Article

A backbone-centred energy function of neural networks for protein design

https://doi.org/10.1038/s41586-021-04383-5

Received: 5 March 2021

Accepted: 23 December 2021

Published online: 09 February 2022

Check for updates

Bin Huang[1,4], Yang Xu[1,4], Xiuhong Hu[1,4], Yongrui Liu[1], Shanhui Liao[1], Jiahai Zhang[1], Chengdong Huang[1,2], Jingjun Hong[1], Quan Chen[1,2]✉ & Haiyan Liu[1,2,3]✉

A protein backbone structure is designable if a substantial number of amino acid sequences exist that autonomously fold into it[1,2]. It has been suggested that the designability of backbones is governed mainly by side chain-independent or side chain type-insensitive molecular interactions[3–5], indicating an approach for designing new backbones (ready for amino acid selection) based on continuous sampling and optimization of the backbone-centred energy surface. However, a sufficiently comprehensive and precise energy function has yet to be established for this purpose. Here we show that this goal is met by a statistical model named SCUBA (for Side Chain-Unknown Backbone Arrangement) that uses neural network-form energy terms. These terms are learned with a two-step approach that comprises kernel density estimation followed by neural network training and can analytically represent multidimensional, high-order correlations in known protein structures. We report the crystal structures of nine de novo proteins whose backbones were designed to high precision using SCUBA, four of which have novel, non-natural overall architectures. By eschewing use of fragments from existing protein structures, SCUBA-driven structure design facilitates far-reaching exploration of the designable backbone space, thus extending the novelty and diversity of the proteins amenable to de novo design.

Computational protein design has exhibited enormous potential, with ground-breaking studies demonstrating the design of de novo proteins with new structures and functions[6–13], most of which were carried out using the state-of-the-art RosettaDesign method[6,13,14]. New backbones were built either by parametrically varying relative geometries between existing structural modules (or templates) to design helix bundles[11,15,16] or repeat proteins[17] or by assembling peptide fragments from existing structures[6,8]. Despite recent improvements[18,19], the template dependence of these approaches for generating backbones still severely restricts the available spectrum of possible new structures[13,20,21], potentially narrowing the scope of the functional activities amenable to design.

An explicit representation of the presumed backbone-centred energy surface that determines designability may provide a basis for a template-free protein design workflow (Fig. 1a); this would be fundamentally distinct from—and may substantially complement—existing methods. Progress to establish such a representation has been slow, probably owing to a lack of methods to represent the relevant molecular interactions to the level of comprehensiveness and precision needed for de novo protein design tasks. Earlier studies have explored simplified backbone energy surfaces, but only to verify the presence of natural backbone-like broad minima[4,22,23], that is, not in the context of attempting backbone design. One exception was a Cα-atom-based statistical potential that emphasized the accurate modelling of local backbone conformations[24]. This model successfully produced a de novo loop designed without using any natural fragment as template[25], indicating the viability of backbone-centred approaches.

The SCUBA (Side Chain-Unknown Backbone Arrangement) model aims to represent factors essential for backbone designability, including the local conformational preferences and hydrogen-bonding geometries of peptide backbones and inter-backbone space required for chirally attached and tightly packed side chains[3–5], doing so with comprehensiveness and precision that support de novo protein design. To achieve this, we represent the various interactions using statistical energy terms, or potentials, trained with a general approach named NC-NN (Fig. 1b–d), with this name denoting a two-step process of first estimating statistical energy values from raw structural data by kernel-based density estimation (that is, neighbour counting) followed by training of neural networks (fully connected three-layer perceptrons; Fig. 1d and Supplementary Methods) to represent the potentials. NC-NN addresses a main technical challenge in constructing statistical potentials, and the resulting potentials, in addition to being continuous and providing easily computable function values (and derivatives) for use in structure sampling and optimization, can represent the complex, high-dimensional and highly correlated distributions of real structural data with high fidelity.

We applied SCUBA-driven stochastic dynamics (SD) simulations[26] together with our data-driven fixed-backbone amino acid sequence

[1]MOE Key Laboratory for Membraneless Organelles and Cellular Dynamics, Hefei National Laboratory for Physical Sciences at the Microscale, School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China. [2]Biomedical Sciences and Health Laboratory of Anhui Province, University of Science and Technology of China, Hefei, China. [3]School of Data Science, University of Science and Technology of China, Hefei, China. [4]These authors contributed equally: Bin Huang, Yang Xu, Xiuhong Hu. ✉e-mail: chenquan@ustc.edu.cn; hyliu@ustc.edu.cn

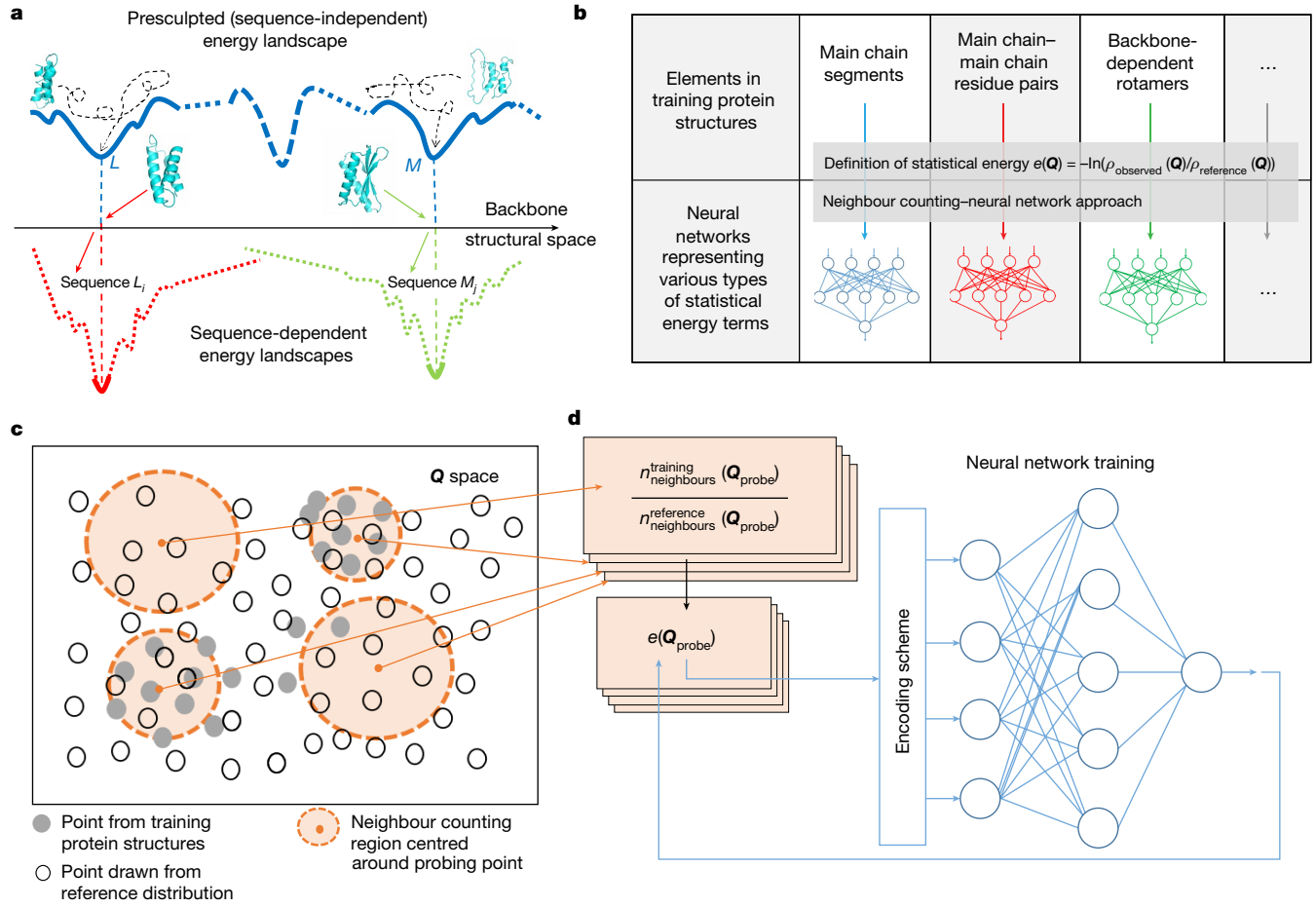


Fig. 1 | Template-free protein design facilitated by explicit representation of the backbone-centred energy landscape. a, The upper blue curve represents the backbone-centred energy landscape, on which the low-lying minima *L* and *M* may closely reproduce the global minima of sequence-specific free energy landscapes (represented by the lower red and green curves for sequences L_i and M_j, respectively). The designable structures *L* and *M* can be found by sampling and optimization starting from initial structures constructed according to user-defined design specifications ('sketches'). Only the solid parts of the energy landscape curves need to be explored during design. **b**, The backbone-centred effective energy function comprises neural networks learned from natural protein structures. Each neural network is trained using the neighbour counting–neural network (NC-NN) approach. **c**, **d**, Principles of NC-NN. Multiple circles corresponding to neighbour counting are shown in **c** to indicate that the single-point energies of many different probing $\boldsymbol{Q}$ points are estimated (separately) by NC. These circles are of different diameter to indicate that the radii of the neighbour-counting kernels can be chosen adaptively according to local distributions of the NC training data.

selection program ABACUS2 (refs [27,28]) to design proteins with topological architectures that fulfil various design specifications or that are novel and were designed from scratch through computational exploration of the designable backbone space using SCUBA. We report the X-ray structures of nine de novo proteins, each with a uniquely designed sequence and structure.

Building statistical potentials with NC-NN

Formally, a statistical potential that depends on a set of structural variables (denoted by $\boldsymbol{Q}$) can be defined as $e(\boldsymbol{Q}) = -\ln(\rho_{observed}(\boldsymbol{Q})/\rho_{reference}(\boldsymbol{Q}))$, where $\rho_{observed}(\boldsymbol{Q})$ is the probability density of the observed data distributed in $\boldsymbol{Q}$ space and $\rho_{reference}(\boldsymbol{Q})$ is the expected probability density of the same variables for an idealized reference system (that is, a system lacking interaction $e(\boldsymbol{Q})$). The workflow of neighbour counting (NC) followed by neural network training (NN) for learning $e(\boldsymbol{Q})$ is illustrated in Fig. 1c, d.

In the NC step, for any selected probing point $\boldsymbol{Q}_{probe}$ in $\boldsymbol{Q}$ space (represented by the box in Fig. 1c), its single-point energy $e(\boldsymbol{Q}_{probe})$ is estimated as the ratio between the number of neighbouring observed data points $n^{observed}_{neighbours}(\boldsymbol{Q}_{probe})$ and the number of neighbouring reference data points $n^{reference}_{neighbours}(\boldsymbol{Q}_{probe})$. By using observed data points from real structures and reference data points drawn computationally, we can estimate the two numbers using a kernel method. The radius of the kernel can be chosen adaptively on the basis of the local density of the training data, such that the trade-off between the resolution and statistical uncertainty of the estimated energies can be balanced.

In the NN step (Fig. 1d), NC-estimated energies at different probing points are used to train a neural network to encode the statistical potential as an analytical function of $\boldsymbol{Q}$. As the NC step only estimates single-point energies in a time-consuming way and yields discrete and noisy values, the NN step is indispensable for obtaining an analytical potential that can be computed efficiently without referring to the original data.

A main advantage of NC-NN over traditional approaches to learning statistical potentials based on placing features into bins is that NC-NN can faithfully retain correlations in high-dimensional space. For example, Extended Data Fig. 1 shows low-dimensional projections of an NC-NN-learned SCUBA energy term depending on 14 variables. The data clearly exhibit strong cross-dimension correlations as captured by NC-NN, thus distinguishing observed (mostly favourable) from computationally drawn (mostly unfavourable) configurations. For both low- and high-dimensional cases, the NC-NN-learned potentials are (1) automatically continuous and (2) of analytical gradients, which makes them suitable for gradient-driven structure sampling and optimization.

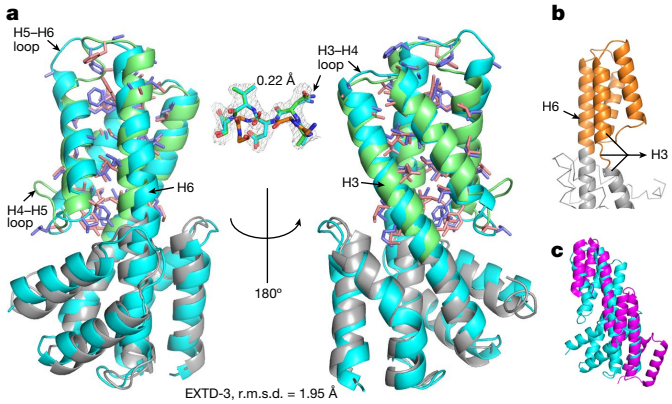


Fig. 2 | The de novo protein EXTD-3 integrates pre-existing and newly designed parts to form a single rigid architecture not yet observed in nature. a, The experimentally solved structure (cyan) superimposed on the designed structure. The pre-existing part (tepsin ENTH domain; PDB, 5WF2) is shown in grey; the newly designed part is shown in green. The r.m.s.d. displayed is for main chain atoms. Internal side chains of the newly designed part are shown as sticks (blue, experimentally observed; orange, designed). The superimposed structures of the H3–H4 loop are shown in the middle (orange, initial backbone; light blue, experimentally determined backbone; green, designed backbone) with the 2*Fo* – *Fc* (at 1.0σ level) electron density surface. **b**, The initial backbone for EXTD-3 (grey, pre-existing part; orange, newly designed part); H3 in the final optimized backbone exists as a helix–loop–helix segment. **c**, EXTD-3 (cyan) aligned with its architecturally closest natural protein (PDB, 5H79; chain D in purple), whose backbone is matched to both the pre-existing and newly designed parts of EXTD-3 with a Dali *z* score of 4.3.

We note that, although NC-NN is built on a previous idea of representing complex energy surfaces with neural networks[29], the use of raw structural data to obtain single-point energies is fundamentally distinct from the use of quantum mechanical calculations.

The backbone-centred SCUBA model

SCUBA was devised to comprehensively consider the dependence of the backbone-centred energy landscape on various structural variables, using five types of NC-NN-learned energy terms (Extended Data Fig. 2a). The training data comprised more than 12,000 non-redundant natural structures[27,30] from the Protein Data Bank (PDB)[31]. The neural network terms were combined with the usual covalent and steric terms to form an effective energy function that analytically depends on atomic Cartesian coordinates (see the Supplementary Methods for details). As the negative gradients of the energy function (with respect to coordinates) can be used as forces, SCUBA can be used to drive Newtonian or Langevin molecular dynamics (MD) or SD simulations of protein structures.

Although developed to be backbone centred, SCUBA contains both backbone-dependent and explicit side chain-dependent terms. Side chains were included because we found 'blurred' positions for backbone atoms when no side chains were specified (which ultimately yielded alanine or glycine residues at more than 50% of positions in subsequent sequence selection). To minimize this blurring while maintaining a useful level of insensitivity to side chain type, we considered explicit side chains in SCUBA with only backbone-dependent rotamer and steric packing terms, leaving out side chain-involving polar interactions and solvation. Modelled in this way, generically chosen side chains (for example, side chains of the 'LVG' form described below) serve as placeholders to fill the backbone inter-space, such that the final optimized backbones can have appropriate inter-space to accommodate tightly packed chiral side chains.

The relative weights of the various energy terms in SCUBA were calibrated using SCUBA-driven SD simulations of 33 natural proteins. When using the finalized energy weights in the simulations, median main chain root-mean-square deviation (r.m.s.d.) values from natural structures were 1.60 Å when native side chains were used and 2.78 Å when LVG-simplified side chains (see below) were used (Extended Data Fig. 2b; when overall structural compactness was restrained in the simulations, the median r.m.s.d. values were 1.25 Å when using native sequences and 2.23 Å when using LVG sequences).

Template-free protein design using SCUBA

We devised a de novo protein design workflow in which we use SCUBA for backbone generation and ABACUS2 (ref. [28]) for sequence selection. Starting from an initial backbone (obtained, for example, by using the procedure depicted in Extended Data Fig. 3), SCUBA-driven SD simulations with simulated annealing (SASDs) are used to obtain optimized backbones that are presumably designable. SASDs can be carried out in two substages: the first without considering side chains (to obtain the backbones of well-formed secondary structures packed in approximately proper ways) and the second considering generically chosen side chains (for example, using leucine homogeneously for α-helices, valine for β-strands and no side chain for loops, resulting in LVG sequences). LVG-simplified side chains are relatively featureless, of medium size and have preferred backbone conformations consistent with corresponding secondary structure types, yielding suitable side chain placeholders in a backbone-centred model. To fine tune a backbone optimized with such side chains to accommodate specific side chains when selecting final sequences, we use an iterative sequence selection–backbone relaxation approach[6], applying ABACUS2 and SCUBA-driven SASD in successive iterations.

We applied the SCUBA-driven SASD protocol to optimize backbones that fulfilled the specifications of sketches of various topological architectures. The sketches (that is, targeted architectures) were defined on the basis of 'periodic table' abstraction of globular protein structures[32], with the initial backbones geometrically constructed accordingly (Supplementary Methods and Extended Data Fig. 3). SCUBA-driven SASD consistently produced backbones that fulfilled the specifications of the sketches while reproducing natural backbones with notable precision (Extended Data Fig. 4).

Using initial structures grossly similar to the desired structure (for example, initial structures built as depicted in Extended Data Fig. 3) as a starting point for SASDs is efficient when designing proteins to fulfil topological architectures because it avoids unnecessarily extensive searching in structure space. Such initial structures can otherwise be arbitrary and even rather 'ill-formed'. In particular, 'strands' do not need to be placed or oriented to form inter-strand hydrogen bonds, because in subsequent SCUBA-driven SASDs hydrogen bonds can be automatically formed, broken and reformed to ultimately produce hydrogen-bond networks across entire β-sheets (Extended Data Fig. 5a). Given that the SASD optimization protocol imposes few restraints on initial backbones (apart from considerations for design specifications and/or efficiency), initial structures can be constructed in diverse ways beyond the example method presented in Extended Data Fig. 3.

In SCUBA-driven SASD, backbones evolve extensively with respect to their initial structures, as can be seen by inspecting the structures in Extended Data Figs. 4 and 5a; this is also evident from the large r.m.s.d. values (exceeding 6 Å) in Extended Data Fig. 5b. In this process, the physical plausibility of the backbones is substantially improved. To demonstrate this in silico, we selected ABACUS2 sequences for both the initial and optimized backbones for the sketches in Extended Data Fig. 4 and calculated the per-residue ABACUS2 energies and Rosetta energies (Extended Data Fig. 5c). The energies determined for the optimized backbones were invariably 1 to 2 energy units lower than those determined for the initial backbones, results consistent with improved designability.

After the first substage of backbone optimization (which does not consider side chains), further SASD optimizations—using LVG side

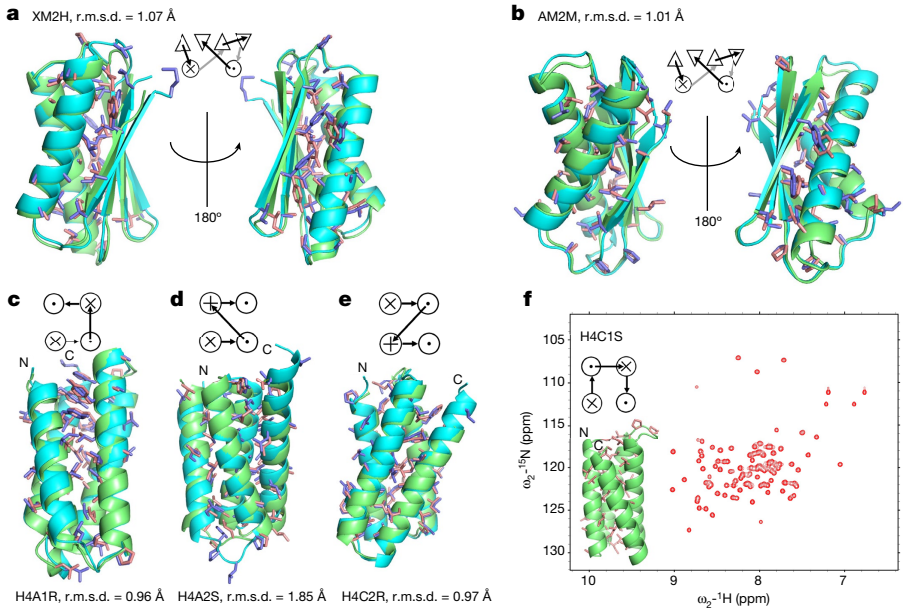


Fig. 3 | Successfully designed two-layered α/β proteins and four-helix bundle proteins. a–e, Experimentally solved structures (cyan) superimposed on designed structures (green) for H2E4 and H4 proteins. The r.m.s.d. values displayed are for main chain atoms. Internal side chains are shown as sticks (blue, experimentally determined; orange, designed). **f**, The designed structure and NMR HSQC spectrum of an H4 protein. Each panel includes the protein's name and a topology diagram. In the topology diagrams, the symbols ⊙ (△) and ⊕ (▽) represent, respectively, outward- and inward-pointing helices (strands); arrows indicate directed connections by loops.

chains before sequence selection and sequence-specific side chains after sequence selection—result in subtle changes in the backbone structures (r.m.s.d. of about 3 Å from adding LVG side chains and 1.1 Å or less in further side chain updates; see an example sketch in Extended Data Fig. 5b). This supports the idea that side chain types do not substantially affect the optimized backbones. However, the backbone adjustments from substage 1 to later optimization stages were necessary for meaningful sequence selection (Extended Data Fig. 5d).

Although we solved a crystal structure for one de novo protein designed with the initial protocol, our experimental results revealed a problem: this protocol does not consistently generate accurate, designable loop structures. In brief, when an entire backbone is optimized in one pass, it probably contains loops that are trapped in high-energy local minima, which understandably could cause failed designs, given that minimally strained loops are understood to be a defining characteristic of designable backbones[8,10,33]. To fix this, we refined the backbone optimization protocol by adding a loop resampling and optimization step (which also uses SCUBA-driven SD; Supplementary Methods). When applied to a two-helix/four-strand (H2E4) sketch (Extended Data Fig. 5a), the refined protocol substantially improved the energies of loop regions (Extended Data Fig. 6a) and improved the success rate of computationally folding the final sequences when using Rosetta biased forward folding[10,15] (Extended Data Fig. 6b, c).

Verifying designs of given architecture

We experimentally tested proteins designed with given topological architectures in two batches (see the detailed process in the Supplementary Methods, the summary in Extended Data Table 1a and the exact sequences in Supplementary Tables 1 and 2). The architectures included an H2E4 sketch (Extended Data Fig. 5a), several four-helix bundle (H4) sketches and an EXTD sketch, in which a loop in a natural protein domain was replaced by a designed segment of four helices (Fig. 2).

The proteins in batch 1 were designed using the backbone optimization protocol without loop resampling (described above). None of the 44 tested H2E4 or H4 designs was successfully crystallized. A solved crystal structure was obtained for one of the eight EXTD designs (EXTD-3) (Extended Data Fig. 7a presents a data summary for the X-ray structures reported in this study). Figure 2a shows the solved structure superimposed with the design model, in which a helix–loop–helix segment in the custom-built initial structure (Fig. 2b) changed to a single continuous helix during SCUBA-driven SASD, yielding a non-natural overall architecture (Fig. 2c). Among the three design-introduced loops in the solved structure, one precisely matches the design model, whereas the other two show large deviations (Fig. 2a).

These discrepancies between experimental and designed structures revealed by the proteins in batch 1 led us to add a loop resampling and optimization step to the backbone optimization protocol. We applied this revised protocol to the H2E4 and H4 sketches and filtered the designed sequences by Rosetta biased forward folding[9,14] (see Extended Data Table 1b for the number of structures or sequences retained at intermediate design stages). The 38 sequences retained after filtering made up the proteins in batch 2. Among them, 12 H2E4 and 4 H4 proteins were purified as monomers and were well folded as indicated by either X-ray crystal structures (for 2 H2E4 proteins and 3 H4 proteins; Fig. 3a–e) or nuclear magnetic resonance (NMR) heteronuclear single quantum coherence (HSQC) spectra (for 10 H2E4 proteins (Extended Data Fig. 7b) and 1 H4 protein (Fig. 3f)). Although the asymmetric crystal units of the two H2E4 proteins contained dimers, size-exclusion chromatography confirmed that these proteins were monomers in solution (Extended Data Fig. 7c).

The SCUBA-designed backbones agreed with the experimentally obtained structures with atomic accuracy, as indicated by the main chain r.m.s.d. values of 0.96–1.85 Å observed between the X-ray and designed structures (Fig. 3a–e). Moreover, 16 of 19 loops in the designed structures agreed with the experimentally refined loops with sub-angstrom precision (Extended Data Fig. 8a–e). Notably, the two H2E4 proteins did not have the same loop conformations, despite their similarly packed secondary structures (Extended Data Fig. 8f).

The designed H2E4 and H4 proteins have low sequence identity with known natural proteins with similar structures, exhibiting identity with structurally aligned natural protein sequences ranging from 3.5–25%

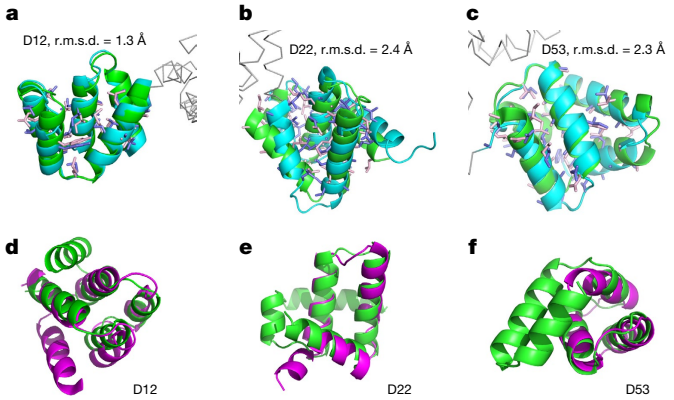


Fig. 4 | Structures of successfully designed de novo proteins that fold into novel architectures. a–**c**, Designed models (green) of the proteins D12, D22 and D53, superimposed on their respective crystal structures (cyan). The r.m.s.d. values displayed are for main chain atoms. Internal side chains are shown as sticks (blue, experimentally determined; orange, designed). The structure of MBP (fused to the N termini of analyte proteins to facilitate protein crystallization) is partially shown (grey). **d**–**f**, Comparison of designed models (green) with their most similar natural partial structure (magenta) found by Dali search of PDB.

(average identity of 14%), much lower than the average sequence identity (~33%) between native sequences and ABACUS2 sequences selected for natural backbones[28]. The designed proteins are of similarly high thermostability as proteins with ABACUS sequences selected for natural backbones[34], as indicated by temperature-dependent circular dichroism spectroscopy data for two example proteins (Extended Data Fig. 7d).

The design success rate for the proteins in batch 2 was ~42% (16 well-folded proteins of 38 experimentally tested ones; we acknowledge that this success rate benefits from contributions from post-design computational filtering). These results support the idea that, alongside the state-of-the-art RosettaDesign method[6,14], SCUBA+ABACUS2 is a useful method for de novo protein design.

Verifying designs of novel architecture

The third batch of proteins we experimentally examined had novel, computationally found architectures (Extended Data Fig. 9). These all-helical structures were generated by SCUBA-driven SASD optimizations of initially randomly placed helical backbone segments, followed by truncation and ordering of the segments and subsequent connection of the segments with SCUBA-optimized loops (Supplementary Methods). Among the 13 proteins in batch 3, 3 were purified in monomer form and were found to be well folded (see the HSQC spectra in Extended Data Fig. 7b). Fusing these three proteins (separately) to mannose-binding protein (MBP) enabled protein crystallization: X-ray structures were solved and are shown in superposition with designed structures in Fig. 4a–c.

The structural novelty of these three proteins is illustrated in Fig. 4d–f, which presents superpositioning of the designed proteins with their most similar natural (partial) structures found in PDB. The success of SCUBA in designing these proteins highlights (1) that currently known protein structures in PDB are limited (at least when considered relative to the immensity of possible protein structures) and (2) that SCUBA-driven designs are not limited by the natural protein architectures used to train the model.

Discussion

Multiple attributes distinguish our approach from existing protein design methods. When using SCUBA, backbone structures are sampled and optimized continuously and with complete flexibility. Moreover, our backbone-centred SCUBA model eliminates the need to search the sequence space at the backbone design stage. Together, these two attributes support facile exploration of entirely novel backbone architectures that have not been observed in nature.

The proteins of novel architecture generated in the present study clearly demonstrate the utility of our approach for designing broader protein geometries than are observed in nature. When designing functional proteins, energy function-driven backbone sampling and optimization can be easily tailored to facilitate both extensive exploration of the structure space (for example, by applying enhanced sampling techniques) and precise control over designed structures (for example, by applying function-related restraints). These approaches enable substantial expansion of the structural diversity and functionalities accessible to de novo protein design.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-04383-5.

1. Li, H., Helling, R., Tang, C. & Wingreen, N. Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666–669 (1996).
2. England, J. L. & Shakhnovich, E. I. Structural determinant of protein designability. *Phys. Rev. Lett.* **90**, 218101 (2003).
3. Hoang, T. X., Trovato, A., Seno, F., Banavar, J. R. & Maritan, A. Geometry and symmetry presculpt the free-energy landscape of proteins. *Proc. Natl Acad. Sci. USA* **101**, 7960–7964 (2004).
4. Rose, G. D., Fleming, P. J., Banavar, J. R. & Maritan, A. A backbone-based theory of protein folding. *Proc. Natl Acad. Sci. USA* **103**, 16623–16633 (2006).
5. Skolnick, J. & Gao, M. The role of local versus nonlocal physicochemical restraints in determining protein native structure. *Curr. Opin. Struct. Biol.* **68**, 1–8 (2021).
6. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
7. Jiang, L. et al. De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).
8. Koga, N. et al. Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
9. Marcos, E. et al. Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
10. Dou, J. et al. De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).
11. Lu, P. et al. Accurate computational design of multipass transmembrane proteins. *Science* **359**, 1042–1046 (2018).
12. Glasgow, A. A. et al. Computational design of a modular protein sense–response system. *Science* **366**, 1024–1028 (2019).
13. Huang, P. S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
14. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
15. Grigoryan, G. & DeGrado, W. F. Probing designability via a generalized model of helical bundle geometry. *J. Mol. Biol.* **405**, 1079–1100 (2011).
16. Thomson, A. R. et al. Computational design of water-soluble α-helical barrels. *Science* **346**, 485–488 (2014).
17. Brunette, T. J. et al. Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015).
18. Jacobs, T. et al. Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352**, 687–690 (2016).
19. Pan, X. et al. Expanding the space of protein geometries by computational design of de novo fold families. *Science* **369**, 1132–1136 (2020).
20. Baker, D. An exciting but challenging road ahead for computational enzyme design. *Protein Sci.* **19**, 1817–1819 (2010).
21. Otten, R. et al. How directed evolution reshapes the energy landscape in an enzyme to boost catalysis. *Science* **370**, 1442–1446 (2020).
22. Zhang, Y., Hubner, I. A., Arakaki, A. K., Shakhnovich, E. & Skolnick, J. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl Acad. Sci. USA* **103**, 2605–2610 (2006).
23. Kukic, P. et al. Mapping the protein fold universe using the CamTube force field in molecular dynamics simulations. *PLoS Comput. Biol.* **11**, e1004435 (2015).
24. MacDonald, J. T., Maksimiak, K., Sadowski, M. I. & Taylor, W. R. De novo backbone scaffolds for protein design. *Proteins Struct. Funct. Bioinf.* **78**, 1311–1325 (2010).
25. MacDonald, J. T. et al. Synthetic β-solenoid proteins with the fragment-free computational design of a β-hairpin extension. *Proc. Natl Acad. Sci. USA* **113**, 10346–10351 (2016).
26. Van Gunsteren, W. F., Berendsen, H. J. C. & Rullmann, J. A. C. Stochastic dynamics for molecules with constraints: Brownian dynamics of n-alkanes. *Mol. Phys.* **44**, 69–95 (1981).

27. Xiong, P. et al. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. *Nat. Commun.* **5**, 5330 (2014).
28. Xiong, P. et al. Increasing the efficiency and accuracy of the ABACUS protein sequence design method. *Bioinformatics* **36**, 136–144 (2020).
29. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
30. Wang, G. & Dunbrack, R. L., Jr. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **33**, W94–W98 (2005).
31. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
32. Taylor, W. R. A 'pperiodic table' for protein structures. *Nature* **416**, 657–662 (2002).
33. Baker, D. What has de novo protein design taught us about protein folding and biophysics? *Protein Sci.* **28**, 678–683 (2019).
34. Liu, R., Wang, J., Xiong, P., Chen, Q. & Liu, H. De novo sequence redesign of a functional Ras-binding domain globally inverted the surface charge distribution and led to extreme thermostability. *Biotechnol. Bioeng.* **118**, 2031–2042 (2021).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Coordinates and structure files for designed proteins have been deposited to PDB under the following accession codes: 7DMF (EXTD-3), 7DKK (XM2H), 7DKO (AM2M), 7DGU (H4A1R), 7DGW (H4A2S), 7DGY (H4C2R), 7FBB (D12), 7FBC (D22) and 7FBD (D53). Other relevant data are available in the main text or the Supplementary Information.

Code availability

Executable computer programs, source code and model parameters for SCUBA and ABACUS2 are available for public download and free non-commercial use from https://doi.org/10.5281/zenodo.4533424.

Acknowledgements This work was supported by the National Key R&D Program of China (2018YFA0900703 to H.L. and 2018YFA0901600 to Q.C.), the National Natural Science Foundation of China (21773220 to H.L., 31971175 to Q.C. and 32090040 to J. Z.) and the Youth Innovation Promotion Association, Chinese Academy of Sciences (2017494 to Q.C.). We thank the staff from the Core Facility Centre for Life Sciences, USTC, and from the BL18U1, BL19U1 and BL02U1 beamlines of the National Facility for Protein Science in Shanghai (NFPS) and the Shanghai Synchrotron Radiation Facility for assistance during crystallographic data collection. We thank the USTC Supercomputing Center for computing resource. We thank Z. Zhu, F. Li, Y. Wang, M. Lv and Y. Yun for assistance with X-ray diffraction data collection and processing and T. Jin for sharing MBP expression plasmids.

Author contributions H.L., B.H. and Y.X. developed computational models and code, and B.H., Y.X., X.H. and Q.C. performed protein design and experimental characterization. S.L. and Y.L. collected and analysed crystallographic data. J.H., J.Z. and C.H. collected and helped process NMR data. H.L. and Q.C. supervised the project. H.L., Q.C. and B.H. wrote the manuscript, and other authors were involved in discussion.

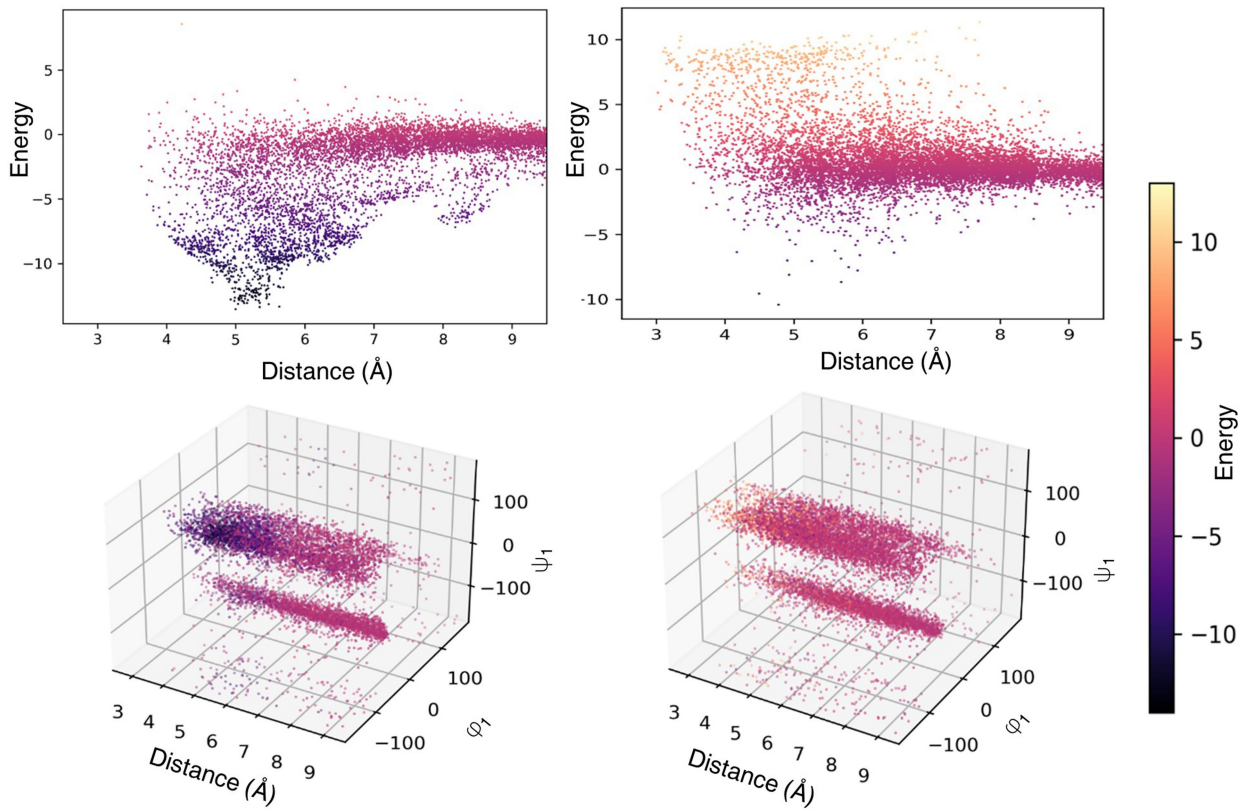
Competing interests H.L., Q.C., B.H., Y.X. and X.H. have filed a patent application (202111197820.0) relating to the template-free protein design method in the name of the University of Science and Technology of China. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-04383-5.

Correspondence and requests for materials should be addressed to Quan Chen or Haiyan Liu.

Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

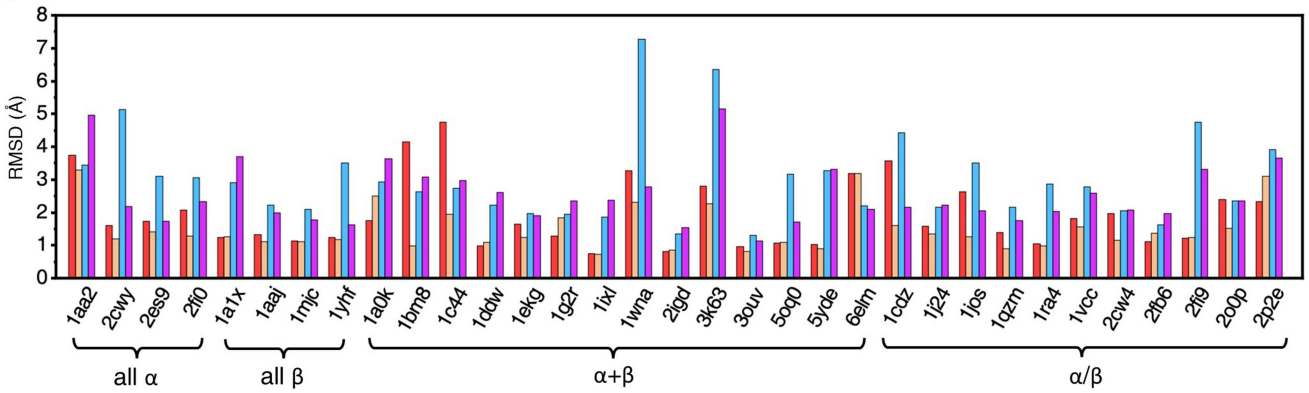


Extended Data Fig. 1 | Statistical energies learned by NC-NN capture correlations in high-dimensional space. a–**d**, Scatter graphs showing projections of a NC-NN-learned term for the through-space interactions between two backbone positions. In total, the term depends on 14 variables. The variables used for projections include the Cα-Cα distance (**a**, **b**) between the two positions, and additionally the mainchain torsional angles φ_1 and ψ_1 at one position (**c**, **d**). Points are colored according to statistical energy values (in arbitrary units) as indicated by the color bar. Points in **a** and **c** correspond to observed configurations, while those in **b** and **d** correspond to configurations randomly drawn according to the reference distribution.

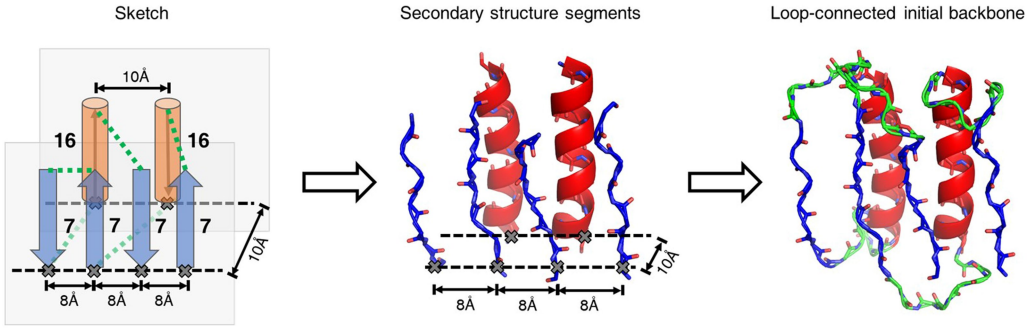
a

Type	Structural variables	Reference distribution	Sidechain involved?
$e_{NC-NN}^{\varphi-\psi\text{-}1}$	2 Ramachandran torsional angles of one residue, encoded by 12 input nodes.	uniform distribution in (φ, ψ) space	No
$e_{NC-NN}^{\varphi-\psi\text{-}5}$	8 Ramachandran torsional angles of a window of 5 consecutive residues, excluding the φ angle of the first and the ψ angle of the last residue, encoded by 64 input nodes.	multiplications of the observed 1-D or 2-D distributions of the Ramachandran angles of individual residues	No
$e_{NC-NN}^{site-pair}$	translations and rotations between the main chain atoms of two residues that are sequentially far-separated (sequence separation > 5) and spatially close (inter-Cα distance below 9.5 Å), and the two sets of four Ramachandran torsional angles, each set centered around one residue, encoded by 191 input nodes.	uniform distributions of inter-residue relative position and orientation, multiplied by the observed distributions of the four surrounding Ramachandran angles surrounding each residue. Sterically clashed configurations excluded	No
$e_{NC-NN}^{local-HB}$	internal geometries of a group of 6 atoms from two peptide units: $C_i, O_i, N_{i+1}, C_j, O_j$, and N_{j+1} in which *i* and *j* are sequential residue numbers and $j - i \leq 5$, encoded by 35 input nodes.	uniform distributions of the relative position and orientation between the two peptide units, excluding sterically clashed configurations	No
$e_{NC-NN}^{rotamer}$	sidechain and backbone rotatable torsional angles of one residue, each angle encoded by 8 input nodes.	uniform distributions of the sidechain torsional angles multiplied by the observed 2-D distribution of the Ramachandran angles	Yes

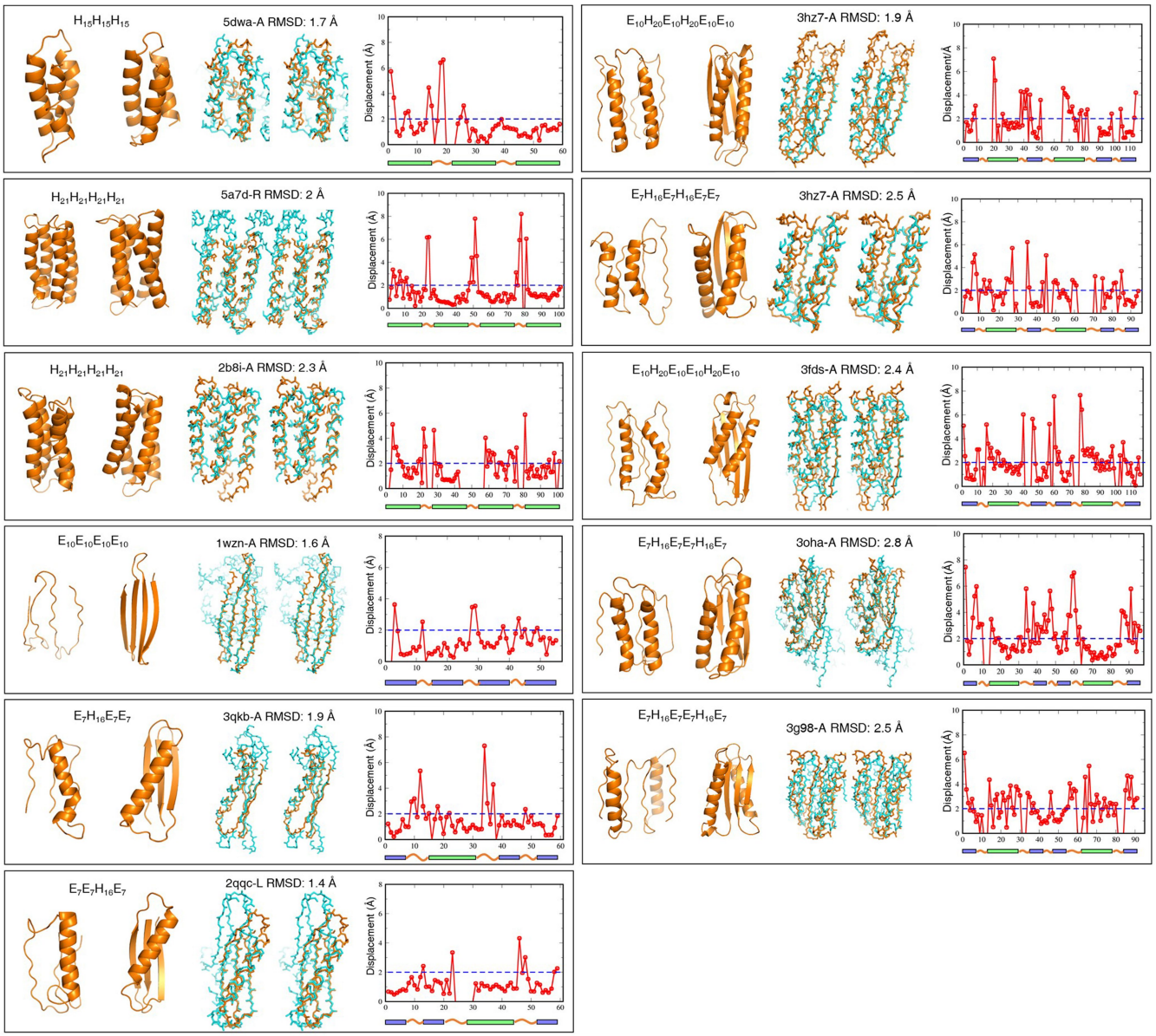
b



Extended Data Fig. 2 | NC-NN-learned components in SCUBA and simulations of natural protein structures by SCUBA. a, Types of NC-NN-learned statistical energy terms in SCUBA. **b**, The deviations of conformations sampled in SCUBA-driven SD simulations from native conformations for 33 natural proteins. Each protein was simulated for 900 *ps* at reduced temperature $T_r = 1.0$ and the r.m.s.d. values (noted as RMSD in the figure) are for mainchain atoms in secondary structures averaged over the last 50 *ps*. Simulations were carried out either with or without a radius of gyration (R_g) restraint, which, as described in the Supplementary Methods, was optionally applied in later backbone design simulations both to bias the sampling of more compact structures and to compensate for thermal expansion in simulated annealing simulations involving higher temperatures. The restraint energy took the form $E_{Rg-restraint}(R_g) = -k_{res} ln\left(\frac{R_g}{R_g^0}\right)$ when $R_g > R_g^0$ and $E_{Rg-restraint}(R_g) = 0$ for $R_g \leq R_g^0$ ($k_{res} = 300$ in reduced energy unit and $R_g^0 = 5$Å). This energy term leads to only weak compressing forces in comparisons with the strong inter-atomic steric repulsions, and does not distort the tightly-packed native-like minimum structures. The median r.m.s.d. values across the 33 proteins are 1.60 Å (native sequences, without R_g restraint, red bars), 1.25 Å (native sequences, with R_g restraint, orange bars), 2.78 Å (LVG sequences, without R_g restraint, blue bars), and 2.23 Å (LVG sequences, with R_g restraint, violet bars).

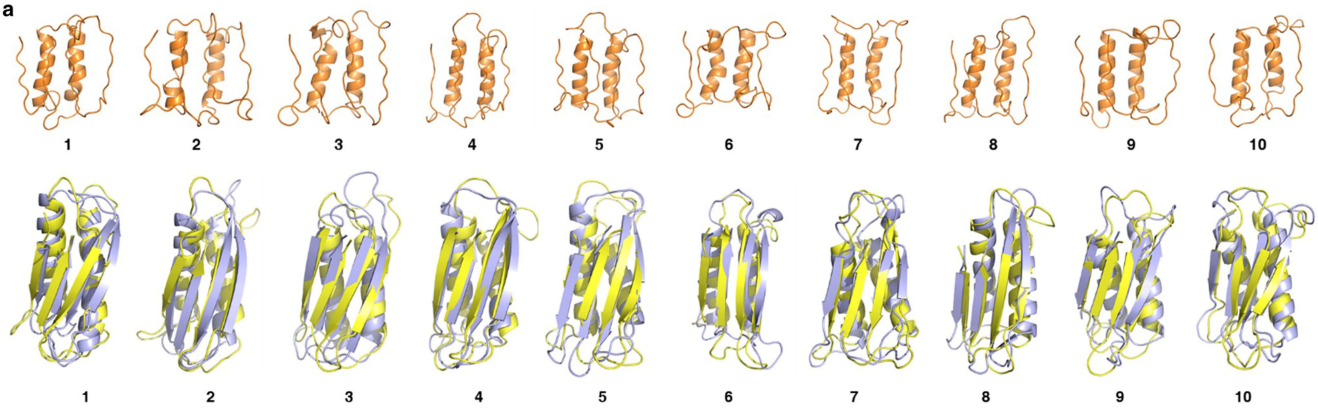


Extended Data Fig. 3 | Generating initial backbone for a given sketch or topological architecture. A sketch is represented as an abstracted architecture comprising regularly arranged layers of secondary structures, the layers in parallel planes. From the abstraction, coordinates of starting or ending positions (indicated by "×") of secondary structure segments are determined as regular grid points on parallel straight lines in different planes. The N to C directions of the segments are perpendicular to the lines. The approximate lengths of the segments may also be pre-specified. Then peptide segments of corresponding local conformations are geometrically generated using coordinates of their terminal positions and directions determined from the sketch. Connecting the segments with closed loops leads to the initial backbone structure to be used by SCUBA-driven SASD.



Extended Data Fig. 4 | SCUBA-driven SASD produced backbones similar to natural proteins. Different boxes correspond to different design sketches. From left to right in each box: initial backbone, optimized backbone, a stereo view of the optimized backbone superposed with the closest natural structure, and deviations of Cα atom positions between the designed and the closest natural backbones. In each box, the text string indicates the type, approximate size, and order of secondary structure segments of the corresponding sketch ("H" for helix, "E" for strand, and the subscripts indicate lengths). The closest natural structures with given PDB IDs and chain IDs were identified using Dali searches. The r.m.s.d. values (noted as RMSD in the figures) are of Cα atoms in aligned secondary structure elements.

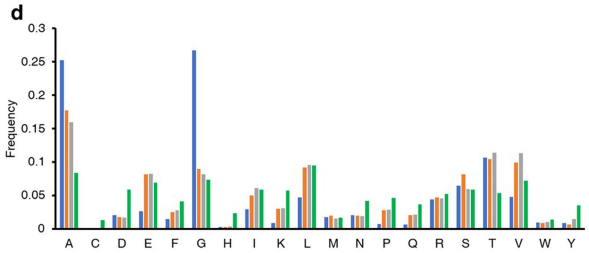
a



b

	Init	Substage-1	Substage-2	Iter1	Iter2
Init	0				
Substage-1	6.39(0.47)	0			
Substage-2	6.11(0.47)	2.99(0.38)	0		
Iter1	6.07(0.45)	2.99(0.36)	0.69(0.14)	0	
Iter2	6.07(0.50)	2.99(0.39)	0.94(0.17)	0.70(0.15)	0
Iter3	6.06(0.54)	2.95(0.40)	1.10(0.20)	0.91(0.17)	0.69(0.13)

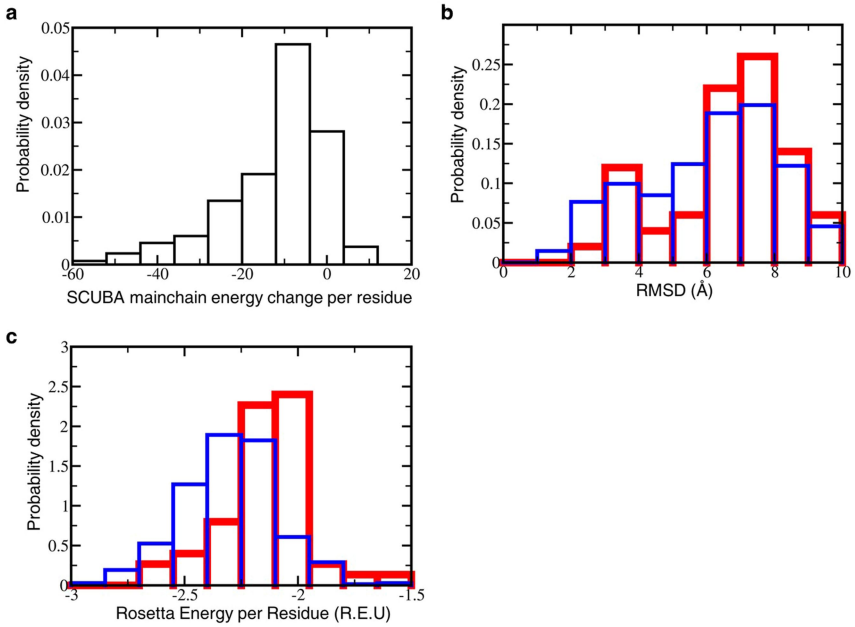
d



c

Architecture		1	2	3	4	5	6	7	8	9	10
ABACUS2 energy per residue	Initial	1.49	1.05	1.02	1.01	0.89	0.82	1.01	0.82	0.96	1.33
	Optimized	0.27	-0.31	-0.13	-0.62	-0.70	-0.68	-0.71	-0.58	-0.63	-0.87
Rosetta energy per residue	Initial	-0.98	-1.42	-1.43	-1.51	-1.54	-1.47	-1.46	-1.51	-1.74	-1.90
	Optimized	-2.65	-2.76	-2.75	-3.05	-3.17	-3.05	-3.22	-3.05	-3.02	-3.17

Extended Data Fig. 5 | Examples of backbone changes at different design stages. a, Initial and optimized backbones for the H2E4 sketch, whose secondary structure sequence is $E_7H_{16}E_7H_{16}E_7E_7$ ("H" for helix, "E" for strand, and the subscripts indicate approximate lengths). The top row shows artificially constructed initial structures, while the bottom row shows substage-1 backbones optimized without sidechain (yellow) superimposed with substage-2 backbones optimized with LVG-simplified sidechains (violet). **b**, The r.m.s.d. values of mainchain atoms (in Å) between the successively generated structures at different design stages of backbone optimization or relaxation. The results have been averaged over the H2E4 designs (standard deviations are given in parentheses). The meanings of the notations are: "Init" for the initial structure, "Substage-1" for substage-1 backbones optimized without sidechains, "Substage-2" for substage-2 backbones optimized with LVG-simplified sidechains, and "Iter1" to "Iter3" for backbones relaxed with the designed sidechains in the sequence design-backbone relaxation iterations. **c**, ABACUS2 and Rosetta energies of ABACUS2-selected sequences for initial and SCUBA-optimized backbones of different topological architectures. The secondary structure compositions of the architectures are: 1: $E_{10}E_{10}E_{10}E_{10}$, 2: $E_7H_{16}E_7E_7$, 3: $E_7E_7H_{16}E_7$, 4: $E_7H_{16}E_7E_7H_{16}E_7$, 5: $E_{10}H_{20}E_{10}H_{20}E_{10}E_{10}$, 6: $E_7H_{16}E_7H_{16}E_7E_7$, 7: $E_{10}H_{20}E_{10}E_{10}H_{20}E_{10}$, 8: $E_7H_{16}E_7E_7H_{16}E_7$, 9: $H_{15}H_{15}H_{15}$, 10: $H_{21}H_{21}H_{21}H_{21}$. For each sketch, 10 initial backbones have been optimized to generate 10 optimized backbones. Sketch 10 led to optimized backbones of both left-handed and right-handed twists, as shown in two boxes in Extended Data Fig. 4. Each energy value has been averaged over 100 sequences selected on a group of 10 initial or optimized backbones (10 sequences selected using ABACUS2 for each backbone), with standard deviations between 0.08 and 0.38. Rosetta energies have been calculated on relaxed structures with selected sequences. **d**, Amino acid usage frequencies in sequences selected with ABACUS2 on the H2E4 backbones at different optimization stages. Averaged values are shown separately for sequences designed using the substage-1 backbones optimized without any explicit sidechain (blue bars), using second stage backbones optimized with LVG-simplified sidechains (orange bars), and using backbones relaxed with the first round ABACUS2-selected sidechains (gray bars) (the sidechain atom radius parameters had been downscaled by multiplying 0.9 to introduce larger sidechains in the first round of sequence selection). The green bars correspond to the distribution in the training proteins.

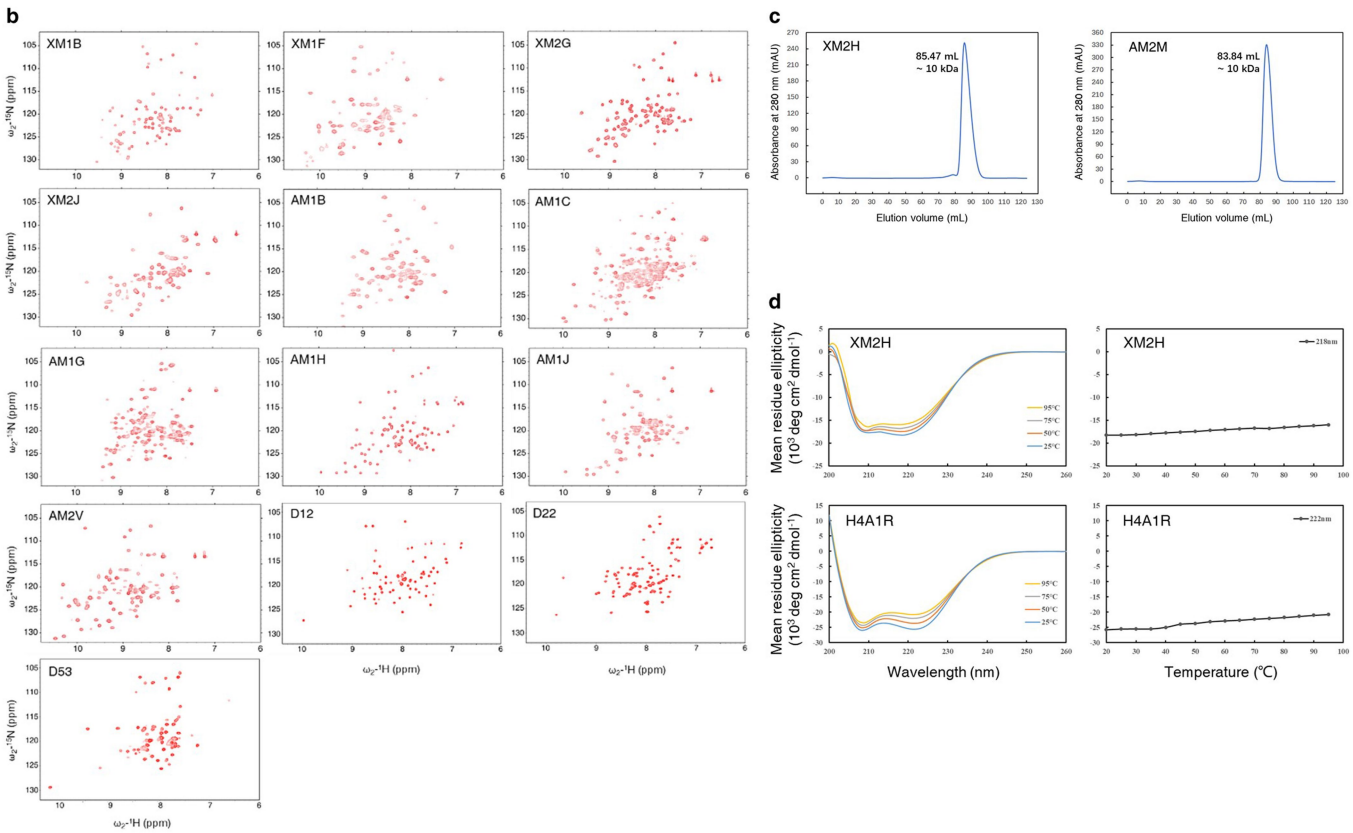


Extended Data Fig. 6 | Effects of loop resampling and optimization. **a**, The distribution of the per-residue SCUBA energy changes of loop residues caused by loop resampling and optimization. For the H2E4 backbone structures, the changes were calculated as the energies after loop re-optimization minus the energies before loop re-optimization. **b**, The distribution of the lowest r.m.s.d. values (noted as RMSD in the figure) of predicted structures from designed structures. **c**, The distribution of per-residue Rosetta energy of the lowest-r.m.s.d. predicted structures. For **b** and **c**, the predictions were carried out using Rosetta biased forward folding for sequences designed from the loop re-optimized H2E4 backbone structures (thinner blue lines), or for sequences designed from H2E4 backbone structures not yet subjected to loop re-optimization (thicker red lines).

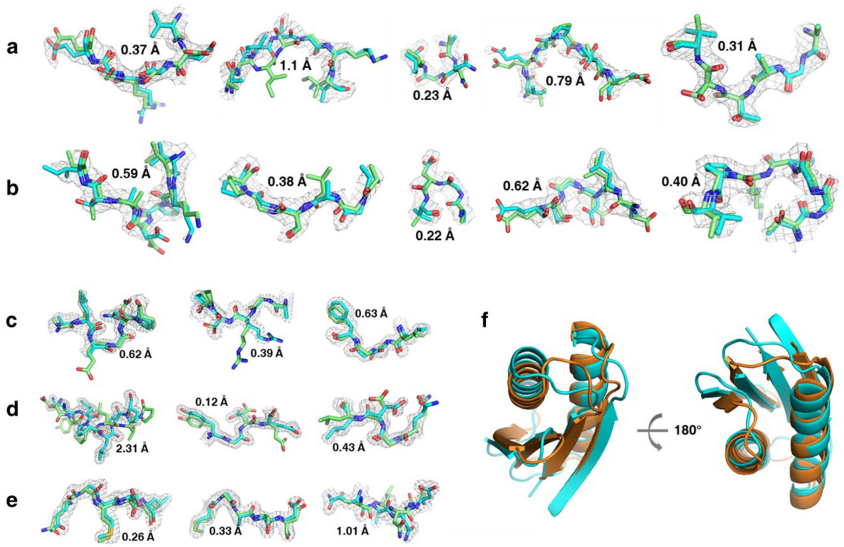
a

	EXTD-3	XM2H	AM2M	H4A1R	H4A2S	H4C2R	D12	D22	D53
Data Collection									
Wavelength(Å)	0.9789	1.5406	1.5406	1.5406	0.9785	0.9785	0.9791	0.9791	1.5406
Space group	$P6_5$	$P2_1$	$I2_1$	$P3_221$	$P2_12_12$	$P4_12_12$	$P2_1$	$P2_12_12_1$	H32
Cell parameters									
a, b , c (Å)	69.73, 69.73, 79.29	48.19, 63.71, 70.74	61.26, 58.62, 108.35	33.41, 33.41, 159.79	46.53, 53.80, 30.55	70.68, 70.68, 64.54	56.08, 112.81, 84.66	61.63, 77.14, 94.88	158.47,158.47, 266.19
α, β , γ (°)	90, 90, 120	90, 101.62, 90	90, 91.34, 90	90, 90, 120	90, 90, 90	90, 90, 90	90, 96.86, 90	90, 90, 90	90, 90, 120
Resolution* (Å)	50-2.20 (2.24-2.20)	13.03-2.10 (2.18-2.10)	12.96-2.60 (2.69-2.60)	12.81-1.75 (1.81-1.75)	35.19-1.35 (1.37-1.35)	49.98-1.80 (1.84-1.80)	67.4-2.31 (2.43-2.31)	48.15-1.85 (1.89-1.85)	13.08-2.85 (3.02-1.85)
R_{merge} *(%)	13.6(56.6)	10.5 (74.5)	12.3 (65.9)	7.2(100.8)	4.1(24.8)	4.3(69.2)	18.1(62.0)	10.6(92.4)	19.8(94.2)
I/σI*	4.0(27.0)	13.44 (0.91)	12.42 (1.15)	18.4(1.5)	29.0(5.6)	45.0(5.2)	6.3(2.8)	11.6(3.1)	10.6(1.9)
Completeness* (%)	99.9(98.5)	94.8 (88.8)	98.2 (97.5)	99.7(100.0)	99.5(92.9)	100.0(100.0)	100.0(100.0)	100.0(100.0)	99.0(100.0)
Redundancy*	19.6(17.2)	4.7 (3.6)	4.5 (3.3)	9.8(6.5)	12.0(7.9)	25.3(23.4)	6.9(6.7)	7.7(8.0)	9.5(6.7)
Refinement									
No. reflections used/free	8396/823	23355/1193	11773/549	11152/1118	17343/1734	15624/1563	45925/2302	39308/1939	29930/1482
Resolution (Å)	50-2.20	13.03-2.10	12.96-2.60	12.81-1.75	35.20-1.35	47.66-1.80	23.91-2.31	48.16-1.85	13.02-2.85
R_{work}/R_{free}(%)	23.19/26.86	21.90/26.30	24.03/28.93	19.84/23.11	19.09/22.67	17.62/19.25	22.53/26.35	18.43/21.50	22.84/27.73
R.M.S.D									
Bondslengths (Å)	0.002	0.008	0.005	0.004	0.004	0.005	0.006	0.009	0.010
Bond angles (°)	0.407	1.149	0.781	0.600	0.720	0.590	0.741	0.892	0.970
B-factors (Å^2)									
Protein	44.80	42.77	50.60	27.40	21.80	34.20	54.66	28.66	33.10
Water	17.72	28.92	26.00	40.70	33.80	48.60	40.29	24.42	NA
No. atoms									
Protein	1478	2628	1899	765	724	756	6680	3612	6747
Water	5	37	13	66	84	85	31	116	NA
Ramachandran plot									
Favored/allowed/outlier(%)	96.79/3.21/0	96.80/3.20/0	98.10/1.90/0	100.0/0.00/0	98.90/1.10/0	100.0/0.00/0	97.59/2.41/0	98.47/1.53/0	97.53/2.47/0

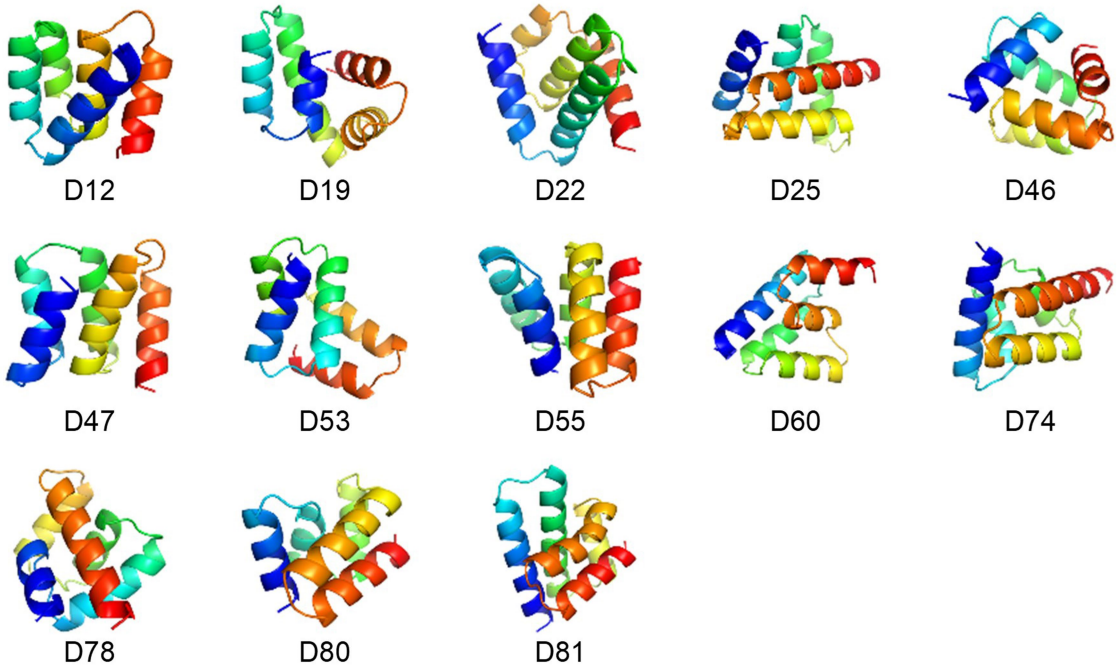
* Values in parentheses are for highest-resolution shell.



Extended Data Fig. 7 | Experimental characterizations of designed proteins. a, X-ray data collection and refinement of crystal structure models. **b**, NMR ^{15}N-^{1}H HSQC spectra of ten designed H2E4 proteins and three novel helical proteins. **c**, Size exclusion chromatography results of the designed H2E4 proteins XM2H (left) and AM2M (right) in solution. The chromatograms were obtained for samples purified by gel filtration, and the molecular weights were estimated from the peak positions. **d**, Circular dichroism spectroscopy of the designed proteins XM2H (top) and H4A1R (bottom) at different temperatures. The slow varying temperature-dependent curves shown on the right suggest that there are only small changes in the secondary structure contents of these proteins over the temperatures range from 25 to 95 °C. For XM2H, its helical content (calculated from the CD curves) decreased from 54.9% at 20 °C to 48.2% at 95 °C, while its β-sheet content changed from 9% to 11%. For H4A1R, its helical content changed from 85.2% at 20 °C to 71.8% at 95 °C.



Extended Data Fig. 8 | The structures of the loops in the H2E4 and H4 proteins. a–**e**, Superimpositions of experimentally determined structures (cyan) with corresponding designed structures (green) for loops in the designed proteins XM2H(**a**), AM2M(**b**), H4A1R(**c**), H4A2S(**d**), and H4C2R(**e**). The 2Fo-Fc (at 1.0 σ level) electron density surfaces are also shown. The r.m.s.d. for main chain atoms are displayed. **f**, The experimentally determined structures of the two H2E4 proteins are superimposed (XM2H in cyan and AM2M in orange) to show their different loop structures connecting similarly arranged secondary structure segments.



Extended Data Fig. 9 | Designed backbone structures of the experimentally examined all-helical proteins in Batch 3. We note that the average per-residue Rosetta energy of the proteins with experimentally solved structures (D12, D22 and D53) is −3.32 ± 0.07 (in arbitrary unit), while for the remaining ten Batch-3 proteins, the same average value is −3.22 ± 0.14.

Extended Data Table 1 | Summary of experimentally examined designs

a

Batch	Architecture	Number	Summary information
1	EXTD	8	6 proteins expressed: EXTD-1, 3, 4, 5, 6, 8 2 proteins soluble: EXTD-1, 3 (EXTD-1, 3 have merged H3 helix) 1 structure solved: EXTD-3
	H2E4	33	24 proteins could not be expressed or purified as monomers 9 proteins purified but not crystalized
	H4	11	7 proteins could not be expressed or purified as monomers 4 proteins purified but not crystalized
2	H2E4	30	20 proteins expressed: XM1A, XM1E, XM2G, XM2J, AM1B, AM1G, XM1B, XM1F, AM1A, AM2M, AM2P, XM2H, AM1J, AM2V, AM1H, XM1D, AM1C, AM2N, AM2R, AM2T 14 proteins soluble: XM2G, XM2J, AM1B, AM1G, XM1B, XM1F, AM2M, AM2P, XM2H, AM1J, AM2V, AM1H, AM1C, AM2T 12 proteins purified as monomers: XM2G, XM2J, AM1B, AM1G, XM1B, XM1F, AM2M, XM2H, AM1J, AM2V, AM1H, AM1C structures of AM2M and XM2H solved, the remaining 10 purified proteins characterized by NMR HSQC
	H4	8	4 proteins expressed, soluble and purified in monomer forms: H4A1R, H4C1S, H4A2S, H4C2R structures of H4A1R, H4A2S, and H4C2R solved. H4C1S characterized by NMR HSQC
3	Novel all-helical	13	4 proteins expressed: D12, D22, D47, D53 structures of three proteins, D12, D22, and D53 solved in MBP-fused form, with isolated proteins characterized by NMR HSQC

b

Sketch	EXTD	H2E4	H4
Number of initial backbones	1000	10	30
Number of SASD optimized backbones	1000	50	116
Number of backbones with re-optimized loops	\	500	3382
Number of ABACUS2 selected sequences	1000	2000	3382
Number of experimentally tested sequences	8	30	8

a, Summary information for the three batches of experimentally examined proteins. **b**, Numbers of generated or examined structures/sequences at different designing/testing stages for EXTD (Batch 1), H2E4 (Batch 2), and H4 (Batch 2).

nature research

Corresponding author(s): Quan Chen; Haiyan Liu

Last updated by author(s): 2021-12-10

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Executables and source codes of the home-made programs (SCUBA and ABACUS2) for protein design described in the manuscript are available from http://doi.org/10.5281/zenodo.4533424 for free non-commercial use.
Data analysis	NMR data were processed using the software NMRDraw/NMRPipe (Version 8.2) and SPARKY 3.115. Crystallographic data of XM2H, AM2M, H4A1R and D53 were processed using the CrysAlisPro software suite (Version 1.171.39.35c). Data sets of EXTD-3, H4A2S, H4C2R, D12 and D22 were processed using XDS56. Model rebuilding was performed in Coot (Wincoot 0.8.9.2) and the final structures were refined by PHENIX 1.14-3260. Structure figures were made with PyMOL version 1.5. The CD spectra data of XM2H and H4A1R were processed and the secondary structure contents were estimated with built-in software suite Pro-Data Viewer v4.5 and Deconvolution v2.1 respectively. Training of the neural network models have been performed with the TensorFlow machine learning package version 1.2.0. PDB structure searches were performed using the Dali server (http://ekhidna2.biocenter.helsinki.fi/dali/). Rosetta energy calculations and biased forward folding were performed using the Rosetta software suite version 3.12.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Coordinates and structure files of designed proteins have been deposited to the Protein Data Bank under the following accession codes: 7DMF (EXTD-3), 7DKK

(XM2H), 7DKO (AM2M), 7DGU (H4A1R), 7DGW (H4A2S), 7DGY (H4C2R), 7FBB(D12), 7FBC(D22), 7FBD(D53). The set of protein structures used to train SCUBA were culled using the PDB sequence culling server PISCES (http://dunbrack.fccc.edu/pisces), and the actual structures for training were downloaded from the Protein Data Bank.The learned parameters for use in SCUBA and ABACUS2 are included in the SCUBA and ABACUS2 packages available from http://doi.org/10.5281/zenodo.4533424. Amino acid sequences and gene sequences of all designed proteins examined by experiments are available as Supplementary Data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The first batch of experiments examined 33, 11, and 8 proteins designed for the H2E4, H4, and EXTD sketches, respectively. One of EXTD sketch structure is determined. The second and third batches of experiments examined 30, 8 and 13 proteins designed for the H2E4, H4 and novel helical protein sketches, respectively. The crystal structures of two proteins designed for the H2E4 sketch, three proteins designed for the H4 sketch and three proteins designed for the (MBP-fused) novel helical protein sketches have been solved. The backbone structure sketches were chosen to include both all-alpha and alpha/beta structure classes, and to serve as representative examples for designs for given architectures and designs by exploring the designable backbone space. The numbers of proteins for experimental examination were chosen based on the estimated workload of the experiments. For each of example cases of the H2E4 sketch, the H4 sketches and the novel helical backbones, more than one final high-resolution structures that agree closely with correponding design models were obtained, which indicates that these numbers of experimentally examined proteins were sufficient to establish the validity and usefulness of the SCUBA +ABACUS2 design protocol.
Data exclusions	No data were excluded.
Replication	Protein expression and solubility was tested once or twice. All attempts to replicate expression and solubility screening experiments for further experimental characterization were successful. The X-ray structures were determined based on single crystals using standard procedures which have internal statistical validations.
Randomization	There was no randomized sample allocation in this work. All tested protein designs received identical treatment.
Blinding	Blinding is not relevant to our study because there is no group allocation.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*

Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for*

transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.