
MS designs	msDPBB_sym1	9.6	✔	✔	9.5	1.0	α/β	✔	>85	
	msDPBB_sym2	9.5	✔	✔	11.0	1.2	α/β	✔	>85	✔ (7DWW)

[a] The proteins could not be purified due to their low stability.

Table S6. Experimental characterization of half-fragmented DPBBs.

Protein name	Mw (kDa)	Experimentally characterized properties							
		Expression	Solubility	SEC analysis		CD analysis			Crystal structure (PDBID)
				Apparent Mw (kDa)	Oligomeric state	2nd structure	Refolding abiity	T_m	
mk1h	4.9	✔	✔	13.4	2.7	α/β	✔	>85	
mk2h	5.1	✔	✔	9.6	1.9	α/β	✔	81.5	✔ (7DXR/7DXT)
ap1h	5.2	✔	✔	13.2	2.6	α/β	✔	81.2	✔ (7DXS)
ap2h	5.3	✔	✔	10.8	2.0	α/β	✔	>85	

Table S7. Crystallization conditions for the synthetic mk2h peptide.

	Reservoir solution (RS)
1	1000mM Potassium sodium tartrate, 100 mMImidazole/Hydrochloric acid pH8.0, 200 mM Sodium chloride
2	400mM Sodium phosphate monobasic/1600mM Potassium phosphate dibasic, 100mM Imidazole/Hydrochloric acid pH8.0, 200mM Sodium chloride
3	1000mM Sodium citrate tribasic, 100mM CHES/Sodium hydroxide pH9.5
4	1000mM Sodium citrate tribasic, 100mM Tris base/Hydrochloric acid pH7.0, 200mM Sodium chloride
5	800mM Sodium phosphate monobasic/1200mM Potassium phosphate dibasic, 100mM Sodium acetate/Acetic acid pH4.5
6	10%(w/v) PEG 3000, 100mM Sodium phosphate dibasic/Citric acid pH4.2, 200mM Sodium chloride
7	20%(v/v) Jeffamine M-600, 100 mM HEPES/Sodium hydroxide pH7.5

Table S8. Quantification of peptide species by LC/MS.

mk1h peptide						
	Before SEC		Aggregation frac.		Dimer frac.	
Monoisotopic mass (Da)	Sum Intensity	Relative Abundance (%)	Sum Intensity	Relative Abundance(%)	Sum Intensity	Relative Abundance(%)
2914.6	2233357	16.7	1538848	19.3	1666535	8.8
3428.9	285446	2.1	50208	0.6	362748	1.9
4004.1	359518	2.7	673725	8.5	11170	0.1
4174.3	110288	0.8	388998	4.9	ND[a]	
4316.4	419224	3.1	1242401	15.6	23622	0.1
4403.4	181392	1.4	519933	6.5	ND[a]	
4408.5	51985	0.4	357992	4.5	ND[a]	
4415.5	2939688	22.0	7965074	100.0	271045	1.4
4431.5	287595	2.2	697384	8.8	46841	0.2
4443.4	97978	0.7	358958	4.5	10669	0.1
4514.5	266774	2.0	950075	11.9	21476	0.1
4559.5	1436130	10.7	3926160	49.3	218946	1.2
4571.5	184029	1.4	471198	5.9	97616	0.5
4601.6	323717	2.4	455629	5.7	19369	0.1
4658.5 (mk1h FL)	**13366797**	**100.0**	**1608175**	**20.2**	**18906107**	**100.0**
4674.5 (mk1h FL met_OX)	558788	4.2	262902	3.3	719772	3.8
mk2h peptide						
	Before SEC		Aggregation frac.		Dimer frac.	
Monoisotopic mass (Da)	Sum Intensity	Relative Abundance(%)	Sum Intensity	Relative Abundance(%)	Sum Intensity	Relative Abundance(%)
1010.6	299575	2.7	370555	2.6	198979	1.2
3575.0	382652	3.5	1076537	7.5	13218	0.1
3600.9	601648	5.5	28685	0.2	70043	0.4
3662.0	621514	5.6	72742	0.5	814803	5.0
4321.4	216421	2.0	337394	2.4	4246	0.0
4477.5	277006	2.5	600300	4.2	ND[a]	
4592.5	311937	2.8	993232	6.9	ND[a]	
4620.5	150073	1.4	387705	2.7	35378	0.2
4633.6	225325	2.0	604106	4.2	ND[a]	
4658.6	306598	2.8	17624	0.1	10857	0.1
4720.6	210070	1.9	589808	4.1	10318	0.1
4748.6	6001828	54.5	14336914	100.0	436117	2.7
4764.6	442905	4.0	688135	4.8	104955	0.6
4819.6	241081	2.2	301508	2.1	180023	1.1
4835.6 (mk2h FL)	**11002630**	**100.0**	**850392**	**5.9**	**16360720**	**100.0**
4851.6 (mk2h FL met_OX)	524838	4.8	150537	1.0	1079166	6.6
4862.6	130898	1.2	285408	2.0	ND[a]	

[a] The peptide could not be detected by LC/MS.

Table S9. Amino acid usage in the designed DPBBs.

		Amino acids																			Total a.a. types
		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
Native DPBBs	taVCP_DPBB	5	1	7	7 (8)	0	7	0	7	6	4	2 (3)	4 (6)	2	0	13	6 (7)	2	12	0	1	16
	mkVCP_DPBB	8	0	7	8	0	8 (9)	0	4	10	4	2 (3)	1	4 (5)	1	9	4	3	12	0	2	16
	apVCP_DPBB	7 (8)	0	8	6	1	9	0	4	4	5	1 (2)	0 (1)	3	1	11	5	4	16	1	2	18
Symmetric designed DPBBs	mkDPBB_sym_65	7	0	7	5 (6)	0	7 (8)	0	2	11	6	2 (3)	1	4 (5)	0	8	6	1	17	0	2	15
	mkDPBB_sym_79	6	0	6	5 (6)	0	7 (8)	0	2	11	6	2 (3)	1	3 (4)	0	10	5	0	19	0	3	14
	mkDPBB_sym_84	8	0	6	7 (8)	0	7 (8)	0	2	12	4	2 (3)	0	3 (4)	0	7	5	0	19	0	5	13
	apDPBB_sym_63	7 (8)	0	7	6	1	8	0	2	6	4	2 (3)	0 (1)	3	1	11	7	3	17	1	2	17
	apDPBB_sym_79	6 (7)	0	7	7	0	7	0	2	7	6	2 (3)	0 (1)	3	1	12	5	1	19	0	3	15
	apDPBB_sym_84	7 (8)	0	7	8	0	7	0	2	9	4	2 (3)	0 (1)	2	1	9	5	1	19	0	5	15
	mkDPBB_sym1	8	0	6	6	0	8 (9)	0	2	12	6	2 (3)	0	2 (3)	0	6	6	0	18	0	4	13
	mkDPBB_sym2	8	0	6	8	0	6 (7)	0	2	12	2	2 (3)	0	2 (3)	0	8	4	0	20	0	6	13
	apDPBB_sym1	6 (7)	0	8	8	0	8	0	2	8	4	2 (3)	0 (1)	2	2	8	6	2	18	0	4	15
	apDPBB_sym2	8 (9)	0	6	8	0	6	0	2	10	4	2 (3)	0 (1)	2	0	10	4	0	20	0	6	13
	rvDPBB_sym1	6	0	8	6	0	10	0	6	10 (11)	6	4	0	4	0	6	4 (5)	0 (1)	10	0	2	13
	rvDPBB_sym2	6	0	6	8	0	10	0	4	10 (11)	6	6	0	4	0	6	4 (5)	0 (1)	10	0	2	13
	rvDPBB_sym3	6	0	8	6	0	10	0	8	10 (11)	4	4	0	2	0	6	4 (5)	2 (3)	10	0	2	14
	rvDPBB_sym4	8	0	6	6	0	8 (9)	2	2	10	2	2 (3)	2	4 (5)	2	6	4	0	18	0	4	16
	rvDPBB_sym5	10	0	8	6	0	8 (9)	0	2	8	2	2 (3)	2	4 (5)	0	6	4	0	18	0	6	14
	rvDPBB_sym6	8	0	6	8	0	8 (9)	0	2	10	4	2 (3)	0	4 (5)	0	6	2	2	18	0	6	14
	rvDPBB_sym7	10	0	6	8	0	8 (9)	0	2	10	2	2 (3)	0	4 (5)	0	6	4	0	18	0	6	13
	msDPBB_sym1	6	0	10	6	0	6	2	4	10	6	6	2	0	0	6	4	0	16	0	2	14
	msDPBB_sym2	6	0	8	8	0	6	2	6	6	10	6	2	0	0	6	4	0	16	0	0	13
Half-fragmented DPBBs	mk1h	4	0	3	3	0	4 (5)	0	1	6	3	1 (2)	0	1 (2)	0	3	3	0	9	0	2	13
	mk2h	4	0	3	4	0	3 (4)	0	1	6	1	1 (2)	0	1 (2)	0	4	2	0	10	0	3	13
	ap1h	3 (4)	0	4	4	0	5	0	1	4	2	1 (2)	0 (1)	1	1	4	3	1	9	0	2	15
	ap2h	4 (5)	0	3	4	0	3	0	1	5	2	1 (2)	0 (1)	1	0	5	2	0	10	0	3	13
Simplified mk2h	mk2h_ΔM	4	0	3	4	0	3 (4)	0	1	6	1	0 (1)	0	1 (2)	0	4	2	0	11	0	3	12
	mk2h_ΔI	4	0	3	4	0	3 (4)	0	0	6	1	1 (2)	0	1 (2)	0	4	2	0	11	0	3	12
	mk2h_ΔL	4	0	3	4	0	3 (4)	0	1	6	0	1 (2)	0	1 (2)	0	4	2	0	11	0	3	12
	mk2h_ΔP	5	0	3	4	0	3 (4)	0	1	6	1	1 (2)	0	0 (1)	0	4	2	0	10	0	3	12
	mk2h_ΔS	4	0	3	4	0	3 (4)	0	1	8	1	1 (2)	0	1 (2)	0	4	0	0	10	0	3	12
	mk2h_ΔY	5	0	3	4	0	3 (4)	0	1	6	1	1 (2)	0	1 (2)	0	5	2	0	11	0	0	12
	mk2h_ΔMIL	4	0	3	4	0	3 (4)	0	0	6	0	0 (1)	0	1 (2)	0	4	2	0	13	0	3	10
	mk2h_ΔMILPS	5	0	3	4	0	3 (4)	0	0	8	0	0 (1)	0	0 (1)	0	4	0	0	13	0	3	8
	mk2h_ΔMILPY	6	0	3	4	0	3 (4)	0	0	6	0	0 (1)	0	0 (1)	0	5	2	0	14	0	0	8
	mk2h_ΔMILYS	5	0	3	4	0	3 (4)	0	0	8	0	0 (1)	0	1 (2)	0	5	0	0	14	0	0	8
	mk2h_ΔMILPYS	6	0	3	4	0	3 (4)	0	0	8	0	0 (1)	0	0 (1)	0	5	0	0	14	0	0	7

The values in parentheses indicate the number including the amino acids on the loop and linker.

Table S10. Experimental characterization of simplified DPBBs.

Protein name	Mw (kDa)	Experimentally characterized properties							
		Expression	Solubility	SEC analysis		CD analysis			Crystal structure (PDBID)
				Apparent Mw (kDa)	Oligomeric state	2nd structure	Refolding abiity	T_m	
mk2h_ΔM	5.1	✔	✔	9.5	1.9	α/β	✔	71.3	
mk2h_ΔI	5.1	✔	✔	9.6	1.9	α/β	✔	74.4	
mk2h_ΔL	5.1	✔	✔	9.6	1.9	α/β	✔	76.2	
mk2h_ΔP	5.1	✔	✔	9.7	1.9	α/β	✔	71.8	✔ (7DXU)
mk2h_ΔS	5.2	✔	✔	13.1	2.5	α/β	✔	78.5	
mk2h_ΔY	5.0	✔	✔	12.8	2.6	α/β		50.9	✔ (7DXV)
mk2h_ΔMIL	5.1	✔	✔	10.4	2.1	α/β		54.6	✔ (7DXW)
mk2h_ΔMILPS	5.1	✔	✔	10.8	2.1	α/β		50.5	✔ (7DXX/7DXY)
mk2h_ΔMILPY	4.9	✔	✔	19.7	4.0	random-coil	-	-	-
mk2h_ΔMILYS	5.0	✔	✔	20.6	4.1	random-coil	-	-	-
mk2h_ΔMILPYS	5.0	✔	✔	20.3	4.1	random-coil	-	-	✔ (7DXZ/7DYC)

Table S11. Gene and primer sequences.

Protein name	Sequence
Isolated DPBBs from VCP proteins	
taVCP_DPBB	AAGTCCTCTTTCAGGGACCCATGGAAAGCAACAACGGTATTATTCTGCGTGTTGCAGAAGCAAATAGCACCGATCCGGGTATGAGCCGTGTTCGTCTGGATGA AAGCAGCCGTCGTCTGCTGGATGCAGAAATTGGTGATGTTGTGGAAATTGAGAAAGTGCGTAAAACCGTTGGTCGTGTTTATCGTGCACGTCCGGAAGATGAA AATAAAGGTATTGTTCGTATCGATAGCGTGATGCGTAATAATTGTGGTGCAAGCATTGGCGATAAAGTGAAAGTTCGTAAAGTGCGCTAAGGATCCGAATTCT GTACAGG
mkVCP_DPBB	AAGTCCTCTTTCAGGGACCCATGCCTGGTCTGCCGATTAAACTGCGTGTTGAAAAAGCATATCCGGAAGATGTTGGTAAACGTGCAGTTCGTATGGATAAAGC AAGCCGTGATCGTATTGGTGTTAGCGAAGGTGATCTGGTGAAAATTACCGGTAGCAAAACCACCGTTGCACGTGTTCTGCCTGCAAAAAAAGAAGATGTAGGC AAAGGTATTGTGCGCATGGATAAATATGAACGTCAGAATGCCGGTGCAAGCGTTGGTGAACCGGTTGAAGTTGATCGTGCAGAATAAGGATCCGAATTCTGTA CAGG
apVCP_DPBB	AAGTCCTCTTTCAGGGACCCATGGCAAATAGCAGCGTTGAACTGCGTGTTAGCGAAGCATATCCGCGTGATGTTGGTCGTAAAATTGTTCGTATTGATCGTCA GACCGCAGCACGTCTGGGTGTTGAAGTTGGTGATTTTGTTAAAGTGAGCAAAGGTGATCGTAGCGTTGTTGCAGTTGTTTGGCCTCTGCGTCCGGATGATGAA GGTCGTGGTATTATTCGTATGGATGGTTATCTGCGTGCAGCACTGGGTGTTACCGTGGGTGATACCGTTACCGTTGAAAAAGCAGAATAAGGATCCGAATTCT GTACAGG
Symmetric DPBBs designed using Symmetric-conservation method (SC-design)	
apDPBB_sym_63	AAGTCCTCTTTCAGGGACCCATGGCAAATAGCAGCGTTGAACTGCGTGTTAGCGAAGCATATCCGGAAGATGTTGGTCGTAAAATTGTGCGTATGGATAAACA GACCCGTGCACGTCTGGGTGTTAGCGTTGGTGATTTTGTTAAAGTTAGCAAAGGTGATCGTAGCGTTGTTGCACGTGTTTGGCCTGCACGTCCTGAAGATGTA GGCCGTGGTATTGTTCGCATGGACAAATATCTGCGTGCAGCACTGGGTGTTTCAGTGGGTGATACCGTTACCGTTGAAAAAGCAGAATAAGGATCCGAATTCT GTACAGG
apDPBB_sym_79	AAGTCCTCTTTCAGGGACCCATGGCAAATAGCAGCGTTGAACTGCGTGTTCTGGAAGCACGTCCGGAAGATGTTGGTCGTAAAATTGTTCGTATGGATAAACA GACCCGTGCACGTCTGGGTGTTAGCGTTGGTGATTATGTTGAAGTGAAAAAAGTTGATCGTAGCGTTGTTGCACGTGTTCTGCCTGCGCGTCCTGAGGATGTT GGCCGTGGTATTGTGCGCATGGACAAATATCTGCGTGCAGCACTGGGTGTTTCAGTGGGCGATTACGTGGAAGTTAAAAAGGTGGAATAAGGATCCGAATTCT GTACAGG
apDPBB_sym_84	AAGTCCTCTTTCAGGGACCCATGGCAAATAGCAGCGTTGAACTGCGTGTTGCCGAAGCATATCCGGAAGATGTTGGTCGTAAAATTGTGCGTATGGATAAACA GACCCGTGCAAAACTGGGTGTTAGCGTTGGTGATTATGTGGAAGTTAAAAAAGTTGATCGTAGCGTTGTTGCACGTGTGGCGGAAGCCTATCCTGAGGATGTA GGCCGTGGTATTGTTCGCATGGACAAATATCTGCGTGCCAAATTAGGTGTTTCAGTGGGCGATTATGTTGAGGTGAAAAAGGTGGAATAAGGATCCGAATTCT GTACAGG
mkDPBB_sym_67	AAGTCCTCTTTCAGGGACCCATGCCTGGTCTGAGCGTTAAACTGCGTGTTGAAAAAGCATATCCGGAAGATGTTGGTAAACGTATTGTGCGTATGGATAAAGC AAGCCGTGCACGTCTGGGTGTTAGCGTTGGTGATCTGGTTAAAGTTACCAAAAGCAAAAGCGTTGTTGCACGTGTTCTGCCTGCAAAACCTGAAGATGTAGGC AAAGGCATTGTTCGCATGGACAAATATGAACGTGCAAATCTGGGTGTGTCAGTGGGTGATCCGGTTGAAGTAGATAAAGCAGAATAAGGATCCGAATTCTGTA CAGG
mkDPBB_sym_81	AAGTCCTCTTTCAGGGACCCATGCCTGGTCTGAGCGTTAAACTGCGTGTTCTGAAAGCACGTCCGGAAGATGTTGGTAAACGTATTGTTCGTATGGATAAAGC AAGCCGTGCACGTCTGGGTGTTAGCGTTGGTGATTATGTTGAAGTGAAAAAGGTTAAAAGCGTTGTTGCACGTGTTCTGCCTGCGCGTCCTGAGGATGTTGGC AAAGGCATCGTGCGCATGGATAAATATGAACGTGCAAATCTGGGTGTGTCAGTGGGCGATTACGTGGAAGTTAAAAAAGTGGAATAAGGATCCGAATTCTGTA CAGG
mkDPBB_sym_86	AAGTCCTCTTTCAGGGACCCATGCCTGGTCTGAGCGTTAAACTGCGTGTTGCCGAAGCATATCCGGAAGATGTTGGTAAACGTATTGTGCGTATGGATAAAGC AAGCCGTGCAAAACTGGGTGTTAGCGTTGGTGATTATGTGGAAGTCAAAAAAGTTAAAAGCGTTGTTGCACGTGTGGCGGAAGCCTATCCTGAGGATGTAGGC AAAGGCATTGTTCGCATGGACAAATATGAACGTGCCAAATTAGGTGTTAGTGTGGGCGATTACGTTGAGGTTAAAAAAGTGGAATAAGGATCCGAATTCTGTA CAGG
mkDPBB_sym1	AAGTCCTCTTTCAGGGACCCATGCCTGGTCTGAGCGTTAAACTGCGTGTTGCCGAAGCATATCCGGAAGATGTTGGTAAAGGTATTGTGCGTATGGATAAAGC AAGCCGTGCAAAACTGGGTGTTAGCGTTGGTGATTATGTGGAAGTTAAAAAAGTTCTGTCAGTGAAACTGCGCGTGGCAGAGGCCTATCCTGAGGATGTAGGC AAAGGCATCGTTCGCATGGACAAAGCATCACGTGCCAAATTAGGTGTTTCAGTGGGCGATTATGTTGAGGTGAAAAAGGTGTAAGGATCCGAATTCTGTACAG G
mkDPBB_sym2	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAGCGTTGTTGCACGTGTTGCCGAAGCATATCCGGAAGATGTTGGTAAACGTATTGTGCGCATGGATAAATA TGAACGTGCAAAACTGGGTGTTAGCGTTGGTGATTATGTGGAAGTTAAAAAAGTTAAAAGTGTGGTTGCCCGTGTGGCAGAGGCCTATCCTGAGGATGTGGGC AAACGCATCGTTCGTATGGACAAATACGAGCGTGCCAAATTAGGTGTTTCAGTGGGCGATTATGTTGAGGTGAAAAAGGTGTAAGGATCCGAATTCTGTACAG G
apDPBB_sym1	AAGTCCTCTTTCAGGGACCCATGGCAAATAGCAGCGTTGAACTGCGTGTTGCCGAAGCATATCCGGAAGATGTTGGTCGTGGTATTGTTCGTATGGATAAACA GACCCGTGCAAAACTGGGTGTTAGCGTTGGTGATTATGTGGAAGTTAAAAAAGTGGACAGCTCAGTGGAACTGCGCGTGGCAGAGGCCTATCCTGAGGATGTA GGCCGTGGCATCGTGCGCATGGACAAACAAACACGTGCCAAATTAGGTGTTTCAGTGGGCGATTATGTTGAGGTGAAAAAGGTCGATTAAGGATCCGAATTCT GTACAGG
apDPBB_sym2	AAGTCCTCTTTCAGGGACCCATGGCAAATCGTAGCGTTGTTGCACGTGTTGCCGAAGCATATCCGGAAGATGTTGGTCGTAAAATTGTGCGCATGGATAAATA TCTGCGTGCAAAACTGGGTGTTAGCGTTGGTGATTATGTGGAAGTTAAAAAAGTTGAACGTAGTGTGGTTGCCCGTGTGGCAGAGGCCTATCCTGAGGATGTG GGACGCAAAATCGTTCGTATGGATAAGTATTTAAGAGCCAAACTGGGCGTTTCAGTGGGCGATTATGTTGAGGTGAAAAAGGTGGAATAAGGATCCGAATTCT GTACAGG
Symmetric DPBBs designed using reverse evolution engineering method (RE-design)	
reDPBB_sym1	AAGTCCTCTTTCAGGGACCCCCGATTAAACTGCGTGTTATGGAAGCATATCCGGAAGATGTTGGTAAAGGTATTGTGCGTATGGATAAAGCAAGCCGTGATAA ACTGGGTGTTTCAGCCGGTGATCTGGTTGAAATTAAAGGTAGCAAAACCCCTATCAAACTTCGCGTGATGGAAGCGTACCCTGAGGATGTAGGCAAAGGCATC GTTCGCATGGACAAAGCCTCACGTGATAAATTAGGTGTGAGCGCAGGCGACCTGGTGGAAATCAAAGGCTAAGGATCCGAATTCTGTACAGG
reDPBB_sym2	AAGTCCTCTTTCAGGGACCCCCGATGAAACTGCGTGTTATGGAAGCATATCCGGAAGATGTTGGTAAAGGTATTGTGCGTATGGATAAAGCAAGCCGTGAAAA ACTGGGTGTTAGTGCCGGTGATCTGGTTGAAATTAAAGGTAGCAAAACCCCGATGAAATTAAGAGTGATGGAAGCGTACCCTGAGGATGTAGGCAAAGGCATC GTTCGCATGGACAAAGCATCACGCGAGAAACTGGGCGTTTCAGCAGGCGACCTGGTGGAAATCAAAGGCTAAGGATCCGAATTCTGTACAGG
reDPBB_sym3	AAGTCCTCTTTCAGGGACCCACCATTAAACTGCGTGTTATGGAAGCATATCCGGAAGATGTTGGTAAAGGTATTGTGCGTATGGATAAAGCAAGCCGTGATAA AATTGGTGTTAGTGCCGGTGATCTGGTGGAAATTAAAGGTAGCAAAACCACAATCAAACTTCGCGTGATGGAAGCGTACCCTGAGGATGTAGGCAAAGGCATC GTTCGCATGGACAAAGCCTCACGCGATAAAATCGGCGTTTCAGCAGGCGATTTAGTTGAAATCAAAGGCTAAGGATCCGAATTCTGTACAGG

reDPBB_sym4	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAGCGTTGTTGCACGTGTTGCACCGGCACATCCGGAAGATGTTGGTAAAGGTATTGTTCGCATGGATAAATA TGAACGCCAGAATCTGGGTGTTAGCGTTGGTGATTATGTGGAAGTTAAAAAAGCCAAATCAGTGGTTGCTCGCGTTGCCCCTGCGCACCCTGAAGATGTAGGC AAAGGCATCGTGAGAATGGACAAATACGAGCGTCAGAACCTGGGCGTTTCAGTGGGCGATTATGTTGAGGTGAAAAAGGCATAAGGATCCGAATTCTGTACAG G
reDPBB_sym5	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAGCGTTGTTGCACGTGTTGCACCGGCATATCCGGAAGATGTTGGTAAAGGTATTGTGCGCATGGATAAATA TGAACGTGCAAATCTGGGTGTTAGCGTTGGTGATTATGTTGAAGTTGATAAAGCCAAATCAGTGGTTGCTCGCGTTGCCCCTGCCTATCCTGAGGATGTAGGC AAAGGCATCGTTCGTATGGACAAATACGAGCGTGCCAACCTGGGCGTTTCAGTGGGCGATTACGTGGAAGTGGACAAAGCATAAGGATCCGAATTCTGTACAG G
reDPBB_sym6	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAACCGTTGTTGCACGTGTTCTGCCTGCATATCCGGAAGATGTTGGTAAAGGTATTGTGCGCATGGATAAATA TGAACGTGCAAAACTGGGTGTTAGCGTTGGTGATTATGTTGAAGTTGAAAAAGCCAAAACCGTGGTGGCTCGCGTGCTGCCAGCCTATCCTGAGGATGTAGGC AAAGGCATCGTTCGTATGGACAAATACGAGCGTGCCAAATTAGGTGTTTCAGTGGGCGATTACGTGGAAGTGGAAAAGGCATAAGGATCCGAATTCTGTACAG G
reDPBB_sym7	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAGCGTTGTTGCACGTGTTGCACCGGCATATCCGGAAGATGTTGGTAAAGGTATTGTGCGCATGGATAAATA TGAACGTGCAAAACTGGGTGTTAGCGTTGGTGATTATGTTGAAGTTGAAAAAGCCAAATCAGTGGTTGCTCGCGTTGCCCCTGCCTATCCTGAGGATGTAGGC AAAGGCATCGTTCGTATGGACAAATACGAGCGTGCCAAATTAGGTGTTTCAGTGGGCGATTACGTGGAAGTGGAAAAGGCATAAGGATCCGAATTCTGTACAG G
Symmetric DPBBs designed using multi-state design method (MS-design)	
msDPBB_sym1	AAGTCCTCTTTCAGGGACCCAGCAGCGTTGTTGCACGTGTTGCACTGGCACATGAAGATGATGTTGGTAAAAACATTGTGCGCATGGATGAAGATCTGATGCG TAAACTGGGTGTTAAAGTTGGTGATTACGTGGAAATCATGAAGAAAAGCTCAGTGGTTGCCCGTGTGGCCTTAGCGCACGAGGATGATGTGGGCAAAAATATC GTTCGTATGGACGAAGATTTAATGAGAAAACTGGGCGTGAAAGTGGGCGACTATGTTGAAATTATGAAAAAATGAGGATCCGAATTCTGTACAGG
msDPBB_sym2	AAGTCCTCTTTCAGGGACCCAGCAGCGTTATTGCACGTGTTGCACTGGCACATGAAGATGATGTTGGTAAAAACATTGTGCGCATGGATGAAGAACTGATGCG TCTGCTGGGTGTTAAAGTTGGTGATCTGGTTGAAATCATGAAAGTGAGCAGTGTGATTGCCCGTGTGGCCTTAGCGCACGAGGATGATGTGGGCAAAAATATC GTTCGTATGGACGAAGAATTAATGCGCCTGTTAGGCGTGAAAGTGGGCGACCTGGTGGAAATTATGAAGGTTTAAGGATCCGAATTCTGTACAGG
Simplified mk2h	
mk2h_ΔM	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAGCGTTGTTGCACGTGTTGCCGAAGCATATCCGGAAGATGTTGGTAAACGTATTGTTCGCGTGGATAAATA TGAACGTGCAAAACTGGGTGTTAGCGTTGGTGATTATGTGGAAGTTAAAAAAGTTAAAAGTGTGGTTGCCCGTGTGGCAGAGGCCTATCCTGAGGATGTGGGC AAACGCATTGTGCGTGTTGACAAATACGAGCGTGCCAAATTAGGTGTTTCAGTGGGCGATTATGTTGAGGTGAAAAAGGTGTAAGGATCCGAATTCTGTACAG G
mk2h_ΔI	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAGCGTTGTTGCACGTGTTGCCGAAGCATATCCGGAAGATGTTGGTAAACGTGTTGTTCGCATGGATAAATA TGAACGTGCAAAACTGGGTGTTAGCGTTGGTGATTATGTGGAAGTTAAAAAAGTTAAAAGTGTGGTTGCCCGTGTGGCAGAGGCCTATCCTGAGGATGTGGGC AAACGCGTGGTGCGTATGGACAAATACGAGCGTGCCAAATTAGGTGTTTCAGTGGGCGATTATGTTGAGGTGAAAAAGGTGTAAGGATCCGAATTCTGTACAG G
mk2h_ΔL	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAGCGTTGTTGCACGTGTTGCCGAAGCATATCCGGAAGATGTTGGTAAACGTATTGTGCGCATGGATAAATA TGAACGTGCAAAAGTTGGTGTTAGCGTGGGTGATTATGTGGAAGTTAAAAAAGTTAAAAGTGTGGTTGCCCGTGTGGCAGAGGCCTATCCTGAGGATGTGGGC AAACGCATCGTTCGTATGGACAAATACGAGCGTGCCAAGGTGGGTGTTTCAGTTGGCGATTATGTTGAGGTTAAAAAGGTGTAAGGATCCGAATTCTGTACAG G
mk2h_ΔP	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAGCGTTGTTGCACGTGTTGCCGAAGCATATGCAGAAGATGTTGGTAAACGTATTGTGCGCATGGATAAATA TGAACGTGCAAAACTGGGTGTTAGCGTTGGTGATTATGTGGAAGTTAAAAAAGTTAAAAGTGTGGTTGCCCGTGTGGCAGAGGCCTATGCCGAGGATGTGGGC AAACGCATCGTTCGTATGGACAAATACGAGCGTGCCAAATTAGGTGTTTCAGTGGGCGATTATGTTGAGGTGAAAAAGGTGTAAGGATCCGAATTCTGTACAG G
mk2h_ΔS	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAAAGTTGTTGCACGTGTTGCCGAAGCATATCCGGAAGATGTTGGTAAACGTATTGTGCGCATGGATAAATA TGAACGTGCAAAACTGGGTGTGAAAGTGGGTGATTATGTTGAAGTGAAAAAGGTGAAAAAAGTGGTGGCACGCGTGGCAGAGGCCTATCCTGAGGATGTGGGC AAACGCATCGTTCGTATGGACAAATACGAGCGTGCCAAATTAGGCGTTAAAGTTGGCGATTACGTGGAAGTCAAAAAAGTGTAAGGATCCGAATTCTGTACAG G
mk2h_ΔY	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAGCGTTGTTGCACGTGTTGCCGAAGCACGTCCGGAAGATGTTGGTAAACGTATTGTTCGTATGGATAAAGC AGAACGTGCAAAACTGGGTGTTAGCGTTGGTGATGTTGTTGAGGTTAAAAAAGTTAAAAGTGTGGTTGCCCGTGTGGCAGAGGCTCGCCCTGAAGATGTGGGC AAACGCATCGTGAGAATGGACAAAGCCGAGCGTGCCAAATTAGGTGTTTCAGTGGGTGATGTGGTGGAAGTGAAAAAGGTTTAAGGATCCGAATTCTGTACAG G
mk2h_ΔMIL	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAGCGTTGTTGCACGTGTTGCCGAAGCATATCCGGAAGATGTTGGTAAACGTGTTGTTCGCGTTGATAAATA TGAACGTGCAAAAGTTGGTGTTAGCGTGGGTGATTATGTGGAAGTCAAAAAAGTTAAAAGTGTGGTTGCCCGTGTGGCAGAGGCCTATCCTGAGGATGTGGGC AAACGCGTTGTGCGTGTGGACAAATACGAGCGTGCCAAAGTGGGCGTTTCAGTTGGCGATTATGTTGAGGTTAAAAAGGTGTAAGGATCCGAATTCTGTACAG G
mk2h_ΔMILPS	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAAAGTTGTTGCACGTGTTGCCGAAGCATATGCCGAAGATGTTGGTAAACGTGTTGTTCGCGTTGATAAATA TGAACGTGCAAAAGTTGGTGTGAAAGTGGGTGATTATGTGGAAGTGAAAAAGGTGAAAAAAGTGGTGGCACGCGTGGCAGAGGCCTATGCAGAGGATGTGGGC AAACGCGTTGTGCGTGTGGACAAATACGAGCGTGCCAAAGTGGGCGTTAAAGTTGGCGATTATGTTGAGGTTAAGAAAGTGTAAGGATCCGAATTCTGTACAG G
mk2h_ΔMILPY	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAGCGTTGTTGCACGTGTTGCCGAAGCACGTGCCGAAGATGTTGGTAAACGTGTTGTTCGTGTTGATAAAGC AGAACGTGCAAAAGTTGGTGTTAGCGTTGGTGATGTTGTGGAAGTGAAAAAGGTGAAATCAGTGGTTGCCCGTGTGGCAGAGGCACGCGCAGAGGATGTGGGC AAACGTGTGGTGCGCGTGGACAAAGCCGAGCGTGCCAAAGTGGGCGTTTCAGTGGGTGATGTGGTTGAGGTTAAAAAGGTGTAAGGATCCGAATTCTGTACAG G
mk2h_ΔMILYS	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAAAGTTGTTGCACGTGTTGCCGAAGCACGTCCGGAAGATGTTGGTAAACGTGTTGTTCGTGTTGATAAAGC AGAACGTGCAAAAGTTGGTGTTAAAGTGGGTGATGTTGTGGAAGTGAAAAAGGTGAAAAAAGTGGTGGCACGCGTGGCAGAGGCTCGCCCTGAAGATGTGGGC AAACGCGTGGTGCGCGTGGATAAAGCCGAGCGTGCCAAGGTGGGCGTGAAAGTTGGAGATGTGGTTGAAGTTAAAAAAGTGTAAGGATCCGAATTCTGTACAG G
mk2h_ΔMILPYS	AAGTCCTCTTTCAGGGACCCATGCCTGGTAAAAAAGTTGTTGCACGTGTTGCCGAAGCACGTGCCGAAGATGTTGGTAAACGTGTTGTTCGTGTTGATAAAGC AGAACGTGCAAAAGTTGGTGTTAAAGTGGGTGATGTTGTGGAAGTGAAAAAGGTGAAAAAAGTGGTGGCACGCGTGGCAGAGGCTCGTGCAGAGGATGTGGGC AAACGCGTGGTGCGCGTGGATAAAGCCGAGCGTGCCAAGGTGGGCGTGAAAGTTGGAGATGTGGTTGAAGTTAAAAAAGTGTAAGGATCCGAATTCTGTACAG G

Primers	
Cloning_upstream	AAGTCCTCTTTCAGGGACCC
Cloning_downstream	CCTGTACAGAATTCGGATCC
mk1/2_half	CCTGTACAGAATTCGGATCCTTAAACTTTTTTAACTTCCAC
ap1_half	CCTGTACAGAATTCGGATCCTTAGTCCACTTTTTTAACTTCCAC
ap2_half	CCTGTACAGAATTCGGATCCTTATTCAACTTTTTTAACTTCCAC
mk2_ΔMIL_half	CCTGTACAGAATTCGGATCCTTAAACTTTTTTGACTTCCACATAATCAC
mk2_ΔY_half	CCTGTACAGAATTCGGATCCTTAAACTTTTTTAACCTCAACAAC
mk2_ΔS_half	CCTGTACAGAATTCGGATCCTTACACCTTTTTCACTTCAACA
mk2_ΔP/M/I/L_half	CCTGTACAGAATTCGGATCCTTAAACTTTTTTAACTTCCACATAATCAC
mk2_ΔMILPS/MILPY/MILYS/MILPYS_half	CCTGTACAGAATTCGGATCCTTACACCTTTTTCACTTCCAC

Table S12. Summary of crystallization methods.

Protein	Concentration (mg/mL)	Reservoir solution (RS)	Cryo solution	Structure solving method	MR model	Beam source	
						Institute	beamline
taVCP_DPBB	12.94	100mM citrate/ phosphate pH5.0, 1.4M Ammonium sulfate	RS + 20% glycerol	SAD (Se)	-	Spring-8	BL26B2
mkVCP_DPBB	21.09	100mM imidazole/HCl pH7.8, 2M NaCl, 200mM Zinc acetate	RS + 20% glycerol	SAD (Zn)	-	Photon factory	NW12A
apVCP_DPBB	34.05	100mM Sodium acetate pH4.5, 20% PEG1000, 200mM Zinc acetate	RS + 20% glycerol	SAD (Zn)	-	Spring-8	BL26B2
apDPBB_sym79	20.28	100mM Tris-HCl, pH8.5, 25%PEG3350, 200mM Lithium sulfate	RS + 20% glycerol	MR	mkVCP_DPBB	Spring-8	BL26B2
mkDPBB_sym86	19.09	100mM HEPES pH7.0, 15% PEG20000	RS + 20% glycerol	MR	mkVCP_DPBB	Spring-8	BL26B2
mkDPBB_sym1	62.57	2100mM D/L-malic acid pH7.0	RS + 20% glycerol	MR	mkDPBB_sym2	Photon factory	NW12A
mkDPBB_sym2	49.54	100mM CHES pH9.5, 1M sodium citrate tribasic	RS + 20% glycerol	MR	mkVCP_DPBB	Photon factory	NW12A
reDPBB_sym1	76.67	100mM Sodium acetate pH3.8, 2M NaCl, 400mM Lithium sulfate	RS + 30% glycerol	MR	Model structure	Spring-8	BL26B2
reDPBB_sym2	47.52	100mM Sodium acetate pH6.2, 35% MPD, 20% PEG1500	RS + 5% glycerol	MR	mkDPBB_sym2	Spring-8	BL26B2
reDPBB_sym4	17.42	200mM Sodium citrate tribasic, 20% PEG3350	RS + 30% glycerol	MR	mkDPBB_sym2	Spring-8	BL32XU
msDPBB_sym2	12.03	100mM CAPS pH10.5, 2M Ammonium sulfate, 200mM Lithium sulfate	RS + 13 % glycerol	MR	mkDPBB_sym2	SLS	X06SA
ap1h	39.85	25% PEG4000, 170mM Ammonium sulfate, 15% glycerol	RS + 15% glycerol (total 30%)	MR	mk2h	Photon factory	BL5A
mk2h	23.75	1200mM Sodium phosphate monobasic/800mM Pottasium phosphate dibasic, 100mM CAPS pH 10.5, 100mM Lithium sulfate	RS + 20% glycerol	MR	mkDPBB_sym2	Spring-8	BL26B2
mk2h(Synthetic peptide)	16.32	800mM Sodium phosphate monobasic/1200mM Pottasium phosphate dibasic, 100mM Sodium acetate pH 4.5	RS + 30% glycerol	MR	mkDPBB_sym2	Photon factory	BL5A
mk2h_ΔP	21.16	2000mM Ammonium sulfate, 100mM CAPS/ Sodium hydroxide, pH 10.5, 200mM Lithium sulfate	RS + 13% glycerol	MR	mk2h	Spring-8	BL26B2
mk2h_ΔY	11.34	100mM SPG buffer, 25% PEG 1500	RS + 30% glycerol	MR	mk2h	Spring-8	BL26B2
mk2h_ΔMIL	24.02	100mM Imidazole/ Hydrochloric acid pH8.0, 200mM Lithium sulfate, 10% (w/v) PEG 3000	RS + 30% glycerol	MR	mk2h	Spring-8	BL26B2
mk2h_ΔMILPS	13.85	2000mM Sodium malonate dibasic	RS + 30% glycerol	MR	mk2h_ΔMIL	SLS	X06SA
mk2h_ΔMILPS(Synthetic peptide)	12.51	400mM Sodium phosphate monobasic/1600mM Pottasium phosphate dibasic, 100mM imidazole_HCl, pH8.0, 200mM NaCl	RS + 30% glycerol	MR	mk2h_ΔMILPS	SLS	X06SA
mk2h_ΔMILPYS(Synthetic peptide, malonate)	10.32	3000mM Sodium malonate dibasic	RS + 20% glycerol	MR	mk2h_ΔMILPS	Spring-8	BL26B2
mk2h_ΔMILPYS(Synthetic peptide, malic acid)	7.62	2100mM D/L-malic acid pH7.0	RS + 30% glycerol	MR	mk2h_ΔMILPYS	Spring-8	BL26B2

References

1. R. Furukawa, M. Nakagawa, T. Kuroyanagi, S. I. Yokobori, A. Yamagishi, Quest for Ancestors of Eukaryal Cells Based on Phylogenetic Analyses of Aminoacyl-tRNA Synthetases. *J. Mol. Evol.* **84**, 51–66 (2017).
2. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang,,W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
3. A. R. D. Voet, H. Noguchi, C. Addy, D. Simoncini, D. Terada, S. Unzai, S. Y. Park, K. Y. J. Zhang, J. R. H. Tame, Computational design of a self-assembling symmetrical β-propeller protein. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15102–15107 (2014).
4. H. Noguchi, C. Addy, D. Simoncini, S. Wouters, B. Mylemans, L. Van Meervelt, T. Schiex, K. Y. J. Zhang, J. R. H. Tame, A. R. D. Voet, Computational design of symmetrical eight-bladed β-propeller proteins. *IUCrJ.* **6**, 46–55 (2019).
5. D. Terada, A. R. D. Voet, H. Noguchi, K. Kamata, M. Ohki, C. Addy, Y. Fujii, D. Yamamoto, Y. Ozeki, J. R. H. Tame, K. Y. J. Zhang, Computational design of a symmetrical β-trefoil lectin with cancer cell binding activity. *Sci. Rep.* **7**, 1–13 (2017).
6. S. Whelan, N. Goldman, A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699 (2001).
7. D. T. Jones, W. R. Taylor, J. M. Thornton, The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*. **8**, 275–282 (1992).
8. H. Ashkenazy, O. Penn, A. Doron-Faigenboim, O. Cohen, G. Cannarozzi, O. Zomer, T. Pupko, FastML: A web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* **40**, 580–584 (2012).
9. A. R. D. Voet, D. Simoncini, J. R. H. Tame, K. Y. J. Zhang, Evolution-inspired computational design of symmetric proteins. *Methods Mol. Biol.* **1529**, 309–322 (2017).
10. F. DiMaio, A. Leaver-Fay, P. Bradley, D. Baker, I. André, Modeling symmetric macromolecular structures in Rosetta3. *PLoS One*. **6** (2011), doi:10.1371/journal.pone.0020450.
11. S. Chaudhury, S. Lyskov, J. J. Gray, PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*. **26**, 689–691 (2010).
12. J. Vucinic, D. Simoncini, M. Ruffini, S. Barbe, T. Schiex, Positive multistate protein design. *Bioinformatics*. **36**, 122–130 (2020).
13. F. Berenger, R. Shrestha, Y. Zhou, D. Simoncini, K. Y. J. Zhang, Durandal: Fast exact clustering of protein decoys. *J. Comput. Chem.* **33**, 471–474 (2012).
14. R. M. Kramer, V. R. Shende, N. Motl, C. N. Pace, J. M. Scholtz, Toward a molecular understanding of protein solubility: Increased negative surface charge correlates with increased solubility. *Biophys. J.* **102**, 1907–1915 (2012).
15. S. Jones, J. M. Thornton, Review Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 13–20 (1996).
16. J. Janin, S. Miller, C. Chothia, Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* **204**, 155–164 (1988).
17. F. Chiti, M. Stefani, N. Taddei, G. Ramponi, C. M. Dobson, Rationalization of mutational effects on protein aggregation rates. *Nature*. **424**, 805–808 (2003).
18. D. Simoncini, T. Schiex, K. Y. J. Zhang, Balancing exploration and exploitation in population-based sampling improves fragment-based *de novo* protein structure prediction. *Proteins.* **85**, 852–858 (2017).
19. J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
20. D. A. Case, R. M. Betz, D. S. Cerutti, T. Cheatham, T. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Götz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. S. Lee,

S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. M. Merz, G. Monard, H. Nguyen, H. T. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, C. Sagui, C. L. Simmerling, W. M. Botello-Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, L. Xiao, P. A. Kollman, Amber 16. *University of California, San Francisco.* (2016).

21. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. Dinola, J. R. Haak, Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
22. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
23. J. P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
24. D. R. Roe, T. E. Cheatham, PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
25. Y. Yamada, N. Matsugaki, L. M. G. Chavas, M. Hiraki, N. Igarashi, S. Wakatsuki, Data management system at the photon factory macromolecular crystallography beamline. *J. Phys. Conf. Ser.* **425** (2013), doi:10.1088/1742-6596/425/1/012017.
26. M. Hiraki, Y. Yamada, L. M. G. Chavas, S. Wakatsuki, N. Matsugaki, Improvement of an automated protein crystal exchange system PAM for high-throughput data collection. *J. Synchrotron Radiat.* **20**, 890–893 (2013).
27. G. Ueno, H. Kanda, R. Hirose, K. Ida, T. Kumasaka, M. Yamamoto, RIKEN structural genomics beamlines at the SPring-8; high throughput protein crystallography with automated beamline operation. *J. Struct. Funct. Genomics*. **7**, 15–22 (2006).
28. S. Ito, G. Ueno, M. Yamamoto, DeepCentering: fully automated crystal centering using deep learning for macromolecular crystallography. *J. Synchrotron Radiat.* **26**, 1361–1366 (2019).
29. N. Okazaki, K. Hasegawa, G. Ueno, H. Murakami, T. Kumasaka, M. Yamamoto, Mail-in data collection at SPring-8 protein crystallography beamlines. *J. Synchrotron Radiat.* **15**, 288–291 (2008).
30. H. Murakami, G. Ueno, N. Shimizu, T. Kumasaka, M. Yamamoto, Upgrade of automated sample exchanger SPACE. *J. Appl. Crystallogr.* **45**, 234–238 (2012).
31. K. Hirata, K. Yamashita, G. Ueno, Y. Kawano, K. Hasegawa, T. Kumasaka, M. Yamamoto, Zoo: An automatic data-collection system for high-throughput structure analysis in protein microcrystallography. *Acta Crystallogr. Sect. D Struct. Biol.* **75**, 138–150 (2019).
32. Y. Nakamura, S. Baba, N. Mizuno, T. Irie, G. Ueno, K. Hirata, S. Ito, K. Hasegawa, M. Yamamoto, T. Kumasaka, Computer-controlled liquid-nitrogen drizzling device for removing frost from cryopreserved crystals. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **76**, 616–622 (2020).
33. W. Kabsch, XDS. *Acta Crystallogr. D. Biol. Crystallogr.* **66**, 125–32 (2010).
34. P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221 (2010).
35. T. C. Terwilliger, P. D. Adams, R. J. Read, A. J. McCoy, N. W. Moriarty, R. W. Grosse-Kunstleve, P. V. Afonine, P. H. Zwart, L. W. Hung, Decision-making in structure solution using Bayesian estimates of map quality: The PHENIX AutoSol wizard. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **65**, 582–601 (2009).
36. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

37. H. Xue, K. L. Tong, C. Marck, H. Grosjean, J. T. F. Wong, Transfer RNA paralogs: Evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life. *Gene*. **310**, 59–66 (2003).
38. Z. Yu, K. Takai, A. Slesarev, H. Xue, J. T. F. Wong, Search for primitive methanopyrus based on genetic distance between Val- and Ile-tRNA synthetases. *J. Mol. Evol.* **69**, 386–394 (2009).
39. J. Nölling, A. Elfner, J. R. Palmer, V. J. Steigerwald, T. D. Pihl, J. A. Lake, J. N. Reeve, Phylogeny of *Methanopyrus kandleri* based on methyl coenzyme M reductase operons. *Int. J. Syst. Bacteriol.* **46**, 1170–1173 (1996).
40. S. Akanuma, Y. Nakajima, S. I. Yokobori, M. Kimura, N. Nemoto, T. Mase, K. I. Miyazono, M. Tanokura, A. Yamagishi, Experimental evidence for the thermophilicity of ancestral life. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11067–11072 (2013).
41. M. N. Nguyen, K. P. Tan, M. S. Madhusudhan, CLICK - Topology-independent comparison of biomolecular 3D structures. *Nucleic Acids Res.* **39**, 24–28 (2011).
42. M. N. Nguyen, M. S. Madhusudhan, Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Res.* **39** (2011), doi:10.1093/nar/gkr348.
43. R. M. Castillo, K. Mizuguchi, V. Dhanaraj, A. Albert, T. L. Blundell, A. G. Murzin, A six-stranded double-psi β barrel is shared by several protein superfamilies. *Structure*. **7**, 227–236 (1999).
44. Z. F. Burton, K. Opron, G. Wei, J. H. Geiger, A model for genesis of transcription systems. *Transcription*. **7**, 1–13 (2016).