nature biotechnology

Article

https://doi.org/10.1038/s41587-024-02395-w

Multistate and functional protein design using RoseTTAFold sequence space diffusion

Received: 5 February 2024

Accepted: 21 August 2024

Published online: 25 September 2024

Check for updates

Sidney Lyayuga Lisanza[1,2,3,10], Jacob Merle Gershon[1,2,4,10], Samuel W. K. Tipps[1,2,10], Jeremiah Nelson Sims[2,5,10], Lucas Arnoldt[1,2,6,10], Samuel J. Hendel[1,2], Miriam K. Simma[7], Ge Liu[1,2], Muna Yase[1,2,4], Hongwei Wu[7], Claire D. Tharp[7], Xinting Li[1,2], Alex Kang[2], Evans Brackenbrough[2], Asim K. Bera[2], Stacey Gerben[2], Bruce J. Wittmann[8], Andrew C. McShan[7] & David Baker[1,2,9]

Protein denoising diffusion probabilistic models are used for the de novo generation of protein backbones but are limited in their ability to guide generation of proteins with sequence-specific attributes and functional properties. To overcome this limitation, we developed ProteinGenerator (PG), a sequence space diffusion model based on RoseTTAFold that simultaneously generates protein sequences and structures. Beginning from a noised sequence representation, PG generates sequence and structure pairs by iterative denoising, guided by desired sequence and structural protein attributes. We designed thermostable proteins with varying amino acid compositions and internal sequence repeats and cage bioactive peptides, such as melittin. By averaging sequence logits between diffusion trajectories with distinct structural constraints, we designed multistate parent–child protein triples in which the same sequence folds to different supersecondary structures when intact in the parent versus split into two child domains. PG design trajectories can be guided by experimental sequence–activity data, providing a general approach for integrated computational and experimental optimization of protein function.

Protein function arises from a complex interplay of sequence and structural features; hence, designing new protein functions requires reasoning over both sequence and structure space. Many protein design methods sample structures and sequences in separate steps, typically by generating protein backbones first and using inverse folding methods to generate sequences. Traditional methods, such as Rosetta flexible backbone protein design[1], alternate between structure and sequence design, whereas recent deep-learning-based approaches[2–5] typically generate backbones first and then use sequence design methods, such as ProteinMPNN (MPNN), to identify sequences that fold into a given backbone[6,7]. Among the latter class of approaches, denoising diffusion probabilistic models[8] (DDPMs), which have shown considerable promise in continuous data domains, allow for the generation of protein backbones subject to a wide range of structural constraints[9]. DDPMs approximate the probability density function over a data distribution by learning to denoise samples corrupted with Gaussian noise,

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA, USA. [4]Department of Molecular Engineering, University of Washington, Seattle, WA, USA. [5]Molecular & Cellular Biology, Medical Scientist Training Program, University of Washington, Seattle, WA, USA. [6]Faculty of Engineering Sciences, Heidelberg University, Heidelberg, Germany. [7]School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA, USA. [8]Office of the Chief Scientific Officer, Microsoft, Redmond, WA, USA. [9]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [10]These authors contributed equally: Sidney Lyayuga Lisanza, Jacob Merle Gershon, Samuel W. K. Tipps, Jeremiah Nelson Sims, Lucas Arnoldt. e-mail: dabaker@uw.edu

enabling the generation of high-quality samples from a Gaussian prior; they have been explored less in categorical domains, such as text and protein sequences[10]. Although powerful, structure-based approaches, such as RFdiffusion[11] and Chroma[12], provide limited opportunities to guide protein generation using sequence-based features and to identify sequences with more than one fold and/or function. Hallucination approaches that apply activation maximization to structure prediction networks[13,14] can generate sequence–structure pairs without additional training, but these solutions can be adversarial and require a large number of steps to converge, and robust experimental success requires subsequent sequence design on the hallucinated backbones[7].

We reasoned that carrying out diffusion in sequence space rather than structure space could enable guidance of design using sequence-based features and explicit design of sequences populating multiple states. To enable conditioning on both sequence and structure features, we start from the RoseTTAFold[15] structure prediction network, which we treat as a mapping from input sequence and structure information to an output sequence and structure, as in the case of RFdiffusion (Fig. 1a and Supplementary Fig. 1). We reasoned that RoseTTAFold could be adapted for sequence space diffusion by noising the sequences of proteins in the Protein Data Bank (PDB; http://www.rcsb.org/) and training to remove the noise while imposing a loss on structure prediction accuracy, thus ensuring that the resulting model has a deep understanding of both sequence and structure.

Results

Categorical DDPM implementation and fine-tuning

We implement diffusion and data noising in categorical space by representing protein sequences as scaled one-hot tensors (for native sequences, true values are set to 1 and all other values set to −1) and embed via a linear layer, allowing for progressive corruption with Gaussian noise $N(\mu = 0, \sigma = 1)$[16,17]. This approach has the advantage over carrying out diffusion in a learned embedding space[10,18] of simplifying the use of raw sequence-based classifiers for guidance. To fine-tune RoseTTAFold, we input protein sequences progressively noised according to a square root schedule[10], the corresponding timestep and optional structural information and train the model to generate ground truth sequence–structure pairs by applying a categorical cross-entropy loss to the predicted sequence (relative to the ground truth sequence) and FAPE[19] structure loss on the predicted structure (Algorithm S1 and Supplementary Fig. 2). Self-conditioning[16] was found to improve training and inference performance. Protein generation begins with an L×20 dimensional sequence of Gaussian noise and a black-hole[19] initialized structure; at each timestep ($\mathbf{x}_t$), the model predicts $\mathbf{x}_0$ from $\mathbf{x}_t$, after which $\mathbf{x}_0$ is noised to $\mathbf{x}_{t-1}$ (Fig. 1b). Sequence-based guidance can be combined with $\mathbf{x}_0$ to guide the model toward a constrained sequence space using activity data, sequence-specific potentials or other information (Fig. 1b)[20]. Fixed motifs in the input sequence are featurized with an extra token to denote that the sequence is not diffused at this position. Secondary structure conditioning information is passed via the one-dimensional (1D) track, whereas three-dimensional (3D) coordinates are embedded via the pair features in the two-dimensional (2D) track and coordinates in the 3D track. Embeddings from these three tracks are linked by cross-attention in the RoseTTAfold architecture, allowing for the output sequence from the 1D track to condition the others.

During inference, we obtain $\mathbf{x}_0$ from $\mathbf{x}_t$ and generate $\mathbf{x}_{t-1}$ by noising $\mathbf{x}_0$; we found sampling $\mathbf{x}_{t-1} \sim q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)$[10] more effective than sampling from $\mathbf{x}_{t-1} \sim q(\mathbf{x}_{t-1} \mid \mathbf{x}_0, \mathbf{x}_t)$[8] (Supplementary Fig. 3 and Algorithm S2). ProteinGenerator (PG) outperforms early hallucination methods in unconditional design accuracy and generates structurally diverse proteins when sampled from different Gaussian mixture models (Supplementary Figs. 4–6). PG readily designs proteins that scaffold specified structural motifs; AlphaFold2 (ref. 19) (AF2)-predicted structures accurately recapitulate (root mean square deviation (RMSD) to design < 2, motif RMSD < 1, AF2 pAE < 5) both the motif and the full design (Fig. 4c and Supplementary Fig. 8). RFdiffusion followed by MPNN performs better on motif scaffolding and unconditional generation of larger proteins (Supplementary Fig. 5c). PG sequence quality as measured by ESM[21] pseudo-perplexity, which was previously shown to be indicative of experimental success[22,23], is indistinguishable from those of native sequences sampled from UniProt[24] and considerably higher than sequences generated using a 640-million-parameter sequence diffusion model, EvoDiff[25] (Supplementary Fig. 5d,e).

Unconditional generation using PG yields sequence–structure pairs with amino acid compositions resembling those of native proteins (Supplementary Fig. 7). Structure predictions from AF2 and ESMFold[21] of the generated sequences are close to the designed structures and confident (6% of designs have AF2 confidence pLDDT > 90 and RMSD < 2 Å; Supplementary Figs. 3 and 5; computational success rates for all design tasks can be found in Supplementary Table 1). We experimentally characterized unconditionally generated 70–80-residue proteins by testing solubility and monomericity via size-exclusion chromatography (SEC), folding by circular dichroism (CD) and stability by CD thermal melts (we use these methods throughout this study to evaluate protein behavior as the design criteria increase in complexity). Of the 42 proteins experimentally tested, 32 were soluble and monomeric by SEC, and CD experiments showed that they had the designed secondary structure and were stable up to 95 °C (Supplementary Fig. 9).

Design of rare amino acid enriched proteins

An advantage of diffusion in sequence space is that sequence-based guiding functions can be readily implemented and applied. To evaluate the ability of PG to reason over sequence–structure relationships outside the PDB training distribution, we sought to design proteins enriched in evolutionarily undersampled amino acids that confer structural or functional properties (Fig. 2a). Given a specification of the desired amino acid content, at each denoising step, sequence positions are ranked based on the frequency of the amino acid of interest, and, for the top N positions (where N is the number of desired occurrences of the amino acid), a bias toward the desired amino acid is added to the update generating $\mathbf{x}_{t-1}$ (Algorithm S3). We used this procedure to generate proteins with high frequencies (20% composition) of tryptophan, cysteine, valine, histidine and methionine (Fig. 2b and Supplementary Fig. 10a–e) with sequences very distinct from those of native proteins (Fig. 2c). Generated designs were filtered for high AF2 confidence (pLDDT > 90) and self-consistency (RMSD to design < 2 Å), and 96 were selected for experimental characterization.

Of the expressed designs, 68 were soluble in *Escherichia coli*, and four of five upweighted cysteine proteins, eight of 19 upweighted tryptophan proteins, 19 of 22 upweighted valine proteins, 10 of 12 upweighted histidine proteins and 10 of 10 upweighted methionine proteins were found to be monomeric by SEC (Supplementary Fig. 10g). CD spectra were obtained for a subset of the monomeric designs, and, in all cases, the indicated secondary structure was consistent with the design and thermostable (Fig. 2e–h). Guiding for high cysteine content at the sequence level resulted in the formation of 3–4 disulfide bonds per protein without any structural conditioning, as indicated by mass spectrometry in the presence and absence of the reducing agent TCEP at 50 mM (Fig. 2e and Supplementary Figs. 11–15). Proteins designed with upweighted tryptophans exhibited high absorbance at 280 nm and had helical CD traces (Fig. 2f). Proteins with upweighted valine exhibited higher beta-sheet content (Supplementary Fig. 10f) by CD, as expected given the secondary structure propensity of valine[26], and were thermostable (Fig. 2h). These results indicate that the model can reason over sequence–structure relationships beyond native protein-like sequence compositions to design folded, thermostable proteins with desired sequence properties.

We further explored the generation of proteins with pre-specified charge composition, isoelectric points and hydrophobicity[27].

https://doi.org/10.1038/s41587-024-02395-w

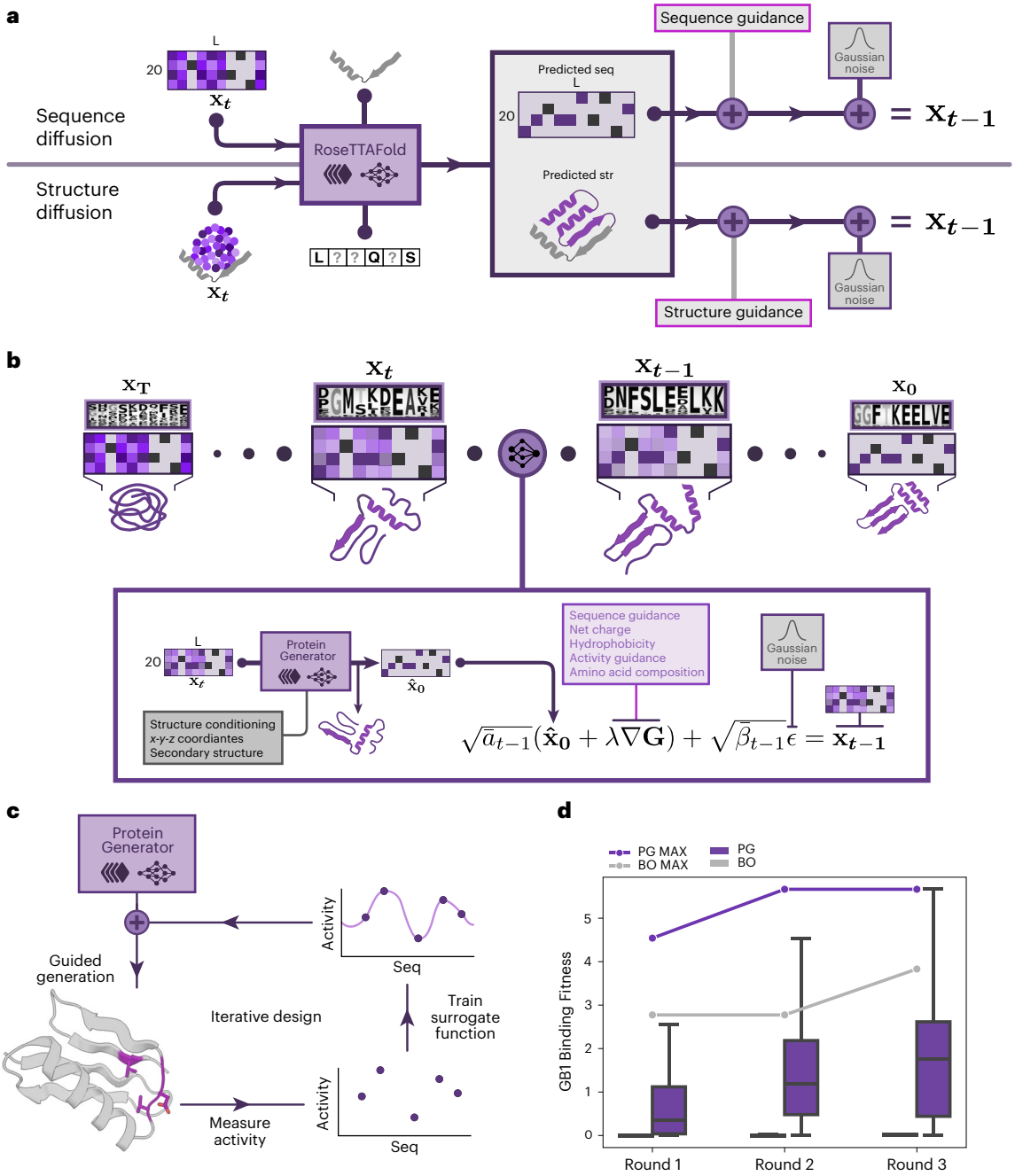


Fig. 1 | Overview of PG. a, Comparison of diffusion in sequence and structure space. PG and RFdiffusion take as input noised sequence (PG) or structure (RFdiffusion) data and problem specific sequence and structure constraints. At each denoising step, the RoseTTAFold architecture generates complete protein sequences and structures, and this is used to generate the next step in the trajectory in sequence (PG) or structure (RFdiffusion) space. Although specific structural or sequence features can be fixed in the input to RoseTTAFold in both approaches, biases toward particular sequence features during the diffusion update at each step are more readily incorporated in PG (as are biases toward structural features, such as symmetry, in RFDiffusion). **b**, Schematic of PG inference trajectory. At each step in the diffusion process the sequence $\mathbf{x}_0$ is predicted from sequence $\mathbf{x}_t$ by RF conditioned on any desired structural information, combined with any desired sequence bias, and noised to generate the $\mathbf{x}_{t-1}$. This process is repeated for T steps as the sequence–structure pair converges on a high-confidence solution shaped by the structural and sequence guidance information. **c**, Iterative design schematic demonstrating how PG can be used in an experimental feedback loop. Designs generated by the model are evaluated for activity; a surrogate function approximating sequence to function relationships is fit; and gradients from the surrogate can then be used to guide PG toward active design space. **d**, In silico demonstration of iterative design using GB1 fitness landscape for binding and comparison with Bayesian optimization (BO). In round 0, not shown in the plot, 96 designs are generated with PG without guidance, and a surrogate function is trained to discriminate high and low activity designs. In rounds 1–3, gradient-based guidance is used to generate 96 designs for each method; a surrogate function is fit; and the process is repeated. Line plots show maximum activity sampled, and box plots show distribution sampled over the batch of 96. Mean activities for each round are statistically significant between the two populations ($P < 0.05$, two-sided Mann–Whitney U-test, $n = 96$ designs per round). Box plots boundaries indicate upper and lower quartiles, and whiskers indicate the nearest quartile + 1.5× interquartile range. seq, sequence; str, structure.

https://doi.org/10.1038/s41587-024-02395-w

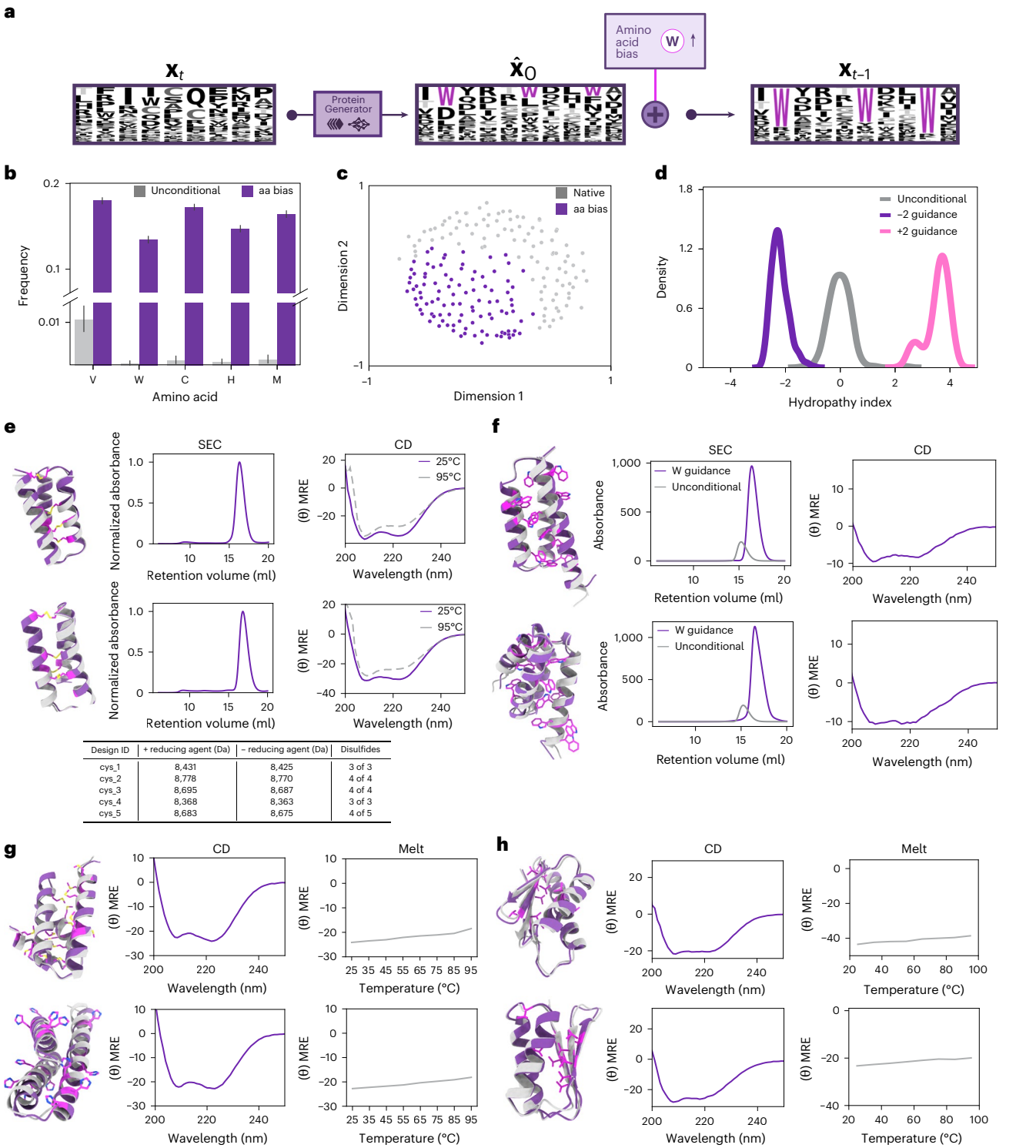


Fig. 2 | Design of proteins with specified sequence composition. a, Amino acid compositional bias schematic. **b**, Comparison of amino acid frequency in unconditional (gray) and amino acid biased (purple) generation; separate PG trajectories were carried out for each enriched amino acid. Error bars are standard deviation. Biased distributions are significantly different from unconditional amino acid frequencies ($P < 0.05$, two-sided Mann–Whitney U-test, $n = 200$ designs per amino acid). Box plot boundaries indicate upper and lower quartiles; whiskers indicate the nearest quartile + 1.5× interquartile range; and the center line is the median. **c**, Multidimensional scaling of native and amino acid biased sequences shows that they occupy distinct regions of sequence space. **d**, Hydropathy guidance. Biasing the sequence toward or away from hydrophobic amino acids results in a shifted distribution of hydropathy scores compared to unconditional generation ($P < 0.05$ two-sided Mann–Whitney U-test, $n = 122$ designs per condition). **e**, Experimental validation of cysteine biased designs (design in gray, AF2 in purple). Proteins are monomeric by SEC and alpha helical by CD at 25 °C and 95 °C. Mass spectrometry indicates the presence of the designed number of disulfide bonds. **f**, Experimental validation of tryptophan biased designs (design in gray, AF2 in purple). Designs are monomeric by SEC, have considerably higher absorbance at 280 nm than unconditional designs and are alpha helical by CD. **g**, Experimental validation of histidine and methionine biased designs (design in gray, AF2 in purple). **h**, Experimental validation of valine biased designs (design in gray, AF2 in purple). Valines highlighted in pink on the designs are present in the beta-fold secondary structure. CD traces and melt curves at 222 nm are to the right of the designs. CD traces and melt curves at 222 nm are to the right of the designs. aa, amino acid.

https://doi.org/10.1038/s41587-024-02395-w

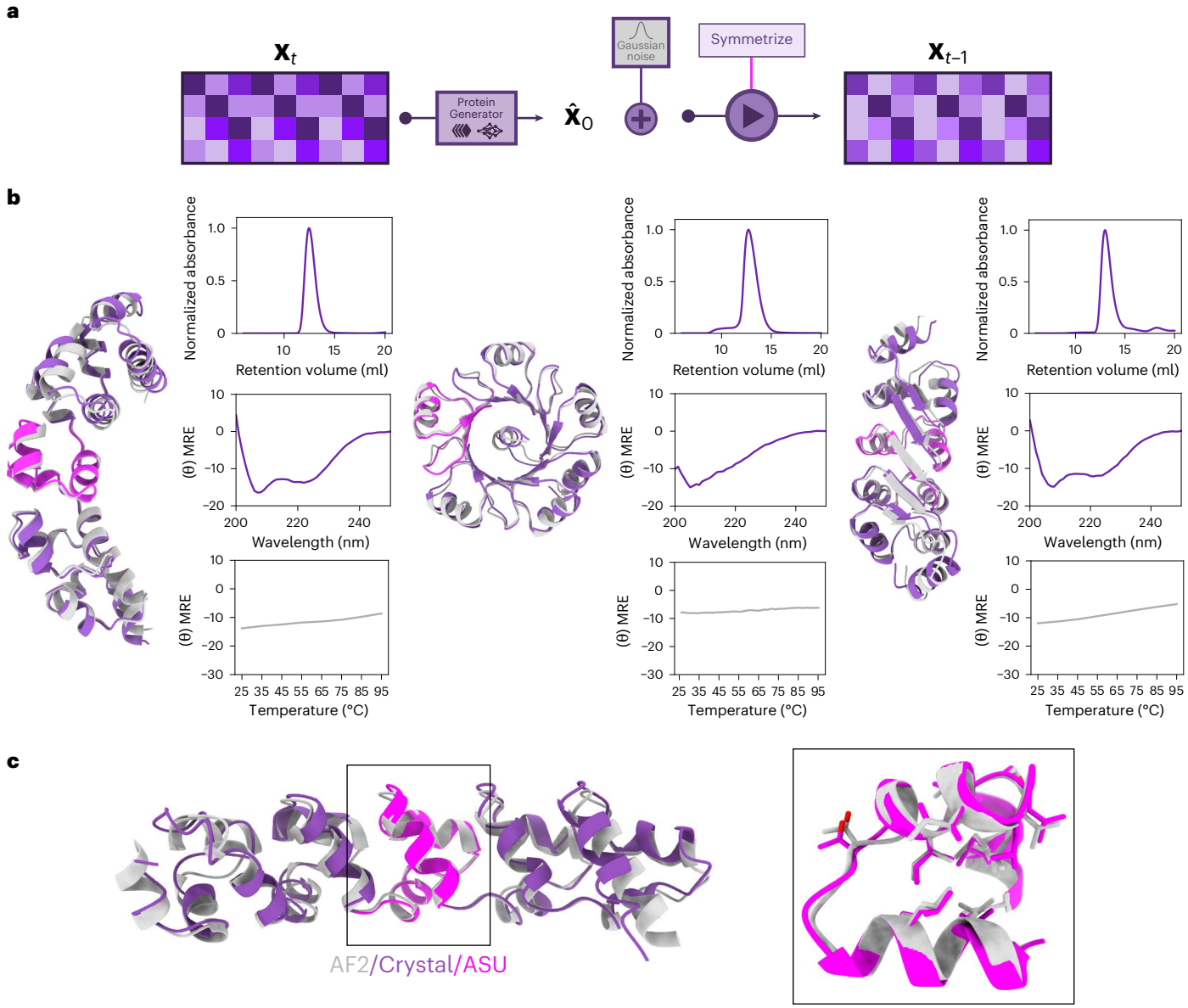


Fig. 3 | Design of sequence repeat proteins with PG. a, Symmetric sequence diffusion to design proteins with sequence symmetry. **b**, Experimental validation of sequence repeat proteins. Designs in gray are overlaid with AF2 predictions in purple, and asymmetric units are highlighted in pink. SEC and CD traces and melting curves demonstrate stability of these designs. **c**, 3.70-Å crystal structure of designed repeat protein: AF2 model in gray, crystal structure in purple and asymmetric unit in pink. Box on the right highlights the accuracy of designed side chains in the asymmetric unit.

We implemented sequence-based potentials to guide[20] the diffusive process toward these characteristics to enable fine-tuned control over physical properties of the output sequence. This approach enabled the design of proteins with a range of user-defined hydrophobicities (Fig. 2c) and isoelectric points (Supplementary Fig. 10h). The ability of PG to control sequence properties during backbone generation should be useful for increasing the developability of therapeutic candidates[28].

Design of sequence repeat proteins

Repeat proteins containing tandem copies of a sequence–structure unit are ubiquitous in nature and play central roles in molecular recognition and signaling[29]. Previous repeat protein design work has required pre-specification of structural features[30] or expensive Markov chain Monte Carlo (MCMC) calculations[7]. We reasoned that PG could be adapted readily to generate repeat proteins given only the sequence length of the repeat unit and number of repeats desired by, at each timestep, applying repeat symmetry to the noised sequence distribution (Fig. 3a). Unconditional generation with this approach yielded largely beta-solenoid structures. To encourage further exploration, we trained PG to condition on secondary structure (Supplementary Fig. 16) and specified secondary structure constraints to yield a wide range of all-alpha, all-beta and mixed alpha/beta designs (Fig. 3b). We added helical caps[31] to a subset of designs to promote stability and reduce aggregation[32,33]. We experimentally characterized 74 repeat proteins with helical caps and 86 repeat proteins without helical caps. Of these, 27 repeats with caps and 10 repeats without helical caps were soluble and monomeric by SEC, and seven of eight proteins evaluated using circular dichroism had the expected secondary structure (Fig. 3b and Supplementary Fig. 17). We solved the crystal structure of a five-repeat unit design composed of a four-helix bundle asymmetric unit and found the design to have atomic accuracy: the C RMSD of design to the crystal structure was 1.38 Å for the whole structure and 0.47 Å for the asymmetric unit (Fig. 3c and Table 1).

Table 1 | Data collection and refinement statistics for the repeat protein crystal structure

	CAP repeat (PDB: 8VD6)
Resolution range	72.08–3.70 (4.05–3.70)
Space group	*P*21
Unit cell	21.62, 45.05, 72.11
	90.00, 91.55, 90.00
Unique reflections	1,543 (359)
Multiplicity	2.0 (2.0)
Completeness (%)	99.8 (99.7)
Mean I/sigma(I)	11.6 (7.6)
Wilson B-factor	76.23
R-merge	0.047 (0.093)
CC1/2	0.997 (0.974)
Reflections used in refinement	1,534 (359)
Reflections used for R-free	79 (79)
R-work	0.2465 (0.2465)
R-free	0.2847 (0.2847)
Number of non-hydrogen atoms	1,384
Macromolecules	1,384
Protein residues	174
RMS (bonds)	0.002
RMS (angles)	0.59
Ramachandran favored (%)	88.95
Ramachandran allowed (%)	7.56
Ramachandran outliers (%)	3.49
Average B-factor	62.38
macromolecules	62.387

Statistics for the highest-resolution shell are shown in parentheses.

Design of conditionally active peptide cages for membrane lysis

The design of proteins with activities conditional on an external input is of considerable interest for the design of therapeutics[34] and biosensors[35] with spatial and temporal control. We used PG to address this challenge by scaffolding bioactive peptide sequences within an inert protein cage (Fig. 4a) by designating a region of the protein chain (usually at the N or C terminus) to be fixed at the bioactive peptide sequence and freely diffusing the remainder of the sequence. Unlike the previous LOCKR[36,37] sensor system, in which the bioactive sequence must be in a helical conformation and make specific interactions with the caging scaffold, neither the structure the peptide adopts nor the structure of the scaffold needs to be pre-specified, enabling caging of a much broader range of peptide sequences (Fig. 4b). Given the peptide sequence and scaffold length, PG generates designs that contain the peptide sequence as an integral part of the protein structure and are predicted to fold with greater than 85 pLDDT and less than 2-Å RMSD to the designed scaffold (Fig. 4c). We used this approach to design proteins caging the pore-forming peptide melittin that can be conditionally released upon proteolytic cleavage of a terminal loop. We specified the sequence of the bioactive peptide melittin with an adjacent furin cleavage site and used secondary structure conditioning to scaffold the peptide in a helical bundle with the cleavage site in a loop to improve protease access (Fig. 4b). Because of the multiple constraints (scaffolding the melittin sequence in an ordered structure and a furin cleavage sequence in a loop), this required increased sampling and filtering (Supplementary Table 1). Despite melittin being disordered in isolation, PG was able to generate solutions with the melittin sequence adopting a helical structure, which we then experimentally tested. Of 13 experimentally characterized designs, five were soluble and monodisperse by SEC, folded with helical secondary structure by CD and thermostable (Fig. 4e,f). We extended the cleavage loop and inserted interface arginine mutations to promote disassociation of the peptide after cleavage (Fig. 4f). Upon addition of furin protease, a band shift from ~18 kD to ~15 kD indicated cleavage of our designed melittin cage (Fig. 4g). Mass spectrometry analysis confirmed release of intact melittin peptide (Fig. 4h and Supplementary Fig. 18). To test conditional membrane lysis of our caged melittin protein, we incubated red blood cells (RBCs) with design D12 in the presence or absence of furin protease and measured absorbance at 450 nm to quantify the presence of heme from lysed RBCs. Samples pre-incubated with furin protease were bright red in color, indicating membrane lysis, whereas little lysis was observed without furin before treatment (Fig. 4l and Supplementary Fig. 19). Due to the bioavailability of endogenous endosomal proteases, such as furin, we anticipate that the design of caged peptides will enable a route toward endosomal escape.

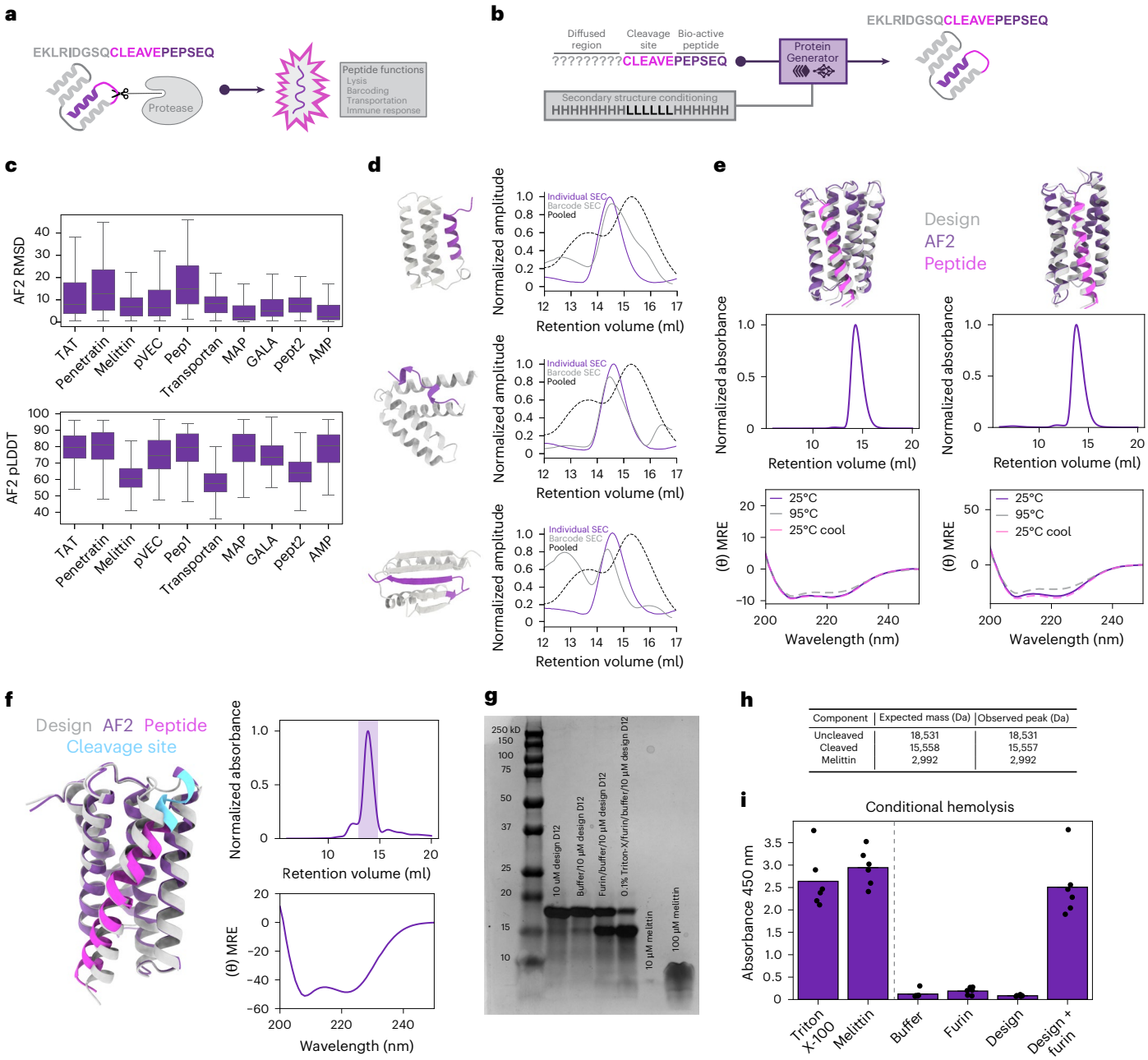
Scaffolding barcode peptide sequences

Peptide barcoding enables large libraries of proteins to be screened in binding assays or SEC, with the identity of individual proteins subsequently read out by mass spectrometry of the barcode after release by proteolysis[38]. Because of the challenge of incorporating a barcode within a folded protein, current approaches attach peptide barcodes to proteins of interest through N-terminal or C-terminal flexible fusions, but these can affect both expression and solubility, and, hence, it can require some 5–10 degenerate barcodes to infer the behavior of the untagged protein. Given the promising results of our bioactive peptide scaffolding experiments, we reasoned that PG could scaffold short barcode-like sequences (7–14 residues) into a protein of interest, thereby removing the need for multiple extrinsic barcodes. We scaffolded a set of validated peptide barcodes[7,39,40] at the C-terminus, flanked by lysine and arginine on the N-termini and C-termini, respectively, to permit facile tandem cleavage with Lys-C and trypsin. PG generates designs that contain the peptide sequence as an integral part of the protein structure and are predicted to fold with reasonable AF2 confidence and RMSD to the design (Fig. 4c and Supplementary Fig. 20). We cloned, expressed and purified a pilot library of 84 pooled designs and used SEC to separate them by size. Barcodes were isolated from each fraction and run on an Orbitrap Lumos to determine the identities of the proteins in each fraction, as described previously[7,39,40]. As a control, we expressed and purified each of these 84 proteins and subjected them to SEC; of these, 64 of 84 (76%) were expressed, and 48 of 64 (75%) of the expressed proteins exhibited monodisperse elution peaks at the expected elution volume (Supplementary Fig. 21a). We overlaid individual SEC (iSEC) elution profiles of expressing designs with corresponding reconstructed iSEC barcode traces (73/84) normalized to maximum MS2 intensity (SEC-MS) for a set of 58 designs. We observed close agreement between SEC-MS and iSEC peak elution volume for 41 of 58 (71%) designs (Supplementary Fig. 21b); exemplary traces are shown in Fig. 4d and Supplementary Fig. 22. Thus, PG can incorporate intrinsic barcodes into protein libraries, resulting in shorter constructs than required with external barcoding approaches.

Multistate design

Designing an amino acid sequence that can adopt distinct structural conformations upon an external trigger is a challenging task, as the energy landscape must contain two discrete minima with free energy differences small enough for a trigger to induce state switching[41]. We reasoned that PG was well equipped for explicit multistate design because guidance can be applied from multiple condition-dependent

https://doi.org/10.1038/s41587-024-02395-w



Component	Expected mass (Da)	Observed peak (Da)
Uncleaved	18,531	18,531
Cleaved	15,558	15,557
Melittin	2,992	2,992

Fig. 4 | Scaffolding bioactive peptides and intrinsic barcodes with PG. **a**, Schematic overview of functional peptide scaffolding for downstream tasks such as protease cleavage for lysis and peptide barcoding. **b**, Sequence-only motif scaffolding and secondary structure conditioning to generate proteins with embedded functional sequences. Cleavage sites can be specified at the N or C terminus of the peptide to allow for protease cleavage. **c**, In silico design metrics for sequence-only bioactive peptide scaffolding. RMSD of AF2 predictions to designs on the top and AF2 pLDDT of designs on the bottom. Box plot boundaries indicate upper and lower quartiles; whiskers indicate the nearest quartile + 1.5× interquartile range; and the center line is the median. $n = 2{,}000$ designs per condition. **d**, Mass spec peptide barcoding assay. Scaffolding barcodes with PG results in soluble and monomeric designs by SEC. SEC traces for individual designs are in gray. When the same designs are expressed in a pooled library (black), and fractions are digested with trypsin, analytical mass spectroscopy of each fraction is able to recapitulate the SEC trace shown in purple. **e**, Melittin scaffolded designs with furin cleavage site. Designs are shown in gray, and AF2-predicted structures are shown in purple, with melittin peptide highlighted in pink. Designs are soluble and monomeric by SEC and folded with helical secondary structure by CD. **f**, Melittin scaffolded design D12. D12 design model is in gray; AF2-predicted structure is overlayed in purple for scaffold; cyan is for the cleavage site; and pink is for melittin. SEC fraction of monomeric D12 used for downstream assays is highlighted with the purple bar. CD trace of D12 is consistent with the designed helical secondary structure. **g**, Representative SDS-PAGE of uncleaved D12 (18 kD), cleaved D12 (15 kD) and melittin peptide (3 kD) ($n = 3$ biological replicates). **h**, Mass spec of the cleavage reaction products confirms the presence of uncleaved D12, cleaved D12 and melittin. Melittin mass was calculated with an additional c-terminal 'GS' due to the expression vector used. **i**, Absorbance at 450 nm for six technical replicates of washed RBCs after incubation with design with and without furin protease. Positive controls Triton X-100 and melittin are shown to the left of the vertical bar. Design with furin lyses RBCs significantly more than samples without design ($P = 0.002$, two-sided Mann–Whitney U-test) or furin ($P = 0.005$, two-sided Mann–Whitney U-test) and is on par with positive controls Triton X-100 ($P = 0.127$, two-sided Mann–Whitney U-test) and melittin ($P = 0.132$, two-sided Mann–Whitney U-test).

https://doi.org/10.1038/s41587-024-02395-w

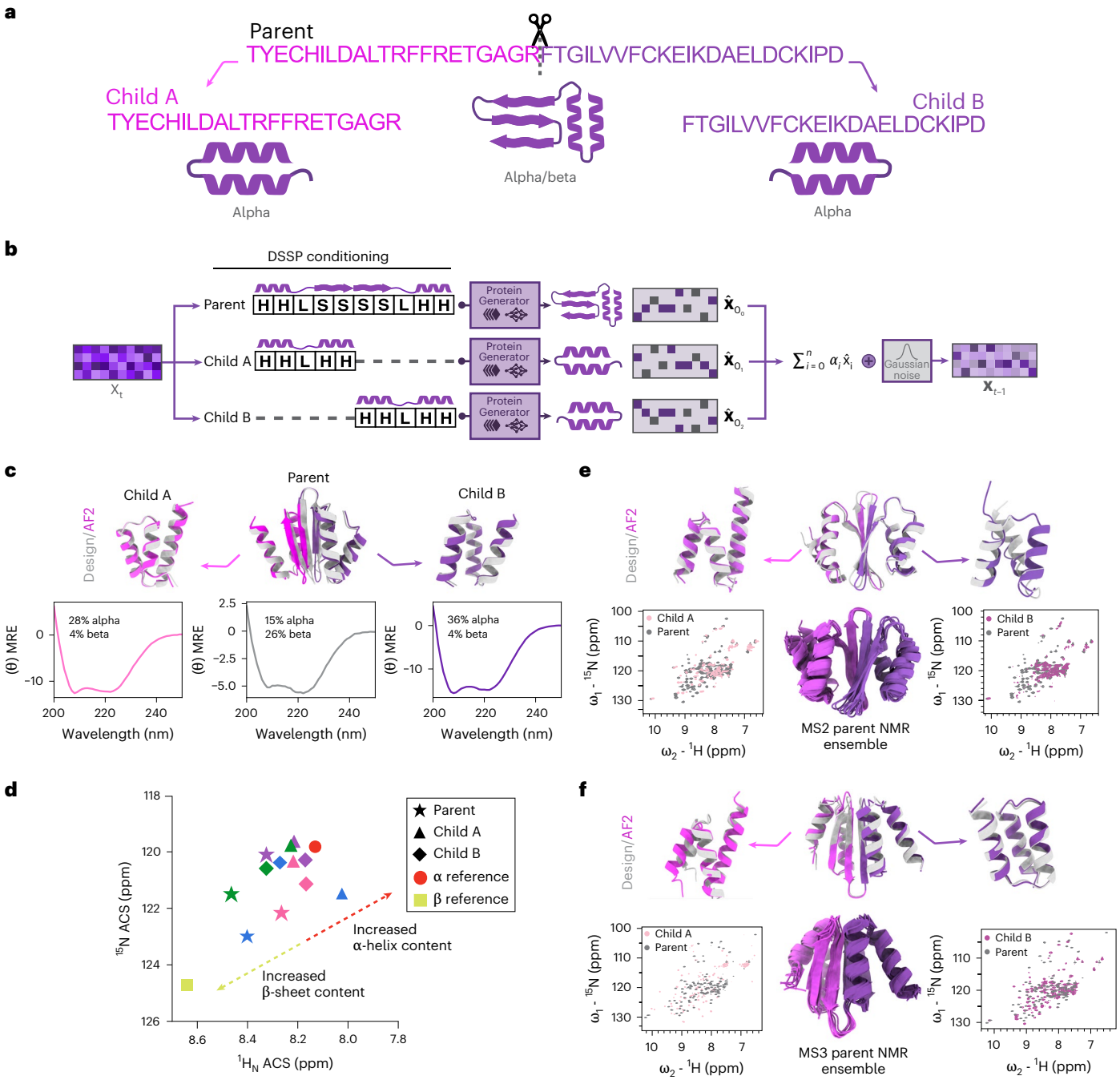


Fig. 5 | Multistate design with PG. a, Multistate DSSP conditioning is used to generate a sequence with an alpha/beta fold in the parent state and all alpha in the child A and child B states. **b**, Implementation of multistate DSSP sequence conditioning. Different DSSP conditioning strings are applied to a full-length parent sequence and two subsequences (child A and child B). RoseTTAFold predictions and model logits are output for parent, child A and child B. A linear combination of output logits is used as a potential to guide the model toward finding one sequence that satisfies all DSSP conditioning strings for parent, child A and child B. **c**, MS1 family adopts distinct folds by CD. Top, high pLDDT design and AF2 models of family MS1. Bottom, CD spectra and deconvolution of family MS1 indicating 26% beta content in the parent compared to 4% beta content in child A and child B, respectively. **d**, ACS of $^{1}H_{N}$ and ^{15}N chemical shifts values obtained from MS1–MS4 HSQC spectra. Reference average ACS values of primarily α-helical proteins (red circle) and primarily β-sheet proteins (yellow square) are shown calculated from $^{1}H_{N}$–^{15}N correlations using chemical shift information obtained from the Biological Magnetic Resonance Bank. ACS values are compared for multistate sequences among parent (α/β mix fold), child A (α-helical fold) and child B (α-helical fold). MS1 in pink, MS2 in purple, MS3 in blue, MS4 in green. MS2 (**e**) and MS3 (**f**) families are designed by PG to adopt distinct folds in the parent and child states with high AF2 confidence (top row). HSQC overlays of MS2 and MS3 child A and B compared to parent (bottom row; ω indicates chemical shift). NMR structures of MS2 and MS3 parent fold into the intended secondary structures with atomic-level accuracy (bottom middle).

structural constraints during the sequence diffusion process. To adapt PG to multistate design, we input to RoseTTAFold the same sequence but different structural conditioning information and take a linear combination of the output logits as input to the next timestep.

We used this approach to design protein sequences that adopt distinct folds when connected in a single chain (the parent) or when expressed separately or cleaved into two chains by a protease (child A and child B) (Fig. 5a). At each timestep, $\mathbf{x}_t$, we use RoseTTAFold to model

Table 2 | NMR structure statistics for designed multistate proteins

	MS3 parent (PDB: 8VL4)	MS2 parent (PDB: 8VL3)
Assignment completeness (backbone atoms N, HN, CA, CB, CO)	93%	94%
NMR distance and dihedral constraints		
Distance constraints		
Total NOEs	59	140
Intra-residue ($\|i–j\| = 0$)	0	0
Sequential ($\|i–jj\| = 1$)	43	65
Medium-range ($\|i–j\| > 1$ and $\|i–j\| < 5$)	9	63
Long-range ($\|i–j\| \geq 5$)	7	12
Intermolecular	0	0
Hydrogen bonds	0	0
Total dihedral angle restraints	192	196
Total residual dipolar coupling (RDC) restraints	0	0
Total chemical shift restraints	464	468
Structure statistics		
Violations		
Distance constraints (total no.)	8	11
Dihedral angle constraints (°)	0	0
Max. dihedral angle violation (°)	0	0
Max. distance constraint violation (Å)	1.65	0.74
Deviations from idealized geometry		
Bond lengths (Å)	0	0
Bond angles (°)	0	0
Impropers (°)	0	0
Average pairwise backbone root means squared (r.m.s.) ensemble (Å)	1.02	1.16
Ramachandran analysis		
Favored	98±1%	95±1%
Allowed	2±1%	5±1%
Disallowed	0±0%	0±0%

the full-length parent sequence along with the cleavage products child A and child B and average the resulting logits followed by noising to $\mathbf{x}_{t-1}$ (Fig. 5b and Algorithm S4). DSSP[42] features are appended to the L×20 sequence representation of each family member to enable conditioning on protein secondary structure (Fig. 5b). We used this approach to generate multistate sequences (MS) that are designed to adopt specific α/β folds in the parent state and different all α-helical folds in the child states; as expected given the greater problem complexity, this required more sampling than the above single sequence design problems (Supplementary Table 1).

We experimentally characterized 72 parent–child triples that AF2 predicted with high confidence and accuracy to be in the parent state when intact and the child states when split (Supplementary Fig. 23), and we selected 4 (MS1–MS4) soluble and monodisperse sequence families for detailed CD and nuclear magnetic resonance (NMR) studies (Supplementary Fig. 24a–d). 2D ^{1}H-^{15}N amide heteronuclear single quantum coherence (HSQC) spectra revealed that all the MS1–MS4 parents and children were well folded and globular proteins (Supplementary Fig. 24). CD spectra of all of the children were consistent with all-alpha proteins; spectral deconvolution suggested higher beta-sheet content in the parents (Fig. 5c and Supplementary Fig. 24, middle rows). As secondary structure estimation by CD can be imperfect, we took advantage of the fact that NMR chemical shifts are influenced by local secondary structure, which leads to distinct averaged chemical shift (ACS) values for primarily α-helical versus β-sheet proteins[43,44]. As expected for proteins with increased β-character, ^{1}H and ^{15}N ACS values of all parent designs were shifted downfield relative to the two associated children (Fig. 5d, dotted yellow arrow), which were shifted upfield into the reference region associated with α-helical proteins (Fig. 5d, dotted red arrow). Clear differences in chemical shift positions of MS1–MS3 child A and child B NMR peaks relative to the parent suggest that they adopt distinct folds (Fig. 5e,f; the differences were smaller for MS4 (Supplementary Fig. 24d), which the ACS values (Fig. 5d) and AF2 predictions (Supplementary Fig. 24e) suggest may not adopt a single fold). Taken together, these data suggest that, as intended, the designed child sequences fold into α-helical supersecondary structures distinct from the parent designs.

To assess the accuracy of the designed alpha/beta parent folds, we obtained high-resolution structures of MS2 parent and MS3 parent (Fig. 5e,f and Table 2). The solution NMR structure of MS2 parent is within 1.06 Å Cα RMSD of the design model, with the central beta-sheet nearly perfectly recapitulated. The solution NMR structure of MS3 parent is within 1.61 Å Cα RMSD of the design model, with the beta-sheet again very close to the design model. These high-resolution structures of the parents, together with the all-alpha helical ACS values and CD spectra of the children, strongly suggest that, as designed, there are large-scale structural rearrangements upon splitting of the polypeptide chain.

Guidance with experimental data

A longstanding goal in directed evolution is the optimization of desired functional attributes, such as enzyme activity, in as few experimental iterations as possible. We investigated the use of PG for experimental data-driven protein function optimization. As a test case, we used an experimental benchmark dataset on the IgG-binding protein GB1, which has the advantage of completeness: activity was measured for every sequence combination (for the four residues that were varied), allowing for evaluation of the activity of any sequence generated for these residues using PG. For other datasets where only a small fraction of sequences have known fitness, it is difficult to do such retrospective comparisons (because the PG-generated sequences will almost always have unknown fitness, like the vast majority of other possible sequences)[45].

We simulated an iterative guidance process by exploring the fitness landscape of the protein GB1 via gradient-based optimization with IgG binding activity-guided diffusion trajectories. At each step, we biased sampling by gradients from classifiers trained on the experimentally determined fitnesses of preceding rounds (Fig. 1c). As the classifiers, we used two-layer multilayer perceptrons (MLPs) optimized for GB1 fitness greater than 2. For broad applicability, we carried out this optimization with standard settings and no extensive hyperparameter turning (96 designs generated and tested per round for three rounds). For comparison, we tested multiple optimization approaches using different acquisition functions on this same problem and chose the best performer to compare to PG. We found that the average and maximum fitness of PG-generated designs increases each round, outperforming a Bayesian optimization baseline with the best identified acquisition function (batched upper confidence bound) (Fig. 1d and Supplementary Figs. 25 and 26). The improved performance likely reflects the rich prior understanding of protein sequence–structure relationships implicit in RoseTTAFold compared to the baseline, which has access to only limited experimental data. Our PG approach can readily incorporate any experimentally measurable fitness attribute and should be useful for machine-learning-assisted directed evolution campaigns[46–48].

Discussion

The in silico and experimental results presented here demonstrate that PG can readily generate a wide variety of de novo (Supplementary Fig. 27) proteins subject to diverse sequence domain constraints, including amino acid composition bias, repeat sequence symmetry, bioactive peptide caging and multistate design. In this section, we compare PG to RFdiffusion and highlight the areas where PG sequence space diffusion is particularly advantageous. Both PG and RFdiffusion take advantage of RoseTTAFold to jointly model protein sequences and structures, and, hence, both PG sequence space diffusion trajectories and RFdiffusion structure space diffusion trajectories can be guided by both sequence and structure information. For example, given 'hard constraints', such as the identities of amino acids at certain positions, or the 3D structure of part of a protein (for an enzyme active site, typically both types of constraints would be provided), both methods will generate proteins satisfying the constraints by providing the relevant sequence and structure input at each RoseTTAFold denoising step along with the current partially denoised sequence (PG) or structure (RFdiffusion). However, 'softer' sequence constraints, such as biases on the number of amino acids of a given type, are more readily implemented in sequence space diffusion, whereas global structural properties, such as overall symmetry, are more readily implemented in structure-based diffusion. As with other protein generative models, obtaining sequences predicted to satisfy specific problem constraints required considerable filtering; future work will seek to increase the fraction of generated sequences that are confidently predicted.

Sequence space diffusion is specifically advantageous in multistate design of protein sequences that fold to two or more distinct structures as in our parent child designs; this can be implemented in sequence space diffusion by logit averaging, but it is not straightforward with structure space diffusion. PG enables deep-learning-based multistate design through the joint search of sequence and structure space without assuming fixed structural priors and should be readily generalizable to the design of more complex conditional state-switching protein systems; evaluation of these designs will benefit from advances in methods for structure ensemble prediction[49–51]. Beyond multistate design, we expect PG, and, more broadly, generative methods enabling direct sequence based guidance, to be useful in generating successive rounds of sequences for experimental characterization in directed evolution campaigns[46,52]. Although classifiers trained on the available experimental data can be used to directly generate sequences using Bayesian optimization and other approaches, using these classifiers instead to guide PG diffusion trajectories has the considerable advantage of being informed by the rich sequence–structure prior information represented within the PG network, which increases the likelihood that the generated sequences will fold and function.

Online content

Any methods, additional references, Nature Portfolio 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/s41587-024-02395-w.

References

1. Huang, P.-S. et al. RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS ONE* **6**, e24109 (2011).
2. Wang, J., Watson, J. L. & Lisanza, S. L. Protein design using structure-prediction networks: AlphaFold and RoseTTAFold as protein structure foundation models. *Cold Spring Harb. Perspect. Biol.* **16**, a041472 (2024).
3. Winnifrith, A., Outeiral, C. & Hie, B. Generative artificial intelligence for de novo protein design. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2310.09685 (2023).
4. Chu, A. E., Lu, T. & Huang, P.-S. Sparks of function by de novo protein design. *Nat. Biotechnol.* **42**, 203–215 (2024).
5. Notin, P., Rollins, N., Gal, Y., Sander, C. & Marks, D. Machine learning for functional protein design. *Nat. Biotechnol.* **42**, 216–228 (2024).
6. Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
7. Wicky, B. I. M. et al. Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).
8. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. Preprint at arXiv https://doi.org/10.48550/arXiv.2006.11239 (2020).
9. Anand, N. & Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv.org* https://doi.org/10.48550/arXiv.2205.15019 (2022).
10. Li, X. L., Thickstun, J., Gulrajani, I., Liang, P. & Hashimoto, T. B. Diffusion-LM improves controllable text generation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2205.14217 (2022).
11. Watson, J. L., Juergens, D. & Bennett, N. R. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
12. Ingraham, J. B., Baranov, M. & Costello, Z. et al. Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
13. Anishchenko, I. et al. De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
14. Wang, J. et al. Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
15. Baek, M. et al. Efficient and accurate prediction of protein structure using RoseTTAFold2. Preprint at *bioRxiv* https://doi.org/10.1101/2023.05.24.542179 (2023).
16. Chen, T., Zhang, R. & Hinton, G. Analog Bits: generating discrete data using diffusion models with self-conditioning. Preprint at arXiv https://doi.org/10.48550/arXiv.2208.04202 (2022).
17. Han, X., Kumar, S. & Tsvetkov, Y. SSD-LM: semi-autoregressive simplex-based diffusion language model for text generation and modular control. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2210.17432 (2022).
18. Dieleman, S. et al. Continuous diffusion for categorical data. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2211.15089 (2022).
19. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
20. Dhariwal, P. & Nichol, A. Diffusion models beat GANs on image synthesis. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2105.05233 (2021).
21. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
22. Hie, B. L. et al. Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* **42**, 275–283 (2024).
23. Verkuil, R. et al. Language models generalize beyond natural proteins. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.21.521521 (2022).
24. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
25. Alamdari, S. et al. Protein generation with evolutionary diffusion: sequence is all you need. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.11.556673 (2023).
26. Fujiwara, K., Toda, H. & Ikeguchi, M. Dependence of α-helical and 13-sheet amino acid propensities on the overall protein fold type. *BMC Struct. Biol.* **12**, 18 (2012).
27. Boswell, C. A. et al. Effects of charge on antibody tissue distribution and pharmacokinetics. *Bioconjug. Chem.* **21**, 2153–2163 (2010).

28. Gruver, N. et al. Protein design with guided discrete diffusion. *Adv. Neural Inf. Process. Syst.* **36**, 12489–12517 (2023).
29. Parmeggiani, F. & Huang, P.-S. Designing repeat proteins: a modular approach to protein design. *Curr. Opin. Struct. Biol.* **45**, 116–123 (2017).
30. Brunette, T. J. et al. Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015).
31. Zorine, D. & Baker, D. De novo design of alpha-beta repeat proteins. Preprint at *bioRxiv* https://doi.org/10.1101/2024.06.15.590358 (2024).
32. Peralta, M. D. R. et al. Engineering amyloid fibrils from 13-solenoid proteins for biomaterials applications. *ACS Nano* **9**, 449–463 (2015).
33. MacDonald, J. T. et al. Synthetic beta-solenoid proteins with the fragment-free computational design of a beta-hairpin extension. *Proc. Natl Acad. Sci. USA* **113**, 10346–10351 (2016).
34. Zeng, Z. et al. Customized reversible stapling for selective delivery of bioactive peptides. *J. Am. Chem. Soc.* **144**, 23614–23621 (2022).
35. Azoitei, M. L. et al. Spatiotemporal dynamics of GEF-H1 activation controlled by microtubule- and Src-mediated pathways. *J. Cell Biol.* **218**, 3077–3097 (2019).
36. Lajoie, M. J. et al. Designed protein logic to target cells with precise combinations of surface antigens. *Science* **369**, 1637–1643 (2020).
37. Quijano-Rubio, A. et al. De novo design of modular and tunable protein biosensors. *Nature* **591**, 482–487 (2021).
38. Egloff, P. et al. Engineered peptide barcodes for in-depth analyses of binding protein libraries. *Nat. Methods* **16**, 421–428 (2019).
39. Kim, D. E. et al. De novo design of small beta barrel proteins. *Proc. Natl Acad. Sci. USA* **120**, e2207974120 (2023).
40. Gerben, S. R. et al. Design of diverse asymmetric pockets in de novo homo-oligomeric proteins. *Biochemistry* **62**, 358–368 (2023).
41. Wei, K. Y. et al. Computational design of closely related proteins that adopt two well-defined but structurally divergent folds. *Proc. Natl Acad. Sci. USA* **117**, 7208–7215 (2020).
42. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
43. Mielke, S. P. & Krishnan, V. V. Characterization of protein secondary structure from NMR chemical shifts. *Prog. Nucl. Magn. Reson. Spectrosc.* **54**, 141–165 (2009).
44. Shen, Y. & Bax, A. Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods Mol. Biol.* **1260**, 17–32 (2015).
45. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5**, e16965 (2016).
46. Wittmann, B. J., Yue, Y. & Arnold, F. H. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst.* **12**, 1026–1045 (2021).
47. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl Acad. Sci. USA* **116**, 8852–8858 (2019).
48. Hie, B. L. & Yang, K. K. Adaptive machine learning for protein engineering. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2106.05466 (2021).
49. Chakravarty, D. & Porter, L. L. AlphaFold2 fails to predict protein fold switching. *Protein Sci.* **31**, e4353 (2022).
50. Wayment-Steele, H. K. et al. Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).
51. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
52. Arnold, F. H. Directed evolution: bringing new chemistry to life. *Angew. Chem. Int. Ed. Engl.* **57**, 4143–4148 (2018).

Methods

Sequence representation

To apply the diffusion framework in sequence space, a continuous representation of the categorical sequence data is needed. To implement this, we represented the sequence, $\mathbf{x}_0$, with dimensions L×20 where L corresponds to the protein length with 20 possibilities for each amino acid type. This takes the form of a one-hot encoded vector that is centered at zero by multiplying the L×20 tensor by 2 and subtracting 1. Each logit within the tensor is a real number, with higher values corresponding to a higher probability for that specific amino acid at that position. With this representation, we noise $\mathbf{x}_0$ to obtain $\mathbf{x}_t$ with the below equation following the Ho et al.[8] formulation for a standard forward process sampling from Gaussian noise with mean at 0 and standard deviation of 1.

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

A critical part of the forward diffusion process is selecting the noising schedule. Determining the correct bin of a categorical distribution is trivial at low timesteps by argmaxing the input sequence. Therefore, more noise should be present at low timesteps to increase the difficulty of the task during training. The square root noise schedule[10] satisfies this requirement and was employed in this study.

Training

To train the model, we began by sampling t uniformly from [0,T], where $t = 0$ is an un-noised sequence and $t = T$ is pure Gaussian noise. We then noise $\mathbf{x}_0$ to $\mathbf{x}_t$ with equation (1) and tasked the model to predict the un-noised sequence $\mathbf{x}_0$ and its corresponding structure $\mathbf{y}$. The timestep feature was added to the sequence template passed to the model. We applied a categorical cross-entropy loss to $\mathbf{x}_0$ and structure losses to $\mathbf{y}$ (FAPE, bond angle, bond length, distogram, lddt). An additional KL loss[10] was applied to the calculated $\mathbf{x}_{t-1}$, as previously demonstrated to stabilize training of discrete diffusion models[10]. Self-conditioning[16] was implemented to allow the model to condition on the previous $\mathbf{x}_0$ prediction and the back-calculated $\mathbf{x}_{t-1}$ during both training and inference. To self-condition in practice, the model was used with gradients turned off to first predict $\mathbf{x}_0$ from $\mathbf{x}_{t+1}$, which was then passed in as a sequence template to the model. During training, RoseTTAFold was allowed 1–3 uniformly sampled 'recycle' steps to refine structure predictions via multiple passes through the model[53]. Pseudo training and inference code is available in the Supplementary Information (Algorithms S1 and S2). In later training iterations, secondary structure conditioning was provided to the model by concatenating a tensor representing DSSP features onto the sequence template. These features were provided 25% of the time and masked uniformly between 0% and 90% when provided.

Along with the standard diffusion task (40% of the time), the model was also challenged with structure prediction (seq2str) and fixed backbone sequence design (30% of the time each). Incorporating these additional tasks during training helped maintain the agreement of sequence–structure pairs diffused by the model. Training examples were conditioned on sequence or structure by either unmasking 1–4 spans of residues, each 4–8 amino acids in length to simulate motif scaffolding, or unmasking randomly selected residues for the model to scaffold as an active site scaffolding problem. Unmasked structure conditioning information was supplied to the input for RoseTTAFold as templates in the 1D sequence track as well as the 2D and 3D structural information tracks.

Inference

During inference starting from $\mathbf{x}_t$, the model predicts $\mathbf{x}_0$ and simultaneously decodes it to $\mathbf{y}$. $\mathbf{x}_0$ is then back-calculated to $\mathbf{x}_{t-1}$ with equation (1) and passed through the network with the previously predicted $\mathbf{x}_0$ to apply self-conditioning. Benchmarking against conditioning on $\mathbf{x}_t$, as done in Ho et al.[8] with the below equation, shows that this approach performs better (Supplementary Fig. 3c), as seen in other categorical diffusion methods[10,17].

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}),$$

where $\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t$ and $\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$

This is done for T steps, but T can be varied and does not have to be what was used during training (inference time for fixed T can be found in Supplementary Table 2). The model finds solutions to some problems in as few as 10 steps (Fig. 1c). Furthermore, clamping the model's output logits from −3,3 gives better agreement with AF2 predictions (Supplementary Fig. 3b). $\mathbf{x}_{t-1}$ is sampled from either a zero-mean normal distribution or a non-Bayesian Gaussian mixture distribution with equal mixing probabilities. For the non-Bayesian Gaussian mixture models, we defined a mixture with two normals centered at [−1, 1] (GMM2) and a mixture with three normals centered at [−1, 0, 1] (GMM3).

Unconditional protein generation

Unconditionally generated proteins were assessed against a set of 1,000 native proteins with a length deviating up to five residues randomly sampled from the RCSB[54] database. For experimental verification, proteins ranging from 70 to 80 amino acids in length with no conditioning information were generated in 25 steps. Designs were filtered by AF2 pLDDT > 90 and AF2 RMSD to design < 2 Å for ordering final constructs. Additionally, proteins with high model confidence but moderate AF2 confidence were ordered by filtering on design pLDDT > 90, AF2 pLDDT < 80 and AF2 RMSD to design < 5 Å.

Compositionally biased protein generation

Proteins ranging from 70 to 80 amino acids in length with an amino acid compositional potential were generated in 25 steps. Designs were filtered by AF2 pLDDT > 90, AF2 RMSD to design < 2 Å and SAP score[55] < 30. The top 10–22 designs were ordered for each upweighted amino acid type (tryptophan, cysteine, valine, histidine and methionine). Pseudocode for the implementation of the amino acid compositional potential is provided in the supplements (Algorithm S3).

Charge biased protein generation

Proteins of 50 amino acids in length with charge potentials applied were generated in 25 steps with charge conditioning information. The ground truth charge for each protein was calculated at pH 7.4 by using the Henderson–Hasselbach equation.

Hydrophobic biased protein generation

Proteins of 50 amino acids in length with hydrophobic potentials applied were generated in 25 steps with hydrophobicity conditioning information. The ground truth hydropathy index for each design was calculated by summing the hydropathy index for each residue and dividing by the sequence length[56].

DSSP guidance

For constructing the DSSP features, we calculated each training example's DSSP based on the structure with helix, strand, loop and masked labels[57]. During training, the calculated per-residue secondary structure features were appended to RoseTTAFold's 1D features and were one-hot encoded for 25% or 50% of the time and masked for 30% or 80% of the time. During inference, DSSP features are appended to the 1D features as necessary and masked when not. Secondary structure representations were input to the model as follows: H, helix; E, sheet; L, loop; X, masked.

Repeat protein generation

Repeat proteins ranging from 125 to 150 amino acids in length were generated in 50 steps with and without DSSP conditioning information. Designed proteins contained five repeat units using one of

the following DSSP strings, where X represents mask, E represents strand and H represents helix:'XXXXEEEEEXXXXXXXXXXXXXX XHHHHHXXXX', 'XXXXEEEEEXXXXXHHHHHXXXXXEEEEEXXXX', 'XXXXHHHHHXXXXXEEEEEXXXXXHHHHHXXXX', 'XXXXHHHHHXXLXXHHHHHXXLXXHHHHHXXXX', 'XXXXEEEEEXXHXXEEEEEXXHXXEEEEEXXXX', 'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX', 'XXXXEXXXXXXXXXEXXXXXXXXXEXXXXXXXX'. Designs were filtered on AF2 pLDDT > 80 and AF2 RMSD to design < 2 Å.

Caging bioactive peptides

Proteins of 155 amino acids in length were generated in 25 steps with spans of helical DSSP conditioning to encourage the model to generate helical bundles and cleavage loops. Additional DSSP features were provided to scaffold the furin protease cleavage site 'GRRKR'. The sequence for melittin was provided without DSSP conditioning as N-terminal 26 amino acids 'GIGAVLKVLTTGLPALISWIKRKRQQ'. Designs were filtered by AF2 pLDDT > 85, AF2 RMSD to design < 2 Å and SAP scores < 40.

Scaffolding barcode peptide sequences

Proteins of 100 amino acids in length were generated in 50 steps with barcodes being set as fixed regions on the C-terminus. Designs were filtered by AF2 pLDDT > 85 and AF2 RMSD to design < 2 Å.

Multistate guidance

Parent and child protein pairings were generated in 25 steps using secondary structure features 'XHHHHHHHHHHHHHLLLHHHHHH HHHHHHHLLLEEEEEELLEEEEEXXXXXXXEEEEELLE EEEEELLLH HHHHHHHHHHHHLLLHHHHHHHHHHHHHX' for parent, 'XHHHH HHHHHHHHHLLLHHHHHHHHHHHHHLLLHHHHHHHHHHHHHH-HHHHX' for child A and 'XHHHHHHHHHHHHHLLLHHHHHHHH-HHHHLLLHHHHHHHHHHHHHX' for child B. A mixing coefficient of 0.25 was used to combine parent and child sequences together at each step. Multistate design pseudocode implementation is available in the supplements (Algorithm S4).

Partial noising

To promote further exploration of known active sequence subspaces, we implemented partial noising during design trajectories. This was done by introducing a temperature parameter, $\lambda \in (0 \text{ to } 1]$, to PG that, at the beginning of a trajectory, sets t to round (T*λ) rather than T (complete noising). This forced the model to do local exploration of the partially noisy sequence subpace. This approach was done in conjunction with normal design trajectories in 'multistate guidance' with λ parameters iterated through [0.2,0.3,0.5,0.8] starting with passing AF2 design sequences.

Iterative guidance

Iterative guidance was employed as described in Fig. 1c. In short, the experimental or in silico characterization of designs is collected to train a classifier, which is used to generate further designs with guidance. In an iterative process, classifiers trained on designs sampled in a preceding round inform network sampling. We investigated the nearly complete fitness landscape of the V39, D40, G41 and V54 amino acid sites of GB1, which is the binding domain of protein G, an immunoglobulin binding protein found in streptococcal bacteria, as provided in the FLIP[58] paper. Three rounds of iterative design were conducted with a batch size of 96.

We performed diffusion on the four mutation sites in the GB1 protein while providing guidance using a scale of 2.0 in the second-round and third-round design processes. In the first round, designs are directly sampled from our protein generator. For each round, a vanilla MLP (two layers, rectified linear unit (ReLU) activation and dropout of 0.25 after the first layer) is trained on the designs generated in the preceding round using a fitness equal to 1 as the classifier boundary. For comparison, batched Bayesian optimization was performed for three rounds using the sequences generated in the first round of the iterative guidance as the initial dataset. For Bayesian optimization, we used Gaussian processes with the radial basis function (RBF) kernel for the surrogate function, and we used the Monte Carlo–based batch upper confidence bound (qUCB) acquisition[59] function with $\lambda = 0.5$.

Sequence and structure quality metrics

All designs used in the ESMFold[21] benchmarks were modeled by using the following curl command: [curl -X POST –data 'SEQUENCE'] (https://api.esmatlas.com/foldSequence/v1/pdb/). All designs that were used in AF2 benchmarks and ordered for experimental characterization were predicted in single-sequence structure prediction mode with model 4. Pairwise backbone RMSDs between the design model and AF2 model were calculated for each design. Sequence quality metrics measured by ESM[21] (pseudo) perplexity were calculated with model: esm2_t33_650M_UR50D.

Sequence identity calculations

Blast alignment was used to examine sequence alignment and similarity using query coverage >90% and target coverage >50%. Alignment to natives was done against UniRef90 (Supplementary Fig. 27).

Plasmid construction

Protein designs were cloned into plasmids as in Watson et al.[11]. In brief, designs were ordered as synthetic genes (eBlocks, Integrated DNA Technologies) with BsaI overhangs compatible with a ccdB-encoding expression vector, LM0627 (ref. 7). Genes cloned into LM0627 result in the following sequence: MSG-design-GSGSHHWGST**HHHHHH** (SNAC cleavage tag and **6×His affinity tag** are indicated). We used the NEBridge Golden Gate Assembly Kit (New England Biolabs) with a total reaction volume of 5 µl and a ratio of 1:2 by mass of LM0627 plasmid DNA to design. We then incubated the reaction mixture at 37 C for 30 min, halted the reaction by incubating the reaction mixture at 60 °C for 5 min and transformed 1 µl of the reaction mixture into 6 µl of BL21 competent cells (New England Biolabs). After heat shock and recovery in SOC media, transformed BL21 cells were grown overnight in 1.0 ml of LB from which glycerol stock was created and small-scale expression cultures were inoculated.

1-ml-scale protein purification

Initially, proteins were expressed with small-scale expression screens as previously reported[7] with small adaptations. In brief, designs were inoculated with 100 µl of overnight growths and 900 µl of auto-induction media (sterile-filtered TBII media supplemented with 50 µg ml^{-1} kanamycin, 2 mM $MgSO_4$, 1 × 5,052) in deep-well 96-well plates. Sixteen hours after inoculation, cells were harvested and lysed in lysis buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM imidazole supplemented with 1× BugBuster, 1 mM PMSF, 0.1 mg ml^{-1} lysozyme, 0.1 mg ml^{-1} DNase). Clarified lysates were added to a 50-µl bed of Ni-NTA agarose resin in a 96-well fritted plate equilibrated with wash buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM imidazole). After sample application and flowthrough, the resin was washed three times with wash buffer, and samples were eluted in 200 µl of elution buffer (50 mM Tris-HCl (pH 8), 0.3 M NaCl, 0.5 M imidazole, 5 mM EDTA (pH 8)). All eluates were sterile filtered with a 96-well 0.22-µm filter plate (Agilent, 203940-100) before SEC. Protein designs were then screened via SEC using an ÄKTA FPLC outfitted with an autosampler capable of running samples from a 96-well source plate. Samples were run on a Superdex S75 Increase 5/150 GL column (Cytiva, 29148722; 3,000–70,000-Da separation range) in a running buffer (20 mM Tris (pH 8), 150 mM NaCl). To improve peak resolution, the SEC column was connected directly in line from the autosampler to the UV detector. Then, 0.25-ml fractions were collected from each run. Absorption spectra were collected by an ÄKTA U9-M at 230 nm and 280 nm.

50-ml-scale protein purification

Proteins selected for further downstream characterization were expressed in 50 ml of auto-induction media[60]. Sixteen hours after inoculation, cells were harvested and lysed in lysis buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM imidazole, 1 mM PMSF, 0.1 mg ml^{-1} lysozyme, 0.1 mg ml^{-1} DNase) through sonication. Clarified lysates were added to a 2-ml bed of Ni-NTA agarose resin in a 20-ml column (Bio-Rad, 7321010) equilibrated with wash buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM imidazole). After sample application and flowthrough, the resin was washed three times with 10 ml of wash buffer, and samples were eluted in 2 ml of elution buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 200 mM imidazole). All eluates were sterile filtered with a 3-ml 0.22-µm filter plate before SEC. Protein designs were then screened via SEC using an ÄKTA FPLC outfitted with an autosampler capable of running samples from a 96-well source plate. Samples were run on a Superdex S75 Increase 10/300 GL column (Cytiva, 29148721; 3,000–70,000-Da separation range) in a running buffer (20 mM Tris (pH 8), 150 mM NaCl). Then, 1-ml fractions were collected from each run. Absorption spectra were collected by the ÄKTA U9-M at 230 nm and 280 nm.

0.5-L-scale protein purification and SNAC cleavage

The best expressing proteins used in high-resolution structural studies were selected for further scale-up and SNAC cleavage[61]. Proteins were expressed in Studiers M2 autoinduction media with 50 µg ml^{-1} kanamycin. Pre-cultures were grown overnight. Cultures were inoculated with 10 ml of pre-culture and grown at 37 °C for 4 h before lowering temperature to 22 °C for 14 h, and cultures were inoculated with 10 ml of pre-culture. Cells were pelleted at 4,000*g* for 10 min, after which the supernatant was discarded. Pellets were resuspended in 30 ml of lysis buffer (100 mM Tris HC (pH 8), 100 mM NaCl, 400 mM imidazole, 1 mM PMSF, 1 mM DNase). Cell suspensions were lysed by sonication for 7.5 min (10 s on, 10 s off) at 80% amplitude using a Qsonica four-prong sonicator. The lysate was clarified at 14,000*g* for 30 min. The His-tagged proteins were batch bound for 1 h to 8 ml of Ni-NTA resin (Qiagen) and washed with 10 ml of lysis buffer and 30 ml of high-salt wash buffer (25 mM Tris HCl (pH 8), 1 M NaCl, 40 mM imidazole) and then 10 ml of SNAC cleavage buffer (100 mM CHES, 100 mM acetone oxime, 100 mM NaCl, 500 mM GnCl (pH 8.6)). Next, 40 ml of SNAC cleavage buffer and 80 µl of 1 M $NiCl_2$ were added, and columns were closed and shaken on a nutator for 12 h to cleave. After cleavage, the flowthrough was collected and concentrated before further purification by SEC/FPLC as described above.

Cysteine bias protein expression

Proteins guided toward high cysteine content were transformed into and expressed in Rosetta-gami B(DE3) competent cells (Novagen, 71137). The 1-ml and 50-ml scale protein purification protocols were otherwise followed.

CD

CD spectra were collected on a Jasco J-1500 CD spectrometer with 1-nm bandwidth, 50-nm permanent scan rate and data integration time of 4 s per read. Sample cuvettes stored in 2% Hellmanex (Hellma, 9-307-011-4-507) were washed with deionized water, 2% Hellmanex, deionized water and then 20% ethanol, after which 300 µl of SEC-purified protein was added for CD spectra measurements. Thermal melts were performed in 10° intervals between 25 °C and 95 °C.

Mass spectrometry

To identify the molecular mass of each protein, intact mass spectra were obtained via reverse-phase liquid chromatography–mass spectrometry (LC–MS) on an Agilent G6230B TOF on an AdvanceBio RP-Desalting column and subsequently deconvoluted by way of BioConfirm using a total entropy algorithm. Disulfide formation was determined by injecting protein at 1.5 mg ml^{-1} in the presence and absence of 50 mM TCEP-HCl (Millipore Sigma, 646547-10X1ML) and detecting the mass shift.

Disulfide bond quantification

To measure the number of cysteines via alkylation, proteins at 1.5 mg ml^{-1} in SEC running buffer (20 mM Tris (pH 8), 150 mM NaCl) were incubated in 50 mM TCEP-HCl at 50 °C for 1 h to reduce disulfide bonds. Simultaneously, an equal amount of protein in SEC running buffer was heated to 50 °C without 10 mM TCEP to maintain formed disulfides. Iodoacetamide (Millipore Sigma, I1149) was added to both conditions to a final concentration of 10 mM and incubated away from light at room temperature for 30 min to alkylate unpaired cysteines. To identify the molecular mass and alkylations status of each protein, intact mass spectra were obtained via reverse-phase LC–MS on the Agilent G6230B TOF on an AdvanceBio RP-Desalting column and subsequently deconvoluted by way of BioConfirm using a total entropy algorithm.

Barcode extraction and liquid chromatography–tandem mass spectrometry

Ni-NTA eluate of the 84-design pool was subjected to SEC with deep fractionation (0.25-ml fractions). From every other fraction, 100 µl was added to fresh wells in a 96-well plate, and fractions were subjected to cleavage in 100 µl of Lys-C buffer (8 M urea, 100 mM Tris HCl (pH 8)) plus 1 µg of endoproteinase LysC (New England Biolabs, P8109S), as previously described. After hexaHis-tagged barcode pulldown with magnetic His-pulldown beads (Thermo Fisher Scientific, 10103D) and subsequent trypsin (New England Biolabs, P8101S) digest to free barcodes, barcodes were diluted 50% in 0.1% trifluoroacetic acid (TFA). Barcode pools corresponding to SEC fractions were separated by hydrophobicity using a previously described tandem guard column–analytical column setup. The guard column was packed to 2 cm with 5 µm of silica (ReproSil-Pur 120 C18Aq, ESI Source Solutions, r15.aq.0001), whereas the analytical column was packed to 14 cm with 1.9 µm of silica (ReproSil-Pur 120 C18Aq, ESI Source Solutions, r119.aq.0001). Peptides were detected using a previously described data independent acquisition (DIA) protocol on a Orbitrap Fusion Lumos Tribrid (Thermo Fisher Scientific) at the UW Proteomic Resource (UWPR).

Solution NMR

Recombinant plasmid DNA (~100 ng) containing synthetic genes encoding for child A, child B and parent for several design families were separately transformed in *E. coli* BL21(DE3) cells. Colonies were grown under kanamycin selection on LB agar media for 16 h. Toward preparation of uniformly ^{15}N-labeled proteins, a streak of colonies was resuspended in 60 ml of 1× M9 minimal media[62] and grown overnight at 37 °C/225 r.p.m., and the inoculum was used to initiate a 1-L 1× minimal media culture supplemented with kanamycin and ^{15}N ammonium chloride (Cambridge Isotope Laboratories, NLM-467) as the nitrogen source. For ^{15}N/^{13}C-labeled proteins, ^{13}C-labeled glucose (Cambridge Isotope Laboratories, CLM-1396) was used as the carbon source. Cultures were incubated at 37 °C/225 r.p.m. until the optical density at 600 nm (OD_{600}) reached 0.6 and then induced with 1 mM IPTG and grown at 37 °C/225 r.p.m. for 6 h. Cultures were harvested by centrifugation (6,000*g*, 15 min, 4 °C), and cell pellets were resuspended with wash buffer (300 mM NaCl, 10 mM imidazole, 50 mM Tris (pH 8)). Cells were lysed by sonication on ice. The lysate was clarified by centrifugation at 10,000*g* for 20 min at 4 °C. The supernatant was loaded onto a 5-ml His-Trap Ni-NTA column. The column was washed extensively with wash buffer, and protein was eluted using a linear gradient from 0% to 100% elution buffer (300 mM NaCl, 500 mM imidazole, 50 mM Tris (pH 8)). Fractions containing protein were pooled and further purified by SEC on a Superdex 200 Increase 10/300 GL column in NMR buffer (100 mM NaCl, 20 mM sodium phosphate (pH 6.5)). All designs were purified into batch-matched NMR buffer and then concentrated to 300 µl in 3-kDa Amicon concentrators. The purity of eluent fractions was confirmed to be greater than 95% by SDS-PAGE. Protein concentrations were

measured by NanoDrop spectrophotometer at 280 nm with extinction coefficient predicted by ExPASY ProtParam. 2D ^{1}H-^{15}N amide HSQC spectra (Bruker pulse sequence hsqcetf3gpsi) were acquired using standard parameters at a ^{1}H field of 800 MHz at 37 °C with recycle delay (d1) set to 1.2 s, sweep width of 30 ppm and acquisition time of 60 ms and number of scans ranging from 8 to 64 on a Bruker AVIIIHD-800 spectrometer equipped with a 3-mm TCI cryoprobe. All data were processed in NMRPipe[63] and analyzed in NMRFAM-SPARKY[64].

Uniformly double-labeled $^{15}N/^{13}C$ design proteins were prepared in NMR buffer as described above at final concentrations of 200 µM to 1,400 µM. Backbone HN, N, Cα, C13 and CO resonances were assigned using sequential assignment strategies[65] via standard triple-resonance experiments with non-uniform sampling with 20% Poisson gap sampling schedule and were reconstructed with istHMS10 (http://gwagner.med.harvard.edu/intranet/hmsIST/). The following experiments were recorded: 3D HNCA (Bruker pulse sequence hncagpwg3d), 3D HNCO (Bruker pulse sequence hncogpwg3d), 3D HNCACB (Bruker pulse sequence hncacbgpwg3d) and 3D CBCACONH (Bruker pulse sequence cbcaconhgpwg3d). Acquisition times were 92 ms in ^{1}H, 15 ms in ^{15}N, 20 ms in ^{13}CO and 10/5 ms in $^{13}Cα$/C13. Recycle delay was set to 1 s in all experiments, which were recorded at a ^{1}H field of 800 MHz at 37 °C. To obtain through-space restraints for structure calculation, 3D amide–amide NOESY experiments (3D SOFAST $_{HNHAro\text{-}NHN}$) were collected with 8–16 scans, 0.6-s recycle delay and 350-ms mixing time[66]. Nuclear Overhauser effect (NOE) cross-peaks were assigned manually in NMRFAM-SPARKY. NMR peak assignments were used by TALOS-N[44] to determine secondary structure information, random coil index order parameter predictions and dihedral angle restraints toward structure calculation. Structure calculations were set up with automated Python scripts using CS-Rosetta[67,68]. We first used TALOS-N to determine psi and phi dihedral angles, and we used protein design sequences and assigned chemical shift values to pick fragments of amino acid lengths 3 and 9. We then used the protein sequence, 3mer/9mer fragments, backbone chemical shifts and amide–amide NOEs as input for the abrelax CS-Rosetta protocol (Rosetta version 3.8 and CS-Rosetta Toolbox version 3.3). From the 30,000 decoys calculated, the 10 lowest energy models were selected to represent the final NMR ensemble structure. The structure calculation was considered converged because the lowest energy models clustered within less than 2 Å from the model with the lowest energy. Final structures were validated with MolProbity.

$^{1}H_N$ and ^{15}N ACS values were determined from peaks in 2D ^{1}H-^{15}N HSQC spectra using the following equation[43,69]:

$$\mathrm{ACS_i} = \frac{1}{N}\sum_{K=1,M}\omega_k$$

where i = $^{1}H_N$ and ^{15}N atoms; N is the total number of peaks picked in the HSQC spectrum; M is the total number of residues in the protein sequence; and wk is the chemical shift of the k-th resonance.

Reference $^{1}H_N$ and ^{15}N ACS values for primarily α-helical proteins and primary 13-sheet proteins were taken from previous reports by Mielke et al.[43,69].

Furin cleavage

To cleave designed proteins, 5 U of furin protease (New England Biolabs, P8077S) was combined with 30 µM design in enzyme buffer (20 mM HEPES, 1 mM $CaCl_2$, 0.2 mM 13-mercaptoethanol) and incubated at 25 °C for 16 h. Cleavage reaction was used for SDS-PAGE (Any kDTM Mini-PROTEAN TGXTM Precast Protein Gels) with protein standards (Precision Plus Protein Dual Color Standards).

Blood cell lysis assay

Hemolysis assay was performed as described previously.[70] Single-donor washed human RBCs (Innovative Research, IWB3ALS40ML) were washed three times by spinning blood at 500*g* for 5 min and discarding supernatant until supernatant appears clear. PBS was used to resuspend the RBCs at 10% hematocrit (v/v). Blood cell lysis was carried out in a 96-well plate at a final hematocrit of 2.5%. Negative controls include PBS, cleavage buffer and cleavage buffer with 0.5 U of furin. Positive controls include 2% Triton-X-100 (Sigma-Aldrich, 9036-19-5) and 15 µM melittin (GenScript, RP20415). Designed proteins were diluted to 15 µM with PBS. Washed RBCs were added to each well and incubated at 37 °C for 1 h, after which the reaction plate was spun down at 500*g* for 5 min. Supernatant from the reaction plate was transferred to a 96-well clear-bottom microplate (Corning, 3598). Absorbance was measured at 450 nm on an Agilent BioTek Epoch 2 TSC microplate reader.

Crystallography

All crystallization experiments were conducted using the sitting drop vapor diffusion method. Crystallization trials were set up in 200-nl drops using the 96-well plate format at 20 °C.

Crystallization plates were set up using a mosquito LCP from SPT Labtech and then imaged using UVEX microscopes and UVEX PS-256 from JAN Scientific. Diffraction quality crystals formed in a mixture of 0.1 M PCB buffer (pH 4) and 25% PEG 1500.

Diffraction data were collected at the National Synchrotron Light Source II. X-ray intensities and data reduction were evaluated and integrated using XDS[71] and merged/scaled using Pointless/Aimless in the CCP4 program suite[72]. Structure determination and refinement starting phases were obtained by molecular replacement using Phaser[73] using the designed model for the structures. After molecular replacement, the models were improved using phenix.autobuild[74]. Structures were refined in Phenix[74]. Model building was performed using Coot[75]. The final model was evaluated using MolProbity[76]. Data collection and refinement statistics are recorded in Table 1. Data deposition, atomic coordinates and structure factors reported in this paper have been deposited in the PDB with accession code 8VD6.

Reporting summary

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

Data availability

Atomic coordinates have been deposited in the Protein Data Bank (http://www.rcsb.org/) with accession codes 8VD6 (ref. 77) (Crystal structure of CAP repeat), 8VL4 (ref. 78) (NMR structure of MS3 parent) and 8VL3 (ref. 79) (NMR structure of MS2 parent). The backbone chemical shift assignments of design proteins have been deposited to the Biological Magnetic Resonance Bank with accession codes 31137 (MS3 parent) and 31136 (MS2 parent).

Code availability

The code[80] for this project, with the exception of training scripts, is available at https://github.com/RosettaCommons/protein_generator. For greater accessibility, we thank Simon Dürr and HuggingFace who supplied a GPU grant to run the model interactively in their browser: https://huggingface.co/spaces/merle/PROTEIN_GENERATOR.

References

53. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
54. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
55. Voynov, V., Chennamsetty, N., Kayser, V., Helk, B. & Trout, B. L. Predictive tools for stabilization of therapeutic proteins. *mAbs* **1**, 580–582 (2009).
56. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

57. Labesse, G., Colloc'h, N., Pothier, J. & Mornon, J. P. P-SEA: a new efficient assignment of secondary structure from Cα trace of proteins. *Comput. Appl. Biosci.* **13**, 291–295 (1997).
58. Dallago, C. et al. FLIP: benchmark tasks in fitness landscape inference for proteins. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.09.467890 (2022).
59. Balandat, M. et al. BoTorch: a framework for efficient Monte-Carlo Bayesian optimization. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1910.06403 (2020).
60. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).
61. Dang, B. et al. SNAC-tag for sequence-specific chemical protein cleavage. *Nat. Methods* **16**, 319–322 (2019).
62. Azatian, S. B., Kaur, N. & Latham, M. P. Increasing the buffering capacity of minimal media leads to higher protein yield. *J. Biomol. NMR* **73**, 11–17 (2019).
63. Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
64. Lee, W., Tonelli, M. & Markley, J. L. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31**, 1325–1327 (2015).
65. Frueh, D. P. Practical aspects of NMR signal assignment in larger and challenging proteins. *Prog. Nucl. Magn. Reson. Spectrosc.* **78**, 47–75 (2014).
66. Rossi, P., Xia, Y., Khanra, N., Veglia, G. & Kalodimos, C. G. ^{15}N and ^{13}C-SOFAST-HMQC editing enhances 3D-NOESY sensitivity in highly deuterated, selectively [^{1}H,^{13}C]-labeled proteins. *J. Biomol. NMR* **66**, 259–271 (2016).
67. Shen, Y. et al. Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl Acad. Sci. U.S.A.* **105**, 4685–4690 (2008).
68. Nerli, S. & Sgourakis, N. G. CS-ROSETTA. *Methods Enzymol.* **614**, 321–362 (2019).
69. Mielke, S. P. & Krishnan, V. V. Protein structural class identification directly from NMR spectra using averaged chemical shifts. *Bioinformatics* **19**, 2054–2064 (2003).
70. Evans, B. C. et al. Ex vivo red blood cell hemolysis assay for the evaluation of pH-responsive endosomolytic agents for cytosolic delivery of biomacromolecular drugs. *J. Vis. Exp.* **73**, e50166 (2013).
71. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
72. Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).
73. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
74. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
75. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
76. Williams, C. J. et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci. Publ. Protein Soc.* **27**, 293–315 (2018).
77. Bera, A. K. Crystal structure of de novo design repeat protein C12. https://www.rcsb.org/structure/8vd6 (2024).
78. McShan, A. C. & Simma, M. K. Solution NMR structure of de novo design protein 312 parent. https://www.rcsb.org/structure/8VL4 (2024).
79. McShan, A. C. & Simma, M. K. Solution NMR structure of de novo design protein F3 parent. https://www.rcsb.org/structure/8VL3 (2024).
80. Lisanza, S., Gershon, J. M., Tipps, S. & Arnoldt, L. ProteinGenerator. https://github.com/RosettaCommons/protein_generator (2023).

Acknowledgements

We would like to acknowledge L. Li, S. Mansoor, D. Zorine, D. Chmielewski, D. Feldman, I. Humphreys, H. Pyles, B. Trippe, D. See, A. Idris, X. Han, M. Said, F. Dou, L. Ann, K. Wu, D. Hicks, E. Kinfu, A. Chazin-Gray, C. Pan and N. Jasti for helpful discussions and support; N. Ennist for help with CD; I. Anishchenko for scripts to run TM—align, sequence similarity and multidimensional scaling plots; J. Wang, J. Dauparas, D. Tischer, J. Watson, D. Juergens and N. Bennett for scripts for benchmarking scripts; M. Baek and F. DiMaio for training scripts and RoseTTAFold code base; and N. King, I. Haydon, L. Stewart, L. Goldschmidt, K. Van Wormer and L. Carter for general operations. This work was supported by Defense Threat Reduction Agency grant HDTRA1-19-1-0003 (X.L.); by funding from the DARPA program Harnessing Enzymatic Activity for Lifesaving Remedies (HEALR) under award HR0011-21-2-0012 (X.L.); by Juvenile Diabetes Research Foundation International grant 2-SRA-2018-605-Q-R (X.L.); by Amgen (S.L.); by Helmsley Charitable Trust Type 1 Diabetes (T1D) program grant 2019PG-T1D026 (X.L.); by Bill and Melinda Gates Foundation grant OPP1156262 (X.L.); by the Audacious Project at the Institute for Protein Design (J.S.); and by the Howard Hughes Medical Institute (J.S.). A.C.M. acknowledges startup funds from the Georgia Institute of Technology. Crystallographic data were collected at the National Synchrotron Light Source II (NSLS2) beamline. The Center for Bio-Molecular Structure is primarily supported by the National Instututes of Health/National Institute of General Medical Sciences through a Center Core P30 Grant (P30GM133893) and by the US Department of Energy (DOE) Office of Biological and Environmental Research (KP1607011). The NSLS2 is a DOE Office of Science User Facility operated under contract DE-SC0012704. This publication resulted from data collected using the beamtime obtained through NECAT BAG proposal number 311950.

Author contributions

Conceptualized the study: S.L., J.G., S.T., J.S., L.A. and D.B. Designed the experiments: S.L., J.G., S.T., J.S., L.A., A.M. and D.B. Conducted the experiments: S.L., J.G., S.T., J.S., L.A., S.H., M.S., G.L., M.Y., H.W., C.T., X.L., A.K., E.B., A.B., S.G. and A.M. Wrote the initial draft of the manuscript: S.L., J.G., S.T., J.S., L.A., S.H., A.M. and D.B. Analyzed the data: S.L., J.G., S.T., J.S., L.A., S.H., A.B., A.M. and D.B. Approved the final manuscript: all authors

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-024-02395-w.

Correspondence and requests for materials should be addressed to David Baker.

Peer review information *Nature Biotechnology* thanks Christian Dallago, Kevin Wu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): David Baker

Last updated by author(s): Aug 16, 2024

Reporting Summary

Nature Portfolio 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 Portfolio 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	ProteinGenerator (this study: https://github.com/RosettaCommons/protein_generator), AlphaFold2, TMalign, Protein-Protein BLAST 2.11.0+, ESMFold, RoseTTAFold Hallucination, AlphaFold2 Hallucination, RFdiffusion 1.0.0, ProteinMPNN 1.0, CCP4 8.0.0, NMRFAM-SPARKY 3.19, nmrPipe 11.5, EvoDiff, python3.
Data analysis	Matplotlib 3.6.2, SciPy 1.9.3, Seaborn 0.11.2, PyMOL 2.5.0, pycorn 0.19, UCSF ChimeraX 1.6.1, BeStSel, GraphPad Prism 10.2.0, Phaser 2.1.2, Phenix 1.21, Coot 1.1 MolProbity 4.5.2, CS-Rosetta toolkit 3.3, TopSpin 3.2, PyTorch 1.9.1.

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Crystal data for the repeat protein in Figure 3C has been deposited in the PDB under accession code 8VD6, and NMR data for MS2 and MS3 in Figure 5 have been deposited in the PDB under accession codes 8VL3 and 8VL4 respectively.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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

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	Variable sample size depending on analysis, detailed in figure legends. But in general were chosen dependent on experimental constraints. (e.g. max 96 designs to fit on a plate)
Data exclusions	None
Replication	Each data set contains many independent measurements as reported in figure legends.
Randomization	Outcomes in experiments are not susceptible to randomization bias.
Blinding	Outcomes in experiments in the manuscript are not susceptible to bias of knowing conditions.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.