eLife

Reviewed Preprint
v1 • August 24, 2026
Not revised

For correspondence:
ava.amini@microsoft.com
yang.kevin@microsoft.com

† Work done principally during an internship at Microsoft Research

* These authors jointly supervised the work

Competing interests: No competing interests declared

Reviewing editor: Bin Zhang, Massachusetts Institute of Technology, United States



# Protein generation with evolutionary diffusion: sequence is all you need

**Sarah Alamdari[1], Nitya Thakkar[2,†], Rianne van den Berg[3], Neil Tenenholtz[1], Robert Strome[4], Alan M Moses[4], Alex X Lu[1], Nicolò Fusi[1], Ava P Amini[1,*], Kevin K Yang[1,*]**

[1]Microsoft Research, Cambridge, United States • [2]Stanford University, Palo Alto, United States • [3]Microsoft Research AI for Science, Amsterdam, Netherlands • [4]Department of Cell and Systems Biology, University of Toronto, Toronto, Canada

## eLife Assessment

This **important** study introduces EvoDiff, an order-agnostic diffusion foundation model for protein sequence generation, and demonstrates its potential across unconditional generation, motif scaffolding, MSA-conditioned design, and IDR inpainting. The evidence supporting the principal claims is **solid**, combining computational evaluation with experimental validation, although the comparative advantages and generality of some applications would benefit from further clarification and calibration.

https://doi.org/10.7554/eLife.112029.1.sa2

## Abstract

Deep generative models are increasingly powerful tools for the *in silico* design of novel proteins. Recently, a family of generative models called diffusion models has demonstrated the ability to generate biologically plausible proteins that are dissimilar to any actual proteins seen in nature, enabling unprecedented capability and control in *de novo* protein design. However, current state-of-the-art diffusion models generate protein structures, which limits the scope of their training data and restricts generations to a small and biased subset of protein design space. Here, we introduce a general-purpose diffusion framework, EvoDiff, that combines evolutionary-scale data with the distinct conditioning capabilities of diffusion models for controllable protein generation in sequence space. EvoDiff generates high-fidelity, diverse, and structurally-plausible proteins that cover natural sequence and functional space. We show experimentally that EvoDiff generations express, fold, and exhibit expected secondary structure elements. Critically, EvoDiff can generate proteins inaccessible to structure-based models, such as those with disordered regions, while maintaining the ability to design scaffolds for functional structural motifs. We validate the universality of our sequence-based formulation by experimentally characterizing intrinsically-disordered mitochondrial targeting signals, metal-binding proteins, and protein binders designed using EvoDiff. We envision that EvoDiff will expand capabilities in protein engineering beyond the structure-function paradigm toward programmable, sequence-first design.

Evolution has yielded a diversity of functional proteins that precisely modulate cellular processes. Recent years have seen the emergence of deep generative models that aim to learn from this diversity to generate proteins that are both valid and novel, with the ultimate goal of then tailoring function to solve outstanding modern-day challenges, such as the rapid development of targeted therapeutics and vaccines or engineered enzymes for the degradation of industrial waste (Fig. 1A) (1, 2). Diffusion models provide a particularly powerful framework for generative modeling of novel proteins, as they generate high-diversity samples and can be conditioned given a wide variety of inputs or design objectives (3–6). Indeed, today's most biologically-plausible instances of *in silico*-designed proteins come from diffusion models of protein *structure* (7–15).

These models – including the current state-of-the-art approach RFdiffusion (10) – fit in the structure-based protein design paradigm of first generating a structure that fulfills desired constraints and then designing a sequence that will fold to that structure. However, sequence, not

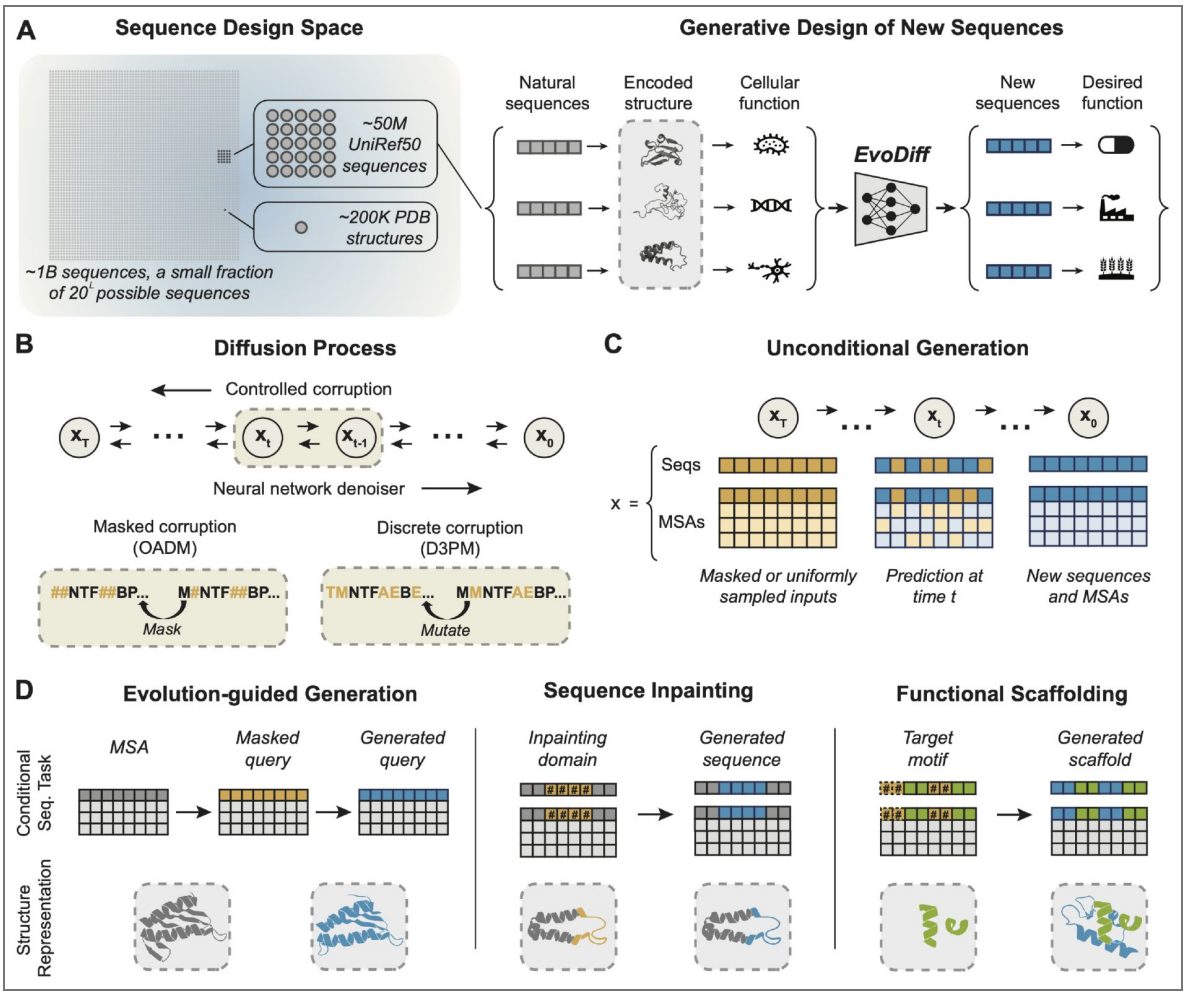



**Figure 1. Protein sequence generation with evolutionary diffusion.**

**(A)** (Left) Evolution has sampled a tiny fraction of possible protein sequences. Experimental structures have been determined for even fewer proteins. (Right) EvoDiff is a generative discrete diffusion model trained on natural protein sequences. Sampling from EvoDiff yields new protein sequences that may perform desired functions. **(B)** Discrete diffusion models consist of controlled corruption and learned denoising processes. In the masked corruption process, input tokens are masked in an order-agnostic fashion (bottom, left). In the discrete corruption process, inputs are corrupted via a Markov process controlled by a transition matrix capturing amino acid mutation frequencies (bottom, right). **(C)** EvoDiff enables unconditional generation of protein sequences or MSAs. Starting from masked or uniformly sampled inputs $x_T$, EvoDiff generates new sequences or MSAs by reversing the corruption process, iteratively denoising $x_t$ into realistic sequences or MSAs $x_0$. **(D)** Controllable protein design with EvoDiff, via conditioning on evolutionary information encoded in MSAs (left); inpainting functional domains from masked portions of a sequence (middle); or scaffolding structural motifs without explicit structural information (right). thermore, EvoDiff-OADM is the only model variant where performance scales with increased model size (Table S1; Fig. S1).

structure, is the universal design space for proteins. Every protein is completely defined by its amino-acid sequence. We discover proteins by finding their coding sequences in genomes, and proteins are synthesized as amino-acid sequences. Sequence then determines function through both an ensemble of structural conformations and the chemistry enabled by the amino acids themselves. However, not every protein folds into a static structure. In these cases, structure-based design is not viable because the function is not mediated by a static structure (16–18), with the most extreme examples being intrinsically disordered regions (IDRs) (19). Therefore, static structures characterized by X-ray crystallography are an incomplete distillation of the information captured in sequence space (20–23). Furthermore, structural data (ca. 200k solved structures in PDB) is scarce and unrepresentative of the full diversity of natural sequences (ca. billions of unique natural protein sequences; Fig. 1A), inherently limiting the capacity of any structure-based generative model to learn the full diversity of protein functional space.

We combine evolutionary-scale datasets with diffusion models to develop a powerful new generative modeling framework, which we term EvoDiff, for controllable protein design from sequence data alone (Fig. 1). Given the natural framing of proteins as sequences of discrete tokens over an amino acid language, we use a *discrete* diffusion framework in which a forward process iteratively corrupts a protein sequence by changing its amino acid identities, and a learned reverse process, parameterized by a neural network, predicts the changes made at each iteration (Fig. 1B). The reverse process can then be used to generate new protein sequences starting from random noise (Fig. 1C). Importantly, EvoDiff's discrete diffusion formulation is mathematically distinct from continuous diffusion formulations previously used for protein structure design (7–15). Beyond evolutionary-scale datasets of single protein sequences, multiple sequence alignments (MSAs) inherently capture evolutionary relationships by revealing patterns of conservation and variation in the amino acid sequences of sets of related proteins. We thus additionally build discrete diffusion models trained on MSAs to leverage this additional layer of evolutionary information to generate new single sequences (Fig. 1C-D).

We evaluate our sequence and MSA models – EvoDiff-Seq and EvoDiff-MSA, respectively – across a range of generation tasks to demonstrate their power for controllable protein design (Fig. 1D). We first show that EvoDiff-Seq unconditionally generates high-quality, diverse proteins that capture the natural distribution of protein sequence, structural, and functional space. Unconditional sampling from EvoDiff can yield proteins that stably express and exhibit expected biophysical properties. Using EvoDiff-MSA, we achieve evolution-guided design of novel sequences conditioned on an alignment of evolutionarily-related, but distinct, proteins. Finally, by exploiting the conditioning capabilities of our diffusion-based modeling framework and its grounding in a universal design space, we demonstrate – through both *in silico* and wetlab experiments – that EvoDiff can reliably generate proteins with functional IDRs, directly overcoming a key limitation of structure-based generative models, and generate scaffolds for functional structural motifs without any explicit structural or binder information.

## Discrete diffusion models of protein sequence

EvoDiff is the first generative diffusion model for protein design trained on evolutionary-scale protein sequence data. We investigated two types of forward processes for diffusion over discrete data modalities (24, 25) to determine which would be most effective (Fig. 1B). In order-agnostic autoregressive diffusion (EvoDiff-OADM, see Methods) (24), one amino acid is converted to a special mask token at each step in the forward process (Fig. 1B). After $T = L$ steps, where $L$ is the length of the sequence, the entire sequence is masked. We additionally designed discrete denoising diffusion probabilistic models (EvoDiff-D3PM, see Methods) (25) for protein sequences. In EvoDiff-D3PM, the forward process corrupts sequences by sampling mutations according to a transition matrix, such that after $T$ steps the sequence is indistinguishable from a uniform sample over the amino acids (Fig. 1B). In the reverse process for both, a neural network model is trained to undo the previous corruption. The trained model can then generate new sequences starting from sequences of masked tokens or of uniformly-sampled amino acids for EvoDiff-OADM or EvoDiff-D3PM, respectively (Fig. 1C).

To facilitate direct and quantitative model comparisons, we trained all EvoDiff sequence models on 42M sequences from UniRef50 (26) using a dilated convolutional neural network architecture introduced in the CARP protein masked language model (27). We trained 38M-parameter and 640M-parameter versions for each forward corruption scheme to test the effect of model size on model performance. As a first evaluation of our EvoDiff sequence models, we calculated each model's test-set perplexity, which reflects its ability to capture the distribution of natural sequences and generalize to unseen sequences (see Methods). We observe that EvoDiff-OADM learns to reconstruct the test set more accurately than two tested EvoDiff-D3PM variants employing uniform and BLOSUM62-based transition matrices (Table S1; Fig. S1). Fur-

To explicitly leverage evolutionary information, we designed and trained EvoDiff MSA models using the MSA Transformer (28) architecture on the OpenFold dataset (29). To do so, we subsampled MSAs to a length of 512 residues per sequence and a depth of 64 sequences, either by randomly sampling the sequences ("Random") or by greedily maximizing for sequence diversity ("Max"). Within each subsampling strategy, we then trained EvoDiff MSA models with the OADM and D3PM corruption schemes. OADM corruption results in the lowest validation set perplexities, indicating that OADM models are best able to generalize to new MSAs (Table S2; Fig. S2). To select a subsampling method, we compared the ability of each model to reconstruct validation set MSAs, finding that maximizing for sequence diversity yields improved performance no matter how the validation MSAs are subsampled (Table S2). We thus selected the OADM-Max model for downstream analysis, hereafter referring to it as EvoDiffMSA.

## Structural plausibility of generated sequences

We next investigated whether EvoDiff could generate new protein sequences that were individually valid and structurally plausible. To assess this, we developed a workflow that evaluates the foldability and self-consistency of sequences generated by EvoDiff (Fig. 2A). We generated 1000 sequences from each EvoDiff sequence model with lengths drawn from the empirical distribution of lengths in the training set. We compared EvoDiff's generations to sequences generated from a left-to-right autoregressive language model (LRAR) with the same architecture and training set as EvoDiff and to sequences generated from protein masked language models such as ESM-2 (30) (Figs. 2B-C, S3, S4; Table S3).

We assessed the foldability of individual sequences by predicting their corresponding structures using OmegaFold (31) and computing the average predicted local distance difference test (pLDDT) across the whole structure (Fig. 2B). pLDDT reflects OmegaFold's confidence in its structure prediction for each residue. In addition to the average pLDDT across a whole protein, we observe that pLDDT scores can vary significantly across a protein sequence (Fig. S5 B). It is important to note that while pLDDT scores above 70 are often considered to indicate high prediction confidence, low pLDDT scores can be consistent with intrinsically disordered regions (IDRs) of proteins (32), which are found in many natural proteins. As an additional metric of structural plausibility, we computed a self-consistency perplexity (scPerplexity) by redesigning each predicted structure with the inverse folding algorithm ESM-IF (33) and computing the perplexity against the original generated sequence (Fig. 2A, C; Table S3). Given that ESM-IF and EvoDiff were both trained on UniRef50 data, it is possible that sequences from EvoDiff's validation set overlap with sequences in the ESM-IF train set; thus we performed the same self-consistency evaluations using ProteinMPNN (34), which is not trained on UniRef50, for inverse folding (Table S3).

While no generative model approaches the test set values for foldability and self-consistency, EvoDiff-OADM outperforms EvoDiff-D3PM and improves when increasing the model size (Fig. 2B-D; Table S3). We therefore selected the 640M-parameter EvoDiff-OADM model for downstream analysis and hereafter refer to it as EvoDiff-Seq. While a left-to-right autoregressive (LRAR) protein language model generates slightly more structurally-plausible sequences (Table S3), EvoDiff-Seq offers the advantage of direct, flexible conditional generation due to its order-agnostic decoding. Unconditional generation from masked language models produces less structurally-plausible sequences because of the mismatch between the training and generation

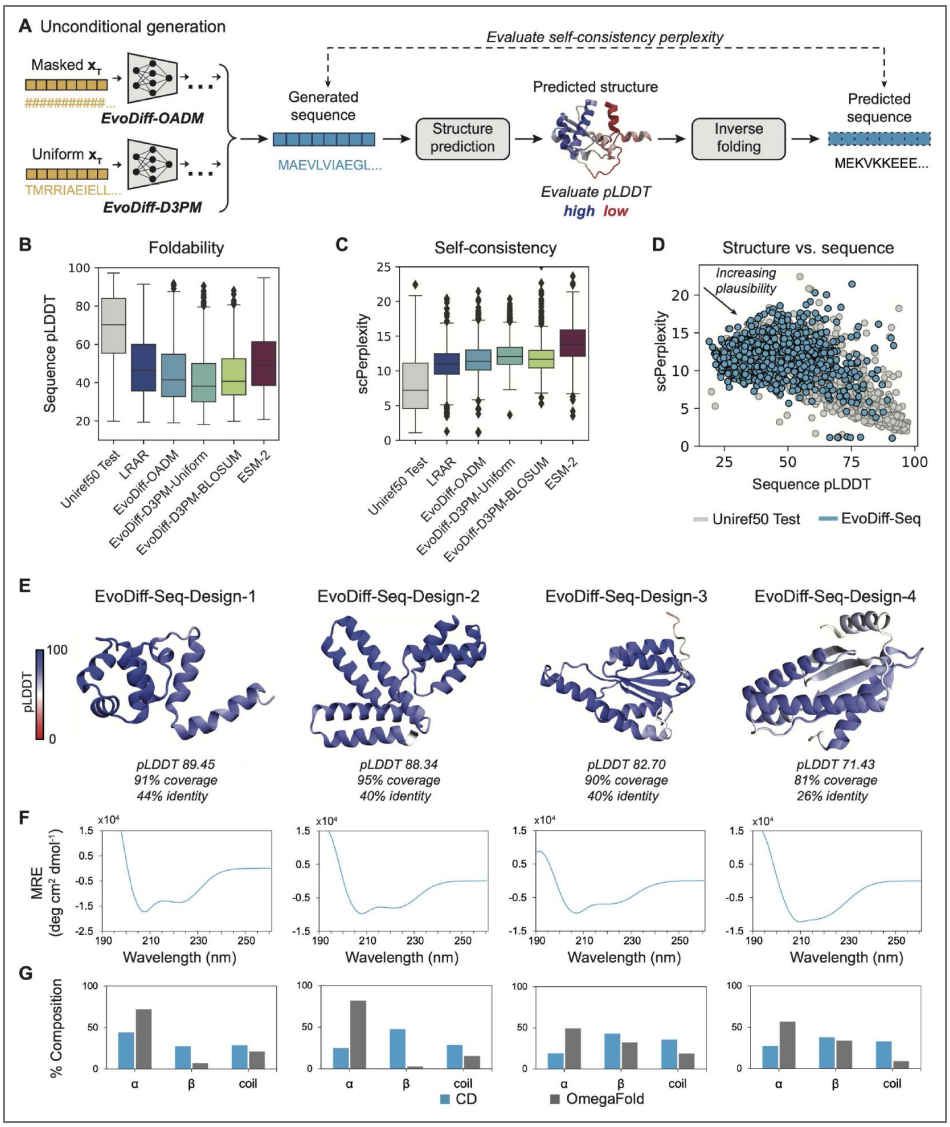



**Figure 2.** **EvoDiff generates realistic and structurally-plausible protein sequences.**

**(A)** Workflow for evaluating the foldability and self-consistency of sequences generated by EvoDiff sequence models. **(B-C)** Distributions of foldability, measured by sequence pLDDT of predicted structures (B), and self-consistency, measured by scPerplexity (C), for sequences from the test set, EvoDiff models, and baselines (*n*=1000 sequences per model; box plots show median and interquartile range). **(D)** Sequence pLDDT versus scPerplexity for sequences from the test set (grey, *n*=1000) and the 640M-parameter OADM model EvoDiff-Seq (blue, *n*=1000). **(E)** Predicted structures and metrics for successfully expressed and characterized unconditional generations from EvoDiff-Seq, the 640M-parameter OADM model. OmegaFold predictions, colored by pLDDT, are shown, and the average pLDDT for each structure is reported. % coverage and % identity to the top BLAST hit is denoted below each design. **(F)** Circular dichroism (CD) spectra for designed sequences from (E). **(G)** Structural composition of each sequence design as inferred from CD spectra (blue) versus from OmegaFold (grey). AlphaFold predictions are included in Fig. S6 for comparison.

tasks (Table S3). Analysis of representative examples of structurally plausible sequences sampled from EvoDiff-Seq across 4 different sequence lengths illustrates their structural plausibility and novelty from sequences in the training set (Fig. S5).

To assess the stability and quality of EvoDiff generations *in vitro*, we used EvoDiff-Seq to unconditionally generate a set of 1000 sequences across four different sequence lengths. We then nominated 25 candidates, based on structural plausibility (OmegaFold pLDDT > 70), for experimental characterization (Supplementary Table 1). Four designs (EvoDiff-Seq-Designs-1...4, Fig. 2E) expressed and had circular dichroism (CD) spectra consistent with mixed alpha helical and beta sheet secondary structure elements (Fig. 2F-G). Two of our successful designs (EvoDiff-Seq-Design-3 and EvoDiff-Seq-Design-4) have OmegaFold pLDDT < 85, but many unsuccessful designs have a pLDDT > 85. In one case (EvoDiff-Seq-Design-1), the OmegaFold and AlphaFold predictions even have significant disagreement (> 1Å) in backbone RMSD, a common metric used to filter for designability in structure-based methods. These observations suggest that structural metrics are not always strong predictors of *in vitro* expressibility. Together, these *in silico* and *in vitro* results demonstrate that EvoDiff generates protein sequences that are individually valid.

## Biological properties of generated sequence distributions

Having shown that EvoDiff's generations are individually foldable and self-consistent, we next evaluated how well the *distribution* of designed protein sequences covered natural protein space. Ideally, generated sequences should capture the natural distribution of sequence, structural, and functional properties while still being diverse from each other and from natural sequences.

Previous work has shown that even without explicit supervision, protein language model embeddings contain information about both sequence and function as captured in GO annotations (35, 36). To evaluate coverage over the distribution of sequence and functional properties, we embedded each generated sequence using ProtT5 (37), a protein language model explicitly benchmarked for imputing GO annotations (35), and calculated the embedding space Fréchet distance between a set of generated sequences and the test set, where lower distance reflects better coverage. We refer to this metric as the Fréchet ProtT5 distance (FPD) and visualize these embeddings and the corresponding FPDs for sequences generated by EvoDiff-Seq and baseline models (Figs. 3A, S7, S8; Table S1). For RFdiffusion, we unconditionally generated 1000 structures with the same lengths as for EvoDiff-Seq and then used ESM-IF (33) to design their sequences. Both qualitatively and quantitatively, EvoDiff-Seq generates proteins that better recapitulate natural sequence and functional diversity than sampling from a state-of-the-art protein masked language model (ESM-2) or predicting sequences from structures generated by a state-of-the-art structure diffusion model (RFdiffusion) (Fig. 3A).

To evaluate the distribution of structural properties in generated sequences, we computed 3-state secondary structures (38) for each residue in generated and natural sequences and quantitatively compared the resulting distributions of structural properties to the distribution for the test set (Figs. 3B, S9). EvoDiff-Seq generates proportions of strands and disordered regions that are much more similar to those in natural sequences, while ESM-2 and RFdiffusion both generate proteins enriched in helices (Fig. 3B). To ensure our models were not memorizing training data, we calculated the Hamming distance between each generated sequence and all training sequences of the same length and reported the minimum Hamming distance, representing the closest match of any generated sequence to any sequence in the train set (Table S1). On average, a sequence generated from EvoDiff-Seq has a Hamming distance of 0.83 from the most similar training distance of the same length. Together, these results demonstrate, via comparison to ESM-2 and RFdiffusion, that EvoDiff's diffusion objective and evolutionary-scale training data are both necessary to generate novel sequences that cover protein sequence, functional, and structural space.

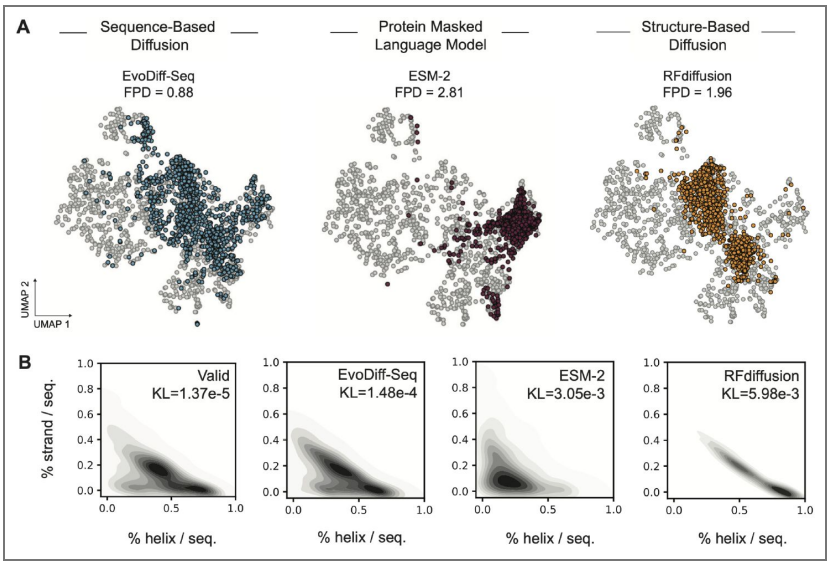



**Figure 3. Generated protein sequences capture natural distributions of protein functional and structural features.**

**(A)** UMAP of ProtT5 embeddings, annotated with FPD, of natural sequences from the test set (grey, *n* =l000) and of generated sequences from EvoDiff-Seq (blue, *n* =l000) and ESM-2 (red, *n* =l000), and inferred sequences inverse-folded from structures from RFdiffusion (orange, *n*=l000). **(B)** Multivariate distributions of helix and strand structural features in generated sequences, based on DSSP 3-state predictions (*n* =l000 samples from each model or the validation set) and annotated with the Kullback-Leibler (KL) divergence relative to the test set.

## Conditional sequence generation for controllable design

EvoDiff's OADM diffusion framework induces a natural method for conditional sequence generation by fixing some subsequences and inpainting the remainder. Because the model is trained to generate proteins with an arbitrary decoding order, this is easily accomplished by simply masking and decoding the desired portions. We applied EvoDiff's power for controllable protein design across three scenarios: conditioning on evolutionary information encoded in MSAs, inpainting functional domains, and scaffolding functional structural motifs (Fig. 1D).

## Evolution-guided protein generation with EvoDiff-MSA

First, we tested the ability of EvoDiff-MSA to generate query sequences conditioned on the remainder of an MSA, thus generating new members of a protein family without needing to train family-specific generative models. We masked the query sequences from 250 randomly-chosen MSAs from the validation set and newly generated these sequences using EvoDiff-MSA. We then evaluated the quality of the resulting conditionally-generated query sequences via our foldability and self-consistency pipeline (Fig. 4A). We find that EvoDiff-MSA generates more foldable and self-consistent sequences than sampling from ESM-MSA (28) or using Potts models (39) trained on individual MSAs (Figs. 4B-C, S10; Table S4). To evaluate sample diversity, we computed the aligned residue-wise sequence similarity between the generated query sequence and the most similar sequence in the original MSA. In contrast to sampling from a Potts model, generating from EvoDiff-MSA yields sequences that exhibit strikingly low similarity to those in the original MSA (Figs. 4D; Table S4) while still retaining structural integrity relative to the original query sequences (Figs. 4E-F; Table S4) To showcase these properties, we visual ize OmegaFold-predicted structures and evaluation metrics for a sample of high pLDDT, low scPerplexity conditionally-generated query sequences that exhibit low sequence similarity to anything in the conditioning MSA (Fig. 4G). These results indicate that EvoDiff-MSA can conditionally generate novel, structurally plausible members of a protein family given guidance from evolutionary information and without further finetuning.

## Generating intrinsically disordered regions

Because it generates directly in sequence space, we hypothesized that EvoDiff could natively generate intrinsically disordered regions (IDRs). IDRs are regions within a protein that lack secondary or tertiary structure; up to 30% of eukaryotic proteins contain at least one IDR, and IDRs make up over 30% of the residues in eukaryotic proteomes (19). IDRs carry out important and diverse functional roles in the cell directly facilitated by their lack of structure, such as subcellular trafficking (40, 41), protein-protein interactions (42, 43), and signaling (44). Altered abundance and mutations in IDRs have been implicated in human disease, including neurodegeneration and cancer (45–47). Despite their prevalence and critical roles in function and disease, IDRs do not fit neatly in the structure-function paradigm and remain outside the capabilities of structure-based protein design methods.

Having observed that unconditional generation using EvoDiff-Seq produced a similar fraction of residues predicted to lack secondary structure as that in natural sequences (Fig. 3B), we used inpainting with EvoDiff-Seq and EvoDiff-MSA to intentionally generate disordered regions via conditioning on their surrounding structured regions (Fig. 5A). To accomplish this, we leveraged a previously curated dataset of computationally predicted IDRs covering the human proteome (48). We selected this dataset because it also curates orthologs for these proteins, enabling construction of MSAs (48). After using EvoDiff to generate putative IDRs via inpainting, we then predicted disorder scores for each residue in the generated and natural sequences using DR-BERT (49) (Figs. 5A, S11). Over 100 generations, we observe that IDR regions inpainted by EvoDiff-Seq and EvoDiff-MSA result in distributions of disorder scores similar to those for natural sequences, across both the IDR and the surrounding structured regions (Figs. 5B, S12). Generations from EvoDiff-MSA exhibit strong correlation in predicted disorder scores to those of true IDRs (Fig. S12). Although putative IDRs generated by EvoDiff-Seq are less similar to their

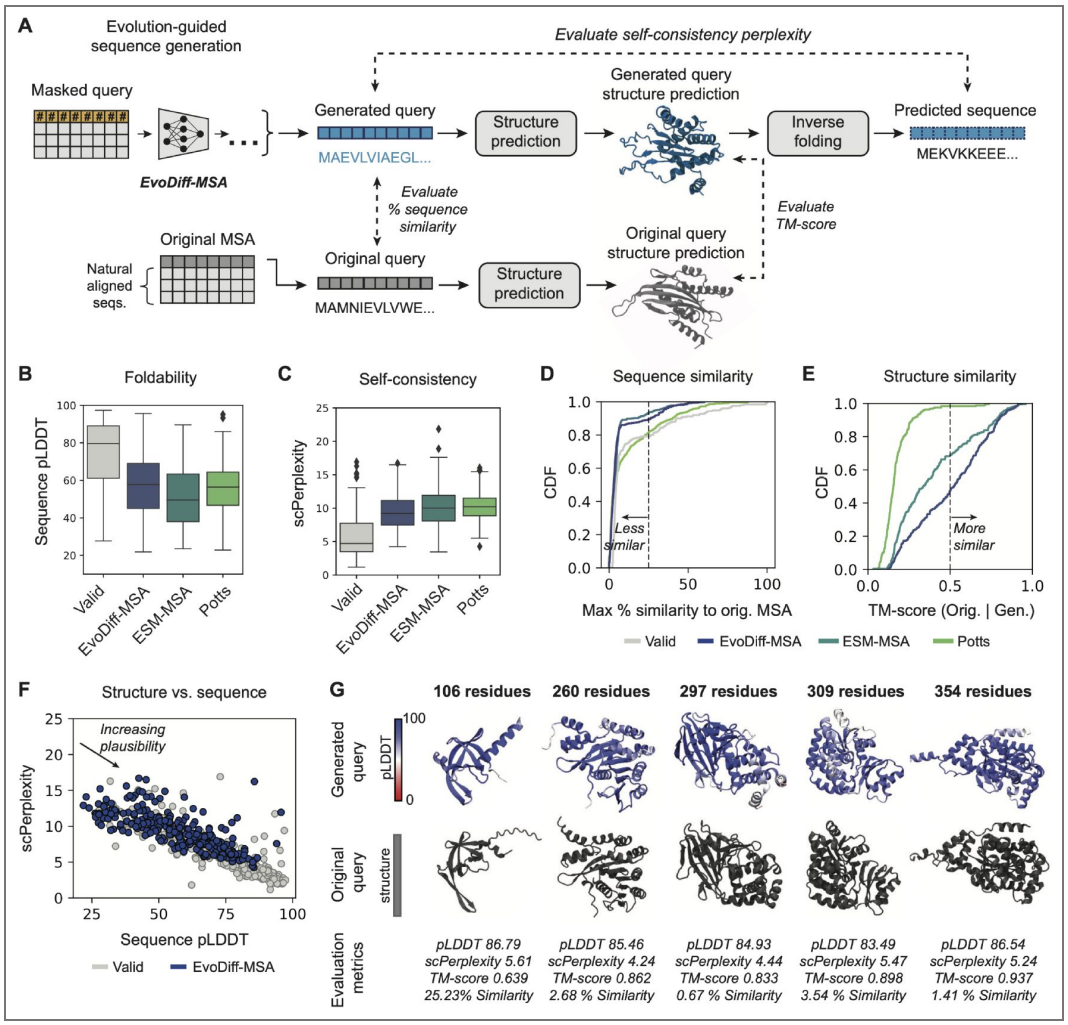



**Figure 4. EvoDiff-MSA enables evolution-guided sequence generation.**

**(A)** A new sequence is generated from EvoDiff-MSA via diffusion over only the query component. Generations are evaluated for diversity and self-consistency and for the quality and consistency of their predicted structures. **(B-E)** Distributions of pLDDT (B), scPerplexity (C), sequence similarity (D; dashed line at 25%), and TM-score (E; dashed line at 0.5) for sequences from the validation set, EvoDiff-MSA, ESM-MSA, and a Potts model (*n*=250 sequences per model; box plots show median and interquartile range). **(F)** Sequence pLDDT versus scPerplexity for sequences from the validation set (grey, *n*=250) and EvoDiff-MSA (blue, *n*=250). **(G)** Predicted structures and metrics for structurally plausible generations from EvoDiff-MSA.

original IDR than those from EvoDiff-MSA (Fig. 5C), both models generated disordered regions that preserve disorder scores over the entire protein sequence and still exhibit low sequence similarity to the original IDR (Fig. S13).

We next sought to validate experimentally that inpainted IDRs maintain cellular function. Many IDRs are targeting signals that drive precise subcellular localization critical to the protein's underlying function. Notably, the vast majority of mitochondrial proteins contain N-terminal disordered mitochondrial targeting signals that are cleaved after import into the mitochondria (40, 41, 50–52). We hypothesized that, given the mature protein as context, EvoDiff can design N-terminal IDRs that drive mitochondrial localization in cells. To test this hypothesis, we masked out the N-terminal IDR (amino acid positions 1-45, inclusive) of the yeast Heme A synthase Cox15 and inpainted 200 IDR candidates of varying lengths from EvoDiff-Seq and EvoDiff-MSA (Rand subsampling) (Fig. 5D). We predicted disorder with DR-BERT and mitochondrial targeting propensity with MitoFates (53) for each candidate (Fig. S14A-B) and selected the top eight generated sequences (Fig. S14C-D), four from each of EvoDiff-Seq and EvoDiff-MSA, for experimental characterization (Supplementary Table 1).

We replaced the native Cox15 N-terminal IDR with each of the generated IDRs and assessed the cellular localization of the resultant proteins using a microscopy screen in yeast cells. Briefly, we tagged the generated constructs with super-folding GFP (sfGFP) and recombined them into a yeast strain expressing endogenous Cox15 tagged with mScarlet, such that overlap between sfGFP (green) and mScarlet (red) fluorescence would indicate co-localization of the designed and endogenous Cox15 and thus demonstrate that the generated IDRs successfully induce mitochondrial transport (Fig. 5E). As negative controls, we used both a knockout of the native Cox15 N-terminal IDR (Δ2-45) as well as the N1 N-terminal sequence, which was designed in a previous IDR study to match the bulk biophysical properties of the Cox15 mitochondrial targeting signal but lacks mitochondrial targeting activity, as predicted by MitoFates and demonstrated experimentally (54). Consistent with our hypothesis, all 8 EvoDiff-IDRCox15 variants exhibit a subcellular localization pattern consistent with the localization of the native Cox15, suggesting that EvoDiff can design IDRs that function in their cellular context (Fig. 5F-G, S15-S17).

## Scaffolding functional motifs with sequence information alone

Thus far, the primary application of deep generative models of protein structure in protein engineering is their ability to scaffold binding and catalytic motifs: given the 3D coordinates of a functional motif, these models can often generate a structural scaffold that holds the motif in precisely the 3D geometry needed for function (10, 14, 55). Given that the fixed functional motif includes the residue identities for the motif, we investigated whether a structural model is actually necessary for motif scaffolding.

We used conditional generation with EvoDiff to generate scaffolds for a diverse set of 17 motif-scaffolding problems (10) by fixing the functional motif, supplying only the motif's amino-acid sequence as conditioning information, and then decoding the remainder of the sequence (Fig. 1D). The problems include simple "inpainting", viral epitopes, receptor traps, small molecule binding sites, protein-binding interfaces, and enzyme active sites. Many of the motifs are not contiguous in sequence space. We compared the performance of EvoDiff, which uses only sequence information, to the state-of-the-art structure model RFdiffusion, and facilitated direct comparisons by using OmegaFold to predict structures for our generated sequences as well as for sequences inverse-folded from RFdiffusion structures. Notably, we use the same EvoDiff models for both unconditional and conditional generation, while the version of RFdiffusion used for scaffolding is finetuned from that used for unconditional generation.

We evaluated the ability of each of EvoDiff-Seq, EvoDiff-MSA, and RFdiffusion to generate successful scaffolds (Fig. 6A-B), where we define a scaffold to be successful if the predicted motif coordinates have less than 1Å RMSD from the desired motif coordinates and if the scaffolded sequence has a pLDDT > 70. Despite operating entirely in sequence space, EvoDiff-Seq and EvoDiff-MSA generate successful scaffolds for 8 and 13 of the 17 problems, respectively (Table S5, S6). EvoDiff-MSA has a higher success rate than EvoDiff-Seq for 10 problems and a higher success rate

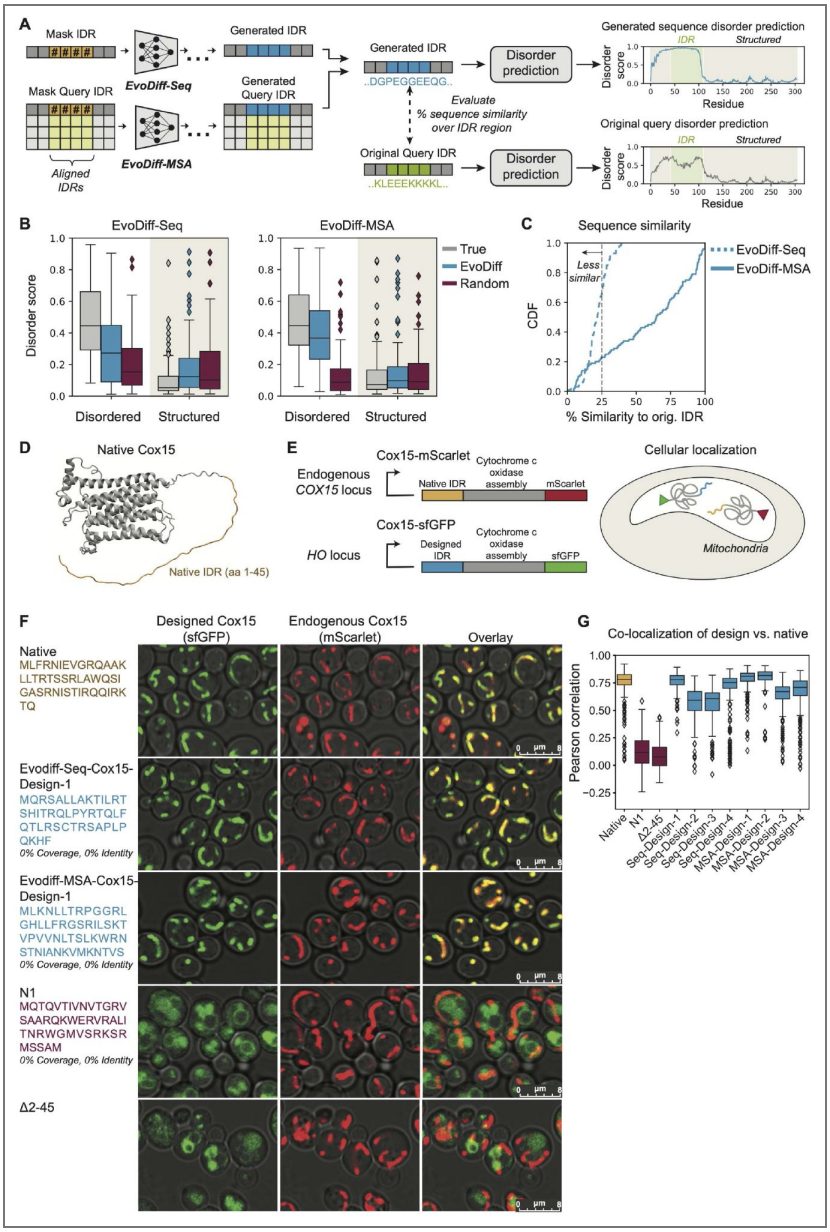



**Figure 5. EvoDiff generates intrinsically disordered regions.**

**(A)** A new IDR sequence is generated from EvoDiff-Seq or EvoDiff-MSA by inpainting disordered residues in the query sequence. DR-BERT is then used to predict disorder scores for the original and regenerated sequences. **(B)** Distributions of disorder scores over disordered and structured regions for sequences with true (grey), inpainted (blue), and randomly-sampled (red) IDRs (*n*=100 sequences per condition; box plots show median and interquartile range). **(C)** Distribution of sequence similarity relative to the original IDR for generated IDRs from EvoDiff-Seq (blue, dashed) and EvoDiff-MSA (blue, solid) (*n*=100; dashed line at 25%). **(D)** Structure of the yeast Cox15 protein, with the disordered N-terminal mitochondrial targeting signal highlighted in yellow. EvoDiff is used to inpaint this disordered region. **(E)** Microscopy assay in yeast to assess cellular localization of Cox15 proteins with EvoDiff-designed targeting signals. Mitochondrial localization is measured by quantifying the fluorescence overlap between endogenous Cox15 (mScarlet) and designed Cox15 (sfGFP). **(F)** Fluorescence microscopy imaging of yeast cells. Columns from left to right show: the N-terminal IDR sequence used for a given Cox15-sfGFP design (rows); signal from the designed Cox15 (sfGFP, colored in green); signal from the endogenous Cox15 (mScarlet, colored in red); overlaid images. All images show brightfield in grey, to visualize cell boundaries. Images for all 8 designs are included in Fig. S15 and S16. Scale bars = 8 $\mu$m. **(G)** Per-cell pixel-wise Pearson correlation coefficients of signal intensities between the mScarlet channel and the sfGFP channel over *n*=183-313 cells segmented by YeastSpotter for each design. Correlations for a biological replicate are shown in Fig. S17.

than RFdiffusion for 6 problems. EvoDiff-Seq has a higher success rate than RFdiffusion for 2 problems and a higher success rate than EvoDiff-MSA for 3 problems. There are two scaffolding problems (1YCR, 3IXT) where EvoDiff-MSA is outperformed by both EvoDiff-Seq and RFdiffusion (Table S5, S6). These are both examples where, for scaffolding, an MSA containing fewer than 64 protein sequences was input to EvoDiff-MSA, which did not see MSAs with fewer than 64 sequences during training.

Interestingly, there is almost no correlation between the problem-specific success rates of EvoDiff and RFdiffusion, and there are very few problems for which both methods have high success rates, showing that EvoDiff may have orthogonal strengths to RFdiffusion (Fig. 6A-B). Due to its conditioning on evolutionary information, EvoDiff-MSA generates scaffolds that are more structurally similar to the native scaffold than EvoDiff-Seq (Fig. 6C). To ensure that EvoDiff is not finding trivial solutions, we show that it outperforms both random generation and the single-order LRAR model (which decodes unconditionally up to and after a motif) (Table S5). ESM-MSA performs similarly to EvoDiff-MSA on this task, as the motif scaffolding task is well-aligned with its training task, and it is trained on approximately 200x more MSAs than EvoDiff-MSA (Table S6). We illustrate examples of successful scaffolds sampled from EvoDiff and note both the qualitative and quantitative quality of generated proteins and predicted structures across a range of functional motifs (Fig. S18).

To validate EvoDiff's ability to design functional scaffolds around structural motifs, we generated scaffolds for two distinct targets – calcium ($Ca^{2+}$) and MDM2 – by conditioning on the binding motif sequences from calmodulin (1PRW) and p53 (1YCR), respectively. We do this without providing any explicit information about the binding target. First, we generated 100 scaffolds each from EvoDiff-Seq, EvoDiff-MSA (Max subsampling), and EvoDiff-MSA (Rand subsampling) for the discontinuous 1PRW binding domain (Fig. 6D). Then, we nominated 13 designs, based on *in-silico* metrics (pLDDT > 70 and motifRMSD < 1Å), to express and evaluate experimentally (Supplementary Table 1). 12 of these designs expressed successfully, and one of these designs (EvoDiff-Seq-1PRW-Design-1) demonstrated a spectroscopic shift in the presence of $Ca^{2+}$ (Fig. 6F, S19), indicative of calcium binding (56). We show the predicted structure of the successful binder in Fig. 6E. Binding experiments were repeated independently in triplicate (Fig. S19).

Next, we scaffolded the transactivation domain of p53 (Fig. 6G), which binds MDM2. Given the importance of p53 as a tumor suppressor in cancer, competitive inhibition of the p53-MDM2 interaction via a synthetic scaffold could promote tumor-suppressing activity of native p53 (57). Unlike in structure-based methods (10), we did not include information about the binder during inference; additionally, we note that p53 contains an intrinsically disordered N-terminal domain (58) (Fig. 6G). Starting with 900 EvoDiff-Seq and 400 EvoDiff-MSA generations, we nominated 24 total designs (Supplementary Table 1). For EvoDiff-Seq we nominated 19 designs with OmegaFold pLDDT > 70 and motifRMSD < 1Å. However, no EvoDiff-MSA generations exceeded the pLDDT threshold, so we nominated 5 candidates from EvoDiff-MSA using only motifRMSD. We screened all 24 designs *in vitro* for binding to full-length human MDM2 using biolayer interferometry (BLI) (Fig. S20-S24). All designs expressed in a cell-free system and four designs (EvoDiff-Seq-p53-Design1 and Evodiff-MSA-p53-Design-1…3) demonstrated strong binding to full-length MDM2, categorized as $K_D$ < 50 nM observed over at least two replicates. The strongest binders, EvoDiff-MSA-p53-Design-2 and EvoDiff-MSA-p53-Design-3, bind with an average $K_D$ of 25.9 nM and 40.2 nM while having 85% and 53% sequence similarity, respectively, to any natural protein sequence (Fig. 6H, S22; Supplementary Table 1). In total, 8 designs demonstrated evidence of binding at varying strengths, as measured across three replicates (Fig. 6I, S21, S22). These *in silico* and *in vitro* results demonstrate that EvoDiff can design functional scaffolds around structural motifs via conditional generation in sequence space alone.

# Discussion

We present EvoDiff, a diffusion modeling framework capable of generating high-fidelity, diverse, and novel proteins with the option of conditioning according to sequence constraints.

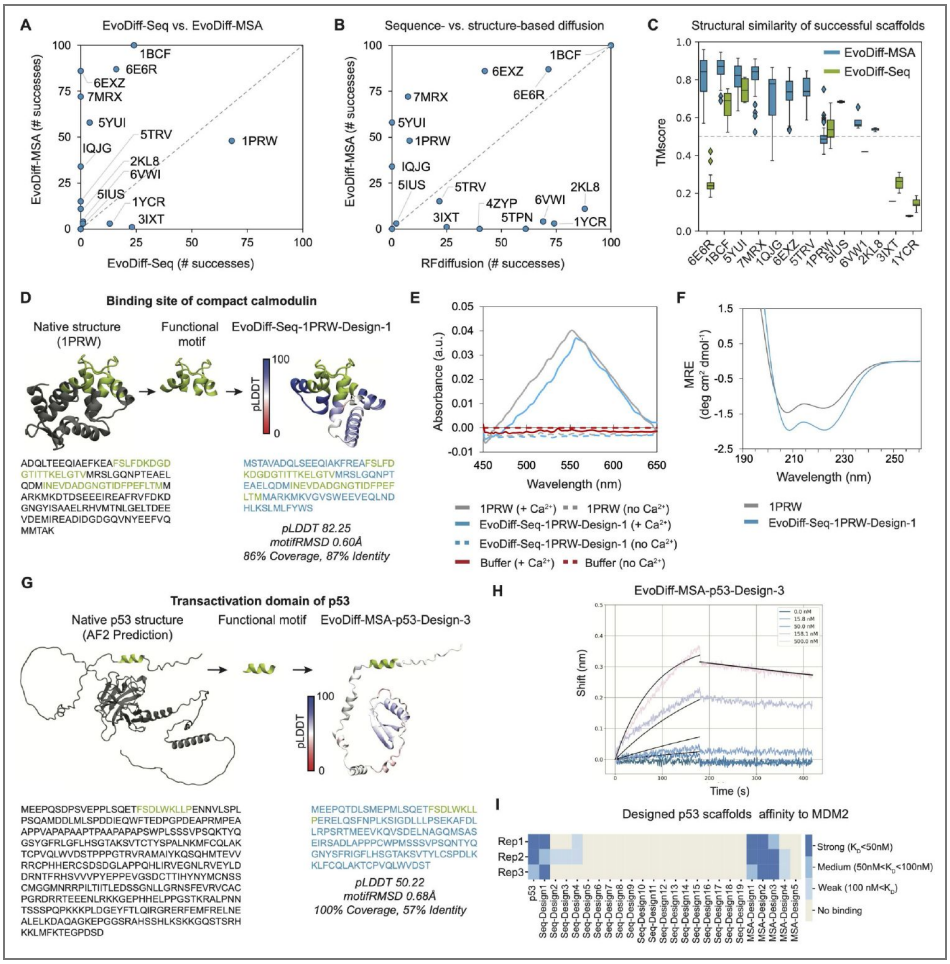



**Figure 6. EvoDiff scaffolds functional motifs without explicit structural information.**

**(A)** Number and identity of successful scaffolds from *n*=100 trials for EvoDiff-Seq (x-axis) versus EvoDiff-MSA (y-axis) across scaffolding problems in which at least one method succeeds. **(B)** Performance comparison of sequence-based scaffolding via EvoDiff-MSA (y-axis) versus structure-based scaffolding via RFdiffusion (x-axis) across scaffolding problems in which at least one method succeeds (*n*=100 trials per problem). **(C)** Distributions of TM-scores of successfully generated scaffolds from EvoDiff models relative to the true structures (dashed line at 0.5; box plots show median and interquartile range). **(D)** EvoDiff-Seq successfully scaffolds the binding site of compact calmodulin. The generated sequence, OmegaFold-predicted structure (colored by pLDDT), and computed metrics for EvoDiff-Seq-1PRW-Design1 are shown. Scaffolding motif is shown in green, and the structure and sequence of calmodulin (1PRW) are shown in dark grey. **(E)** Background-corrected absorbance spectra of the native 1PRW (grey), EvoDiff-Seq-1PRW-Design-1 (blue), and control buffers (red) with (solid line) and without (dashed lines) $Ca^{2+}$. **(F)** CD spectra of native 1PRW (grey) and EvoDiff-Seq-1PRW-Design-1 (blue). **(G)** EvoDiff successfully scaffolds the transactivation domain of p53. Generated sequence, OmegaFold-predicted structure (colored by pLDDT), and computed metrics for EvoDiff-MSA-p53-Design-3, are shown. The transactivation domain is shown in green, and the AlphaFold-predicted structure and sequence of p53 are shown in dark grey. **(H)** BLI measurements for EvoDiff-MSA-p53-Design-3 measured over 5 concentrations. **(I)** *In vitro* binding affinities of all EvoDiff designs against the full-length human MDM2. Measurements for all designs were performed in triplicate (Fig. S20-S24).

We validate EvoDiff's generative capabilities both *in silico* and in the lab. Because it operates in the universal protein design space, EvoDiff can unconditionally sample diverse structurally-plausible proteins that express and exhibit stable secondary structure, generate functional intrinsically disordered regions, and scaffold structural motifs using only sequence information, challenging a paradigm in structure-based protein design.

EvoDiff is the first suite of deep learning models to demonstrate the power of diffusion generative modeling over evolutionary-scale protein sequence space. Unlike previous attempts to train diffusion models on protein structures (7–15) and/or sequences (59–63), EvoDiff is trained on a large, diverse sample of all natural sequences, rather than on smaller protein structure datasets or sequence data from a specific protein family. Previous protein generative models trained on global sequence space have been either left-to-right autoregressive (LRAR) models (64–67) or masked language models (MLMs) (27, 28, 30, 37, 68). EvoDiff's OADM training task generalizes the LRAR and MLM training tasks. Specifically, the OADM setup generalizes LRAR by considering all possible decoding orders, while the MLM training task is equivalent to training on one step of the OADM diffusion process.

This generalized mathematical formulation yields empirical benefits, as EvoDiff-Seq produces sequences that better cover protein functional and structural space than sampling from state-of-the-art protein MLMs (Fig. 3). While an LRAR model learned to fit the evolutionary sequence distribution better (Table S1), the fixed decoding order of traditional left-to-right autoregression restricts conditional generation. EvoDiff directly addresses this barrier by enabling different forms of conditioning, including evolution-guided generation (Fig. 4) as well as inpainting and scaffolding (Figs. 5-6). We report the first demonstrations of these programmable generation capabilities from deep generative models of protein sequence alone.

Future work may expand these capabilities to enable conditioning via guidance, in which generated sequences can be iteratively refined to fit desired properties (69, 70). While we observe that OADM generally outperforms D3PM in unconditional generation, likely because the OADM denoising task is easier to learn than that of D3PM, conditioning via guidance intuitively fits into the EvoDiff-D3PM framework because the identity of each residue in a sequence can be edited at every decoding step. OADM and existing conditional LRAR models, such as ProGen (64), both fix the identity of each amino acid once it is decoded, limiting the effectiveness of guidance. Guidance-based conditioning of EvoDiff-D3PM should enable the generation of new protein sequences specifying functional objectives, such as those specified by sequence-function classifiers.

Because EvoDiff only requires sequence data, it can readily be extended for diverse down-stream applications, including those not reachable from a traditional structure-based paradigm. As a first example, we have demonstrated EvoDiff's ability to generate functional IDRs – overcoming a prototypical failure mode of structure-based predictive and generative models – via inpainting without fine-tuning. Fine-tuning EvoDiff on application-specific datasets, such as those from display libraries or large-scale screens, may unlock new biological, therapeutic, or scientific design opportunities that would be otherwise inaccessible due to the cost of obtaining structures for large sequence datasets. Experimental data for structures is much sparser compared to sequences, and while structures for many sequences can be predicted using AlphaFold and similar algorithms, these methods do not work well on point mutants and can be overconfident on spurious proteins (71, 72).

While we demonstrated some coarse-grained strategies for conditioning generation through scaffolding and inpainting, to achieve even more fine-grained control over protein function, with future development EvoDiff may be conditioned on text, chemical information, or other modalities. For example, text-based conditioning (73) could be used to ensure that generated proteins are soluble, readily expressed, and non-immunogenic. Future use cases for this vision of controllable protein sequence design include programmable modulation of nucleic acids via conditionally-designed transcription factors or endonucleases, improved therapeutic windows via biologics optimized for *in vivo* delivery and trafficking, as well as newly-enabled catalysis via zero-shot tuning of enzyme substrate specificity.

In summary, we present an open-source and experimentally-validated suite of discrete diffusion models that provide a foundation for sequence-based protein engineering and design. EvoDiff models can be directly deployed for unconditional, evolution-guided, and conditional generation of protein sequences and may be extended for guided design based on structure or function. We envision that EvoDiff will enable new abilities in controllable protein design by reading and writing function directly in the language of proteins.

# Methods

## Diffusion models

Diffusion models are a class of generative models that learn to generate data from noise. They consist of a forward corruption process and a learned reverse denoising process. The forward process is a Markov chain of diffusion steps $q(x_t|x_{t-1})$ that corrupts an input ($x_0$) over $T$ timesteps such that $x_T$ is indistinguishable from random noise. The learned reverse denoising process $p_\theta(x_{t-1}|x_t)$ is parameterized by a model such as a neural network and generates new data from noise. Discrete diffusion models have previously been developed over binary random variables (3), developed over categorical random variables with uniform transition matrices (74, 75), linked to autoregressive models (24), and optimized for use with transition matrices (25).

This work presents models from two different discrete diffusion frameworks – order-agnostic autoregressive diffusion models (OADMs) and discrete denoising diffusion probabilistic models (D3PMs) – on protein sequences and multiple sequence alignments (MSAs).

## Discrete Denoising Diffusion Probabilistic Models (D3PMs)

Discrete denoising diffusion probabilistic models (D3PMs) operate by defining a transition matrix $Q$ such that, over $T$ timesteps, discrete inputs (i.e. protein amino-acid sequences for EvoDiff) are iteratively corrupted via a controlled Markov process until they constitute samples from a uniform stationary distribution at time $T$. This section describes the D3PM process and loss for a single categorical variable $x$ in one-hot format. The forward corruption process is described by:

$$q\left(x_t \mid x_{x-1}\right) = \mathrm{Cat}\left(x_t; p = x_{x-1}Q_t\right). \tag{1}$$

This allows for efficient training via efficient computation of $q(x_t|x_0)$ and $q(x_{t-1}|x_t)$. The

D3PM approach can emulate a masked modeling process by choosing a transition matrix with an absorbing state (e.g., [MASK]; (25)). However, in this work, the D3PM formulation is only used for discrete corruption because masking corruption via OADM generally outperforms absorbing-state D3PM (24). EvoDiff includes two discrete corruption schemes: one based on a uniform transition matrix (D3PM-Uniform) and one based on a biologically-informed transition matrix (D3PM-BLOSUM).

EvoDiff-D3PM models are trained via a hybrid loss function

$$\mathscr{L}_\lambda = \mathscr{L}_{vb} + \lambda\mathscr{L}_{ce}. \tag{2}$$

This loss combines a variational lower bound $\mathscr{L}_{vb}$ on the negative log likelihood

$$\mathscr{L}_{vb} = \mathbb{E}_{q(x_0)} + \underbrace{[D_{KL}\left[q\left(x_T \mid x_0\right) \| p\left(x_T\right)\right]}_{\mathscr{L}_T} \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(x_t|x_0)}\left[D_{KL}\left[q\left(x_{t-1} \mid x_t, x_0\right) \| p_\theta\left(x_{t-1} \mid x_t\right)\right]\right]}_{\mathscr{L}_{t-1}} + \underbrace{-\mathbb{E}_{q(x_1|x_0)}[\log\left(p_\theta\left(x_0 \mid x_1\right)\right]}_{\mathscr{L}_0}. \tag{3}$$

and a cross-entropy loss $\mathscr{L}_{ce}$ on $p_\theta(x_0|x_t)$. Investigation of the impact of $\lambda$ on model performance revealed minimal improvement to sample generation quality when $\lambda > 0$, consistent with the findings of the original D3PM paper (25). Thus $\lambda$=0 and $T$ =500 were used in all D3PM experiments.

$\mathscr{L}_{vb}$ has three terms. $\mathscr{L}_T$ measures whether the corruption reaches the stationary distribution $p(x_T)$ at time $T$ and does not depend on $\theta$. Next consider the remaining two terms $\mathscr{L}_{t-1}$ and $\mathscr{L}_0$, which depend on $\theta$. Following the original D3PM paper, $\tilde{p}_\theta\left(\tilde{x}_0 \mid x_t\right)$ is directly predicted by the neural network. To compute the loss at timesteps $0 < t < T$, the terms $q(x_{t-1}|x_t, x_0)$ and $p_\theta(x_{t-1}|x_t)$ must be computed from $x_t$, $x_0$, and $\tilde{p}_\theta\left(\tilde{x}_0 \mid x_t\right)$ using Markov properties

Defining $Q_t = Q_1 Q_2 \cdots Q_t$:

$$q\left(x_{x-1} \mid x_t, x_0\right) = \mathrm{Cat}\left(x_{t-1}; p = \frac{x_t Q_t^\top \odot x_0 Q_{t-1}}{x_0 \bar{Q}_t x_t^\top}\right) \tag{4}$$

$$p_\theta\left(x_{t-1} \mid x_t\right) \propto \sum_{\tilde{x}_0} q\left(x_{t-1}, x_t \mid \tilde{x}_0\right) \tilde{p}_\theta\left(\tilde{x}_0 \mid x_t\right) \tag{5}$$

where $\odot$ represents an element-wise product. For Equation 5 rules of conditional probability and Markov properties are used to define $q\left(x_{t-1}, x_t \mid \tilde{x}_0\right)$ in terms of $x_t$ and $\tilde{x}_0$:

$$q\left(x_{x-1}, x_t \mid \tilde{x}_0\right) = \mathrm{Cat}\left(x_{x-1}; p = x_t Q_t \odot \tilde{x}_0 \bar{Q}_{t-1}\right) \tag{6}$$

Putting everything together, at each step of training a corruption timestep is sampled according to $t \sim \mathcal{U}$ (1, ..., $T$ − 1). $x_t$ is then sampled via $q\left(x_t \mid x_0\right) \sim \mathrm{Cat}\left(x_t; p = x_0 Q_t\right)$ for every residue in the input protein, and the neural network predicts $\tilde{p}_\theta\left(\tilde{x}_0 \mid x_t\right)$. Note that, while the corruption and loss are computed independently over each residue, the neural network predicts $p$ in the context of the entire sequence. If $t$ = 1, only the loss $\mathscr{L}_0$ is used, reflecting a standard negative log likelihood. Otherwise, Equations 4 and 5 are used to compute the loss $\mathscr{L}_{t-1}$.

Sampling from a trained model begins with the noised $x_T$, where each residue is randomly sampled from a uniform distribution over amino acids. $x_{t-1}$ is then iteratively sampled via $p_\theta(x_{t-1}|x_t)$ as described in Equation 5. For all models, generated sequences are sampled to match the distribution of sequence lengths in the training set, going up to 2048 residues as the maximum length.

## EvoDiff-D3PM-Uniform

Many strategies exists to schedule corruption in D3PMs. EvoDiff-D3PM-Uniform employs the simplest case – a uniform corruption scheme. Specifically, EvoDiff-D3PM-Uniform models implement a doubly stochastic, uniform transition matrix $Q_t$ with a corruption schedule $(T - t + 1)^{-1}$ from Sohl-Dickstein et al. (3), so that information is linearly corrupted between $x_t$ and $x_0$ for all $t < T$.

## EvoDiff-D3PM-BLOSUM

EvoDiff-D3PM-BLOSUM implements a transition matrix derived from BLOSUM62 matrices of amino acid substitution frequencies (76). BLOSUM matrices are derived from observed alignments across highly conserved regions of protein families and thus provide the relative frequencies of amino acids and their substitution probabilities.

Rows that represent uniform transition probabilities for non-standard amino acid codes (J, O, U) and for < GAP > tokens in the MSA input case are included in addition to standard amino acids. BLOSUM substitution frequencies are converted to a matrix of transition probabilities by performing a softmax over the frequencies and then normalizing over rows and columns via the Sinkhorn-Knopp algorithm to obtain a doubly stochastic matrix. In this scheme, the gradual corruption of a single sequence to random noise is simulated in a way that prioritizes conserved evolutionary relationships of amino acid mutations. A $\beta$-schedule was implemented to taper the number of mutations over time for timesteps up to $T$ =500, specifically via an empirical schedule that corrupts half the sequence content by half of $T$ ($t$=250) (Fig. S25). This schedule was chosen to approximate the linear rate of mutations observed over 500 timesteps in the uniform transition matrix case, shown in Fig. S25b.

## Order-Agnostic Autoregressive Diffusion Models (OADMs)

Order-agnostic autoregressive diffusion models (OADMs) generalize absorbing-state D3PM and left-to-right autoregressive models (LRARs) (24). This section describes the OADM process and loss for a sequence $x$ of $L$ categorical variables. In the case of EvoDiff, $L$ is the sequence length.

LRARs factorize a high-dimensional joint distribution $p(x)$ into the product of $L$ univariate distributions using the probability chain rule:

$$\log p(\boldsymbol{x}) = \sum_{t=1}^{L} \log p\left(x_t \mid \boldsymbol{x}_{<t}\right) \tag{7}$$

where $x_{<t} = x_1, x_2, \ldots x_{t-1}$. LRARs are typically parametrized using a triangular dependency structure, such as causal masking in a transformer or CNN, in order to allow parallelized computation of all the conditional distributions in the likelihood during training. LRARs learn to generate sequences in a pre-specified left-to-right decoding order, which may be non-obvious for modalities such as proteins and does not allow conditioning on arbitrary fixed subsequences.

LRARs can be expanded into a diffusion framework via two subtle changes. Following the exposition in Hoogeboom *et al.*, (24), the first change is to allow order-agnostic decoding. In an order-agnostic autoregressive model, a decoding order $\sigma$ is first sampled uniformly from all possible decoding orders $S_L$. At time step $t$ in the forward process, $x_{\sigma(L-t)}$ is masked. The log-likelihood for an order-agnostic autoregressive model is derived using Jensen's inequality:

$$\begin{aligned}\log p(\boldsymbol{x}) = \log \mathbb{E}_{\sigma\sim\mathcal{U}(S_L)} p(\boldsymbol{x} \mid \sigma) &\geq \mathbb{E}_{\sigma\sim\mathcal{U}(S_L)} \log p(\boldsymbol{x} \mid \sigma) \\ &\geq \mathbb{E}_{\sigma\sim\mathcal{U}(S_L)} \sum_{t=1}^{L} \log p\left(x_{\sigma(t)} \mid \boldsymbol{x}_{\sigma(<t)}\right)\end{aligned}$$

The next change involves an objective that optimizes over arbitrary decoding orders one timestep at a time in the style of modern diffusion models, without requiring a neural network that enforces a triangular or causal dependency structure. This is accomplished by replacing the summation over $t$ by an expectation that is appropriately re-weighted.

$$\begin{aligned}\log p(\boldsymbol{x}) &\geq \mathbb{E}_{\sigma\sim\mathcal{U}(S_L)} \sum_{t=1}^{L} \log p\left(x_{\sigma(t)} \mid \boldsymbol{x}_{\sigma(<t)}\right) \\ &= \mathbb{E}_{\sigma\sim\mathcal{U}(S_L)} L \cdot \mathbb{E}_{t\sim\mathcal{U}(1,\ldots,L)} \log p\left(x_{\sigma(t)} \mid \boldsymbol{x}_{\sigma(<t)}\right) \\ &= L \cdot \mathbb{E}_{t\sim\mathcal{U}(1,\ldots,L)} \mathbb{E}_{\sigma\sim\mathcal{U}(S_L)} \frac{1}{L-t+1} \sum_{k\in\sigma(\geq t)} \log p\left(x_k \mid \boldsymbol{x}_{\sigma(<t)}\right)\end{aligned}$$

The overall expected log likelihood log $p(x)$ can be thought of according to a series of likelihoods, each captured in the loss at step $t$, $\mathcal{L}_t$:

$$\mathcal{L}_t = \frac{1}{L-t+1} \mathbb{E}_{\sigma\sim\mathcal{U}(S_L)} \sum_{k\in\sigma(\geq t)} \log p\left(x_k \mid \boldsymbol{x}_{\sigma(<t)}\right). \tag{8}$$

Thus, the overall expected log likelihood is lower bounded as:

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{t\sim\mathcal{U}(1,\ldots,L)} \left[L \cdot \mathcal{L}_t\right] \tag{9}$$

A neural network can be efficiently trained to learn the reverse process $p_\theta(x_{\sigma(t)} | x_{\sigma(<t)})$ by randomly masking a set of $t$ tokens at each iteration and minimizing the reweighted loss, allowing the model to learn from predictions of all masked positions at each timestep. By learning one model over all possible decoding orders, OADM allows for conditioning by fixing arbitrary subsequences at generation time. Sequences were generated unconditionally from OADM models by beginning with an all-mask sequence as input, randomly sampling a decoding order, and sampling each token from the predicted probability distribution.

## Left-to-right autoregressive and masked language models are diffusion models

The connection between autoregressive models and diffusion models has been described previously (24, 25). Left-to-right autoregressive (LRAR) diffusion models implement a masked modeling process that is akin to a process which iteratively and deterministically masks all tokens to the right of the sampled token $x_t$, where the current diffusion timestep $t$ is equivalent to the number of tokens masked over the entire sequence length, with all tokens masked at the final timestep $T = L$.

Likewise, masked language models (MLMs) are equivalent to only learning one step $t$ of OADM:

$$\mathcal{L}_{\text{MLM}} = \frac{1}{L-t+1} \mathbb{E}_{\sigma\sim\mathcal{U}(S_L)} \sum_{k\in\sigma(\geq t)} \log p\left(x_k \mid \boldsymbol{x}_{\sigma(<t)}\right). \tag{10}$$

Thus, the OADM setup generalizes LRAR models by considering all possible decoding orders rather than left-to-right decoding, while the MLM learning task is equivalent to only training on one step of the OADM diffusion process.

## Datasets

Sequence-only EvoDiff models were trained on UniRef50 (26) which contains approximately 45 million protein sequences. The UniRef50 release and train/validation/testing splits from CARP (27) were used to facilitate comparisons between models. Sequences longer than 1024 residues were randomly subsampled to 1024 residues. This split includes the random assignment of 82,929 sequences to the validation set and 209,426 sequences to the test set. Multiple sequence alignment (MSA) EvoDiff models were trained on OpenFold (29), which contains 401,381 MSAs for 140,000 unique Protein Data Bank (PDB) chains and 16,000,000 UniClust30 clusters. To construct the MSAs used to train EvoDiff, insertions, denoted by lowercase characters, were removed to restore the alignments, as the queries do not contain gap characters. Next, MSAs that contained sequences

with more than 512 consecutive < GAP > tokens as well as MSAs that contained fewer than 64 sequences per alignment were filtered out. This filtering resulted in 382,296 total MSAs, which were then randomly split into 372,296 training and 10,000 validation MSAs.

## MSA subsampling for training EvoDiff-MSA models

To optimize for memory constraints during training, MSAs were subsampled to 64 sequences and a maximum sequence length of 512. MSAs shorter than 512 sequences were padded to a sequence length of 512, but MSAs containing fewer than 64 sequences were excluded from training. For MSAs with more than 64 sequences, two subsampling schemes were implemented: random ("Rand.") and MaxHamming ("Max"). The random subsampling scheme ("Rand.") randomly samples 64 sequences from the MSA, making sure that the reference/query sequence (i.e. the first sequence) is always included. The "Max" subsampling scheme greedily selects for sequence diversity in the 64 sequence subset by iteratively selecting the sequence that maximizes the minimum Hamming distance to the sequences already selected. The Hamming distance measures the distance between two sequences, denoted by the number of amino acids that differ between aligned sequences. Subsampling to maximize the Hamming distance enabled input of an MSA rich with evolutionarily diverse sequences to EvoDiff-MSA models.

## Modeling, architecture, and training details

For sequences, the EvoDiff denoising model adopts a ByteNet-style CNN architecture (77) previously shown to perform similarly to transformers for protein sequence masked language modeling tasks (27). All models are implemented in PyTorch (78). In EvoDiff-OADM models, the diffusion timestep is implicitly encoded in the number of masked positions. EvoDiff-D3PM models use a 1D sinusoidal encoding (79) to denote the timestep for each input. All sequence models were trained with the Adam optimizer (80), a learning rate of 1e-4 with linear warmup over 16,000 steps, and dynamic batching that adaptively selects the number of samples to maximize GPU utilization. EvoDiff's small sequence models implement a ByteNet-style architecture with ca. 38M parameters. Large models were scaled to a ByteNet architecture of ca. 640M parameters by increasing the model dimension $d$ from 1020 to 1280, increasing the encoder hidden dimension from $d/2$ to $d$, and increasing the number of layers from 16 to 56.

38M parameter models were trained on 8 32GB NVIDIA V100 GPUs; 640M parameter models were trained on 32 (2×16) 32GB NVIDIA V100 GPUs. The maximum number of tokens per GPU in each batch was reduced from 40,000 to 6,000 to accommodate training the larger 640M parameter models. 38M parameter models were trained for approximately 2 weeks and saw ca. 3e14 tokens over 700,000 training steps. 640M parameter models were trained to the extent permitted by available compute resources to maximize model performance. Models saw between ca. 1e10 and 1e17 tokens over ca. 400,000-2,000,000 training steps. The D3PM-BLOSUM model ceased to improve after approximately 12 days of training. The D3PM-Uniform and OADM models were trained for 23 days without reaching convergence.

For MSAs, the EvoDiff denoising model adopts a 100M parameter MSA Transformer architecture (28). As with the single sequence models, EvoDiff-MSA-OADM models implicitly encode the diffusion timestep; EvoDiff-MSA-D3PM models include an additional sinusoidal timestep embedding. All MSA models were trained with the Adam optimizer with a learning rate of 1e-4 and linear warmup over 15,000 steps. EvoDiff MSA models were trained on 16 32GB NVIDIA V100 GPUs for 10 days and saw ca. 3e9 tokens over 55,000 training steps.

## Baseline models

To enable direct comparison, the left-to-right autoregressive (LRAR) and CARP baselines were trained with the same CNN architectures on the same dataset as EvoDiff sequence models. For LRAR, the convolution modules have a causal mask to prevent information leakage. For additional MLM baselines, sequences were sampled from the protein MLMs ESM-1b (68) and ESM-2 (30), which were trained on different releases of UniRef50. ESM-1b and ESM-2 both generated many "unknown" amino acids (X); performance was improved results by manually setting the logits for

X to `-inf`. Sequences were sampled from MLMs by treating the MLM as an OADMs and beginning from an all-mask state. For the structure-based diffusion baselines, sequences were obtained from FoldingDiff (8) and RFdiffusion (10) by first unconditionally generating structures and then using ESM-IF (33) to design their sequences. For MSA baselines, new query sequences were generated from ESM-MSA (28) by treating it as an OADM and sampling from an all-mask starting query sequence. CCMgen (39) with default parameters was used to train and generate from Potts models of validation MSAs from OpenFold.

## Computation of test-set perplexities

Perplexity was calculated by uniformly sampling a timestep for each test sequence, corrupting the sequence according to each diffusion model, predicting the sequence $x_0$ at $t = 0$ by passing inputs once through each trained model, and then computing the perplexity. For D3PM models, the perplexity is:

$$\mathrm{Perp}_{\mathrm{D3PM}} = \mathbb{E}_{t\sim\mathcal{U}(1,\ldots,T)} \exp\left(-\frac{1}{L}\sum_{t}^{L}\log p_\theta\left(x_0 \mid x_t\right)\right) \tag{11}$$

For OADMs, the perplexity is:

$$\mathrm{Perp}_{\mathrm{OADM}} = \mathbb{E}_{t\sim\mathcal{U}(1,\ldots,L)}\mathbb{E}_{\sigma\sim\mathcal{U}(S_D)} \exp\left[\frac{-1}{L-t+1}\sum_{k\in\sigma(\geq t)}\log p\left(x_k \mid \boldsymbol{x}_{\sigma(<t)}\right)\right] \tag{12}$$

To enable model comparison, perplexities for MLMs (CARP, ESM-1b, ESM-2) were computed as if they are OADMs.

And for LRAR models, the perplexity is:

$$\mathrm{Perp}_{\mathrm{LRAR}} = \exp\left[-\frac{1}{L}\sum_{t=1}^{L}\log p\left(x_t \mid \boldsymbol{x}_{<t}\right)\right] \tag{13}$$

Calculated D3PM perplexities were on average higher as $t \to T$ and lower as $t \to 1$, and masked perplexities were similarly higher for a greater number of masked tokens per sequence, i.e., as $t \to L_{masked}$ (Fig. S1, S2). Lower perplexities indicated improved performance and generalization capacity.

## Evaluation of structural plausibility

The structural plausibility pipeline (Fig. 2A) evaluates both the foldability and self-consistency of a given sequence. Foldability was evaluated by averaging the per-residue confidence score, reported as pLDDT by OmegaFold, across the entire sequence. Sequence self-consistency, denoted scPerplexity, describes how likely the generated sequence is to correspond to the predicted structure. Self-consistency was measured by taking structures predicted for a sequence from OmegaFold, running them through ESM-IF, and calculating the perplexity between the ESM-IF predicted-sequence and the original generated sequence.

The novelty of generated sequences was evaluated relative to training data seen by the model, by computing the Hamming distance between each generated sequence and every trainingset sequence of the same sequence length. The Hamming distance reported here is the distance between two sequences of equal length, normalized by sequence length. A distance of 1 indicates that there are no shared residues between two sequences, while a distance of 0 indicates that all the residues are shared between two sequences. The minimum of these Hamming distances, representing the closest sequence seen by the model during training, was reported for each sequence.

## BLAST protein homology search

Statistics on homology to natural sequences for all experimentally tested designs are provided in Supplementary Table 1. Each sequence was evaluated using the `blastp` tool in BLAST (81), comparing each protein sequence against BLAST's non-redundant protein database, which includes GenBankCDS translations, RefSeq, PDB, Swissprot, PIR, and PRF. The % query coverage and % sequence identity to the top hit identified are reported. In some cases, there are no homologous sequences identified; these are denoted as having "no blast hits".

## Computation of functional and structural features

To evaluate sequence coverage, ProtT5 embeddings were computed for each of 1,000 generated protein sequences and 10,000 sequences sampled from the test set using the Tools from Protein Prediction for Interpretation of Hallucinated Proteins (PPIHP) package (82). The resulting distributions of sequence embeddings (i.e., representing the corresponding distributions of sequences) were compared via the Fréchet ProtT5 distance (FPD),

$$\mathrm{FPD} = \|\mu_{\mathrm{test}} - \mu_{\mathrm{gen}}\|^2 + \mathrm{Tr}\left(C_{\mathrm{test}} + C_{\mathrm{gen}} - 2\sqrt{C_{\mathrm{test}}\, C_{\mathrm{gen}}}\right) \tag{14}$$

where, given the embedding space feature vectors for the test and generated distributions, $\mu$ is the feature-wise mean for each set of sequences, $C$ is the respective covariance matrix, and Tr refers to the trace linear algebra operation, defined as the sum of the elements along the main diagonal of a square matrix. FPD quantifies the distance between the two distributions of ProtT5 embeddings and is a computationally tractable approximation of the Wasserstein distance between them. Embeddings were visualized in 2D via uniform manifold approximation and projection (UMAP), fit to the test data and with `n neighbors`=25. The number of neighbors hyperparameter was selected to favor local similarities in place of global ones, in order to appropriately visualize the corresponding differences in embedding space and FPD measured for each model.

Structural features of generated sequences were evaluated via the ProtTrans (37) CNN predictor model to assign a 3-state secondary structure definition from DSSP (helix, strand, or other) to each residue in a protein. The fraction of predicted 'helix', 'strand', or 'other' was computed (the three values sum to 1 per sequence). The resulting multivariate distributions of secondary structure features (computed over 1000 generated or natural sequences) were visualized via kernel density estimation. The KL divergence between the mean values across the 3-state predictions for the generated and test sets was used to quantitatively evaluate the distribution of secondary-structures assigned for each model.

## Evolution-guided generation with EvoDiff-MSA

Starting with either a random or Max-Hamming subsampled MSA, new query sequences were generated by sampling from an allmask starting query sequence. The generated query sequence was evaluated relative to the corresponding original query sequence using the same tools and workflow described in *Evaluation of structural plausibility*. Each generated sequence was additionally evaluated for similarity relative to its reference MSA, which is comprised of a query sequence and alignment sequences. The % similarity of each generated sequence relative to its parent MSA was computed as the maximum % similarity over all sequences in the original MSA. Specifically, for a pair of sequences, the % similarity was computed by calculating the number of shared residue identities (accounting for both amino-acid identity and position index in the sequence), and for a given generated sequence the maximum value of these % similarities was determined. Across generated sequences both the CDF and mean of maximum % similarity were reported. Generated sequences were additionally evaluated for structural similarity relative to their original query sequences. Structures were predicted for each of the generated query

sequences and the original query sequences using OmegaFold. Structural similarity was measured via the template modeling score (TM-score) (83) for the two predicted structures following structural alignment:

$$\text{TM-score} = \max\left[\frac{1}{L_{\text{gen}}}\sum_{i}^{L_{\text{common}}}\frac{1}{1+\left(\frac{d_i}{d_0(L_{\text{true}})}\right)^2}\right] \quad (15)$$

where $L_{\text{gen}}$ is the length of the generated query sequence; $L_{\text{common}}$ is the number of shared residues; $d_i$ is the distance between the $i^{th}$ pair of residues; $L_{\text{true}}$ is the length of the true query sequence; and $d_0\left(L_{\text{true}}\right) = 1.24\sqrt[3]{L_{\text{true}}} - 15 - 1.8$ is a distance scale for normalization.

## Overview of protein expression, purification, and functional assays

Designs were expressed in mammalian (Chinese Hamster Ovary, CHO), bacterial (*Escherichia coli*, *E. coli*), yeast (*Saccharomyces cerevisiae*, *S. cerevisiae*), or cell-free systems; a mapping of design to expression system is provided in Supplementary Table 1. Unconditional generations were expressed, purified, and assayed by WuXi Biologics. Designed N-terminal IDRs for Cox15 were tested via cell-based assays in *S. cerevisiae*. Designs scaffolding the calmodulin $Ca^{2+}$-binding motif were expressed, purified, and assayed by WuXi Biologics. Designs scaffolding the p53 transactivation domain were expressed and assayed by Adaptyv Bio. All experimental details are provided in the corresponding Methods sections.

## Mammalian cell expression and downstream purification of unconditional generations and calmodulin scaffolds

A set of unconditional generations and scaffolds of the calmodulin $Ca^{2+}$-binding motif were expressed by WuXi Biologics in CHO cell expression systems (Supplementary Table 1). The WuXi Biologics Proprietary CHO K1 cell line was adapted from CHO-K1 cells (ATCC) to grow as suspension culture in WuXi's proprietary culture and feeding media (WuXi Biologics). Briefly, proteins were initially tested by WuXi's Ultra 96 platform, which allows for the expression of a greater number of proteins at a lower scale. Designs that expressed successfully were then produced by WuXi's WuXian transient platform in CHO K1 cells (WuXi Biologics) scaled up to 0.2-1.0L scale; purification followed successful expression at scale. When expressed successfully in CHO cells, protein samples used in downstream assays were derived from scaled expression in CHO, though certain designs also expressed in *E. coli* (Supplementary Table 1). Brief details on scaled expression in CHO cells and subsequent purification are provided below.

DNA sequences of target proteins were codon optimized and inserted into a proprietary backbone plasmid (WuXi Biologics). Synthesized constructs contained a CMV promoter inframe to express and secrete target proteins into the cell culture media. Constructed vectors were transformed into chemically competent *E. coli* TOP10 cells following manufacturer's instruction, and plasmids were isolated from successful colonies and then sequence confirmed. DNA was transfected into CHO K1 host cells, and transfected cultures were incubated by shaking at 150 rpm for 7 days.

Cell culture supernatants were harvested by primary centrifugation at 10,000 x g for 40 min, followed by sterile filtration. Supernatants were purified by affinity capture chromatography with Ni-Excel resin (Cytiva), with step imidazole elution, conducted on the AKTA Pure 25 column chromatography system (Cytiva). Following affinity capture chromatography, samples were further purified via size exclusion chromatography on Superdex 75 pg resin (Cytiva) in PBS. Relevant fractions were pooled, and final products were subject to quality control testing, including via SDS-PAGE, HPLC (Vanquish Duo UHPLC; ThermoFisher Scientific) and LC-MS (6545XT AdvanceBio LC/Q-TOF; Agilent). Final products were stored at -80C.

## *E. coli* expression of unconditional generations

For *E. coli* expression of unconditional generations (EvoDiff-Seq-Design-2; EvoDiff-Seq-Design-3), assembly with NcoI and XhoI was used to insert constructs (Genewiz) into the pET28a vector containing 6xHis tags. Plasmids were transformed into *E. coli* BL21(DE3). Large-scale cultures were grown at 5L (EvoDiff-Seq-Design-2; EvoDiff-Seq-Design-3) scale in auto-induction medium (84) for 2.5hrs at 37C pre-induction and 20hrs at 20C for induction. Though EvoDiff-Seq-Design-1 also successfully expressed in *E. coli* (Supplementary Table 1), its scaled-up expression occurred in CHO, followed by downstream purification as described above.

Following growth, cultures were harvested via centrifugation at 8,000 x g for 20 min at 4C. Pellets were suspended at 5 mL / 1 g pellet in lysis buffer (for EvoDiff-Seq-Design-2, EvoDiff-Seq-Design-3: 20 mM Tris, 300 mM NaCl, 20 mM Imidazole, 1 mM TCEP, 10% v/v glycerol, 0.1% v/v Triton X-100, with EDTA-free protease inhibitor cocktail, pH 8.0; for 1PRW and EvoDiff-Seq-1PRW-Design-5: 50 mM Tris, 150 mM NaCl, 1 mM TCEP, 10% v/v glycerol, with EDTA-free protease inhibitor cocktail, pH 8.0) and incubated at 4C for 1hr with stirring. High-pressure homogenization was used to lyse the cells (3x homogenization at 1000 bar), and the material was pelleted via centrifugation at 40,000 x g for 60 min at 4C. For EvoDiff-Seq-Design-2 and EvoDiff-Seq-Design-3, pellets were dissolved in a second lysis buffer (5 mL / 1 g pellet; 20 mM Tris, 300 mM NaCl, 20 mM Imidazole, 1 mM TCEP, 3 M Urea, 10% v/v glycerol, 0.1% v/v Triton X-100, pH 8.0), incubated at 4C for 1hr with stirring, then homogenized (2x homogenization at 1000 bar). Following centrifugation, the clarified supernatant was collected.

## Purification of *E. coli*-expressed unconditional generations

Following centrifugal lysate clarification, proteins were first purified by affinity chromatography at 4C. Clarified supernatant was equilibrated into equilibration buffer (EvoDiff-Seq-Design-2, EvoDiff-Seq-Design-3: 20 mM Tris/HCl, 300 mM NaCl, 20 mM Imidazole, 0.5 mM TCEP, 3 M Urea, 10% v/v glycerol, pH 8.0), loaded onto Ni-NTA 3 mL affinity chromatography matrix (Ni-NTA Superflow, Qiagen), washed once with equilibration buffer, washed twice with Ni-NTA wash buffer (20 mM Tris/HCl, 300 mM NaCl, 20 mM Imidazole, 0.5 mM TCEP, 10% v/v glycerol, pH 8.0), and eluted with Ni-NTA elution buffer (20 mM Tris/HCl, 300 mM NaCl, 250 mM Imidazole, 0.5 mM TCEP, 10% v/v glycerol, pH 8.0).

Following affinity chromatography, purification of *E. coli*-expressed unconditional generations proceeded as follows. Throughout, intermediate samples were verified by SDS-PAGE. The purity and molecular weight of final samples were confirmed by liquid chromatography-mass spectrometry (LC-MS; Agilent).

For EvoDiff-Seq-Design-2 and EvoDiff-Seq-Design-3, following affinity chromatography, samples were further purified by Q-HP anion exchange chromatography (HiTrap Q HP anion exchange chromatography column, Cytiva) at 4C. EvoDiff-Seq-Design-2 and EvoDiff-Seq-Design-3 samples were diluted 6x in dilution buffer (20 mM Tris/HCl, 0.5 mM TCEP, 3 M Urea, 10% v/v glycerol, pH 8.0), equilibriated, loaded onto the column, washed with equilibriation buffer (20 mM Tris/HCl, 50 mM NaCl, 0.5 mM TCEP, 3 M Urea, 10% v/v glycerol, pH 8.0), and eluted (20 mM Tris/HCl, 1 M NaCl, 0.5 mM TCEP, 3 M Urea, 10% v/v glycerol, pH 8.0) using a concentration gradient of 0-60%, 60-100% for elution.

For EvoDiff-Seq-Design-2, following Q-HP purification, monomeric fractions were pooled, and samples were buffer exchanged overnight at 4C into a final buffer (50 mM Tris/HCl, 300 mM NaCl, 0.5 mM TCEP, pH 7.5) and stored.

For EvoDiff-Seq-Design-3, following Q-HP purification, the protein was further purified at scale-up via reverse-phase (RP) chromatography. For RP chromatography, pooled protein from Q HP purification was adjusted to pH of 4.0 with acetic acid in a final solution with 5% v/v acetonitrile. The sample was loaded onto a Sepax Bio-C8 column (Sepax Technologies), washed with a solution of 0.1% v/v TFA, 5% v/v acetonitrile in ddH2O, and then eluted (0.1% v/v TFA in acetonitrile) using a concentration gradient of 5-100%, with sample eluting at a concentration of ca. 50% elution buffer. The final sample was frozen in liquid nitrogen and freeze-dried in a vacuum dryer.

## Structural characterization by UV circular dichroism

Circular dichroism (CD) was used to characterize higher-order structures in EvoDiff-designed proteins. All CD data collection was performed with the Chirascan V100 CD spectrometer (Applied Photophysics), and analysis was conducted via the Chirascan software (Applied Photophysics). Briefly, secondary structure characteristics were determined via measurement of the far-UV CD spectra. Prior to measurement, protein samples were concentrated and then diluted with ultrapure water to a final concentration of 0.1 mg/mL. Measurements were acquired in the far-UV spectra ranging from 190 to 260nm; all reported measurements were acquired within the linear range of the instrument. Analysis of secondary structural composition was determined using the CDNN software with a net setting using 33 basespectra. CD experiments were conducted by WuXi Biologics.

## Generation of intrinsically disordered regions (IDRs)

IDR generation and analysis leveraged a publicly available dataset of 15,996 human IDRs and their orthologs (48). This dataset was generated by running SPOT-Disorder v1 (85) on the human proteome and applying the predicted IDR positions to an MSA of likely-similar-function orthologs (determined using an evolutionary distance heuristic), curated from the larger set of orthologs contained in the OMA database (86). The resulting dataset only contained IDRs, not full protein sequences, and thus IDR sequences were mapped back to the MSAs of full protein sequences in OMA in order to provide context about the sequence regions surrounding the IDRs.

For input to EvoDiff models, the full sequence of an IDR-containing human protein was treated as the query sequence, and a corresponding MSA was constructed by subsampling 63 other sequences from all the query's orthologs. All sequences were subsampled to 512 residues in length, with the following criteria maintained. Subsampling criteria were that the subsampled query sequence contain at least 1 IDR, and that the total IDR region was less than half the total length of the subsampled sequence ($L_{\text{IDR}} \leq 256$). For IDR generation from EvoDiff-Seq, the query sequence with the IDR region masked was provided as the only input to EvoDiff-Seq, which then generated new residues for the masked region (i.e., the region corresponding to the true IDR). For IDR generation from EvoDiff-MSA, the query sequence with the IDR region masked, aligned to the rest of the MSA, was provided as input to EvoDiff-MSA, which then generated new residues for the masked region.

The resulting generations, containing putative IDRs, were input to DR-BERT, a protein language model fine-tuned for disorder prediction (49), to obtain per-residue disorder scores ranging from 0-1 (less to more disordered). A single-sequence IDR predictor (DR-BERT) was used in place of MSA-based IDR scoring methods, because of an observed bias towards higher disorder scores with MSA-based methods – e.g., random uniform sampling of residues in the masked query positions still resulted in a prediction of disorder given the presence of the orthologs in the alignment. Disorder scores for true IDRs, generated IDRs, scrambled IDRs, and randomly generated IDRs were computed to evaluate the performance of DR-BERT predictions. The randomly-sampled baseline was constructed by randomly sampling amino acids over an IDR region; the scrambled baseline was constructed by shuffling the existing amino acids over an IDR region into a scrambled permutation. In all cases (true IDRs, generated IDRs, scrambled and random baselines), the entire protein sequence was input to DR-BERT for scoring. Since DR-BERT is for single-sequences, for putative IDRs generated by EvoDiff-MSA, the entire query sequence was inputted into DR-BERT, with < GAP > tokens eliminated, to obtain per-residue disorder scores. Lastly, a direct comparison between the original IDR and the generated putative IDR was conducted by calculating the % sequence similarity between the fraction of shared residues between the two IDR regions.

## Design of N-terminal IDRs for Cox15

The yeast Heme A synthase Cox15 was used as a target for the design of synthetic N-terminal IDRs. A total of 600 inpainted designs for the disordered N-terminal IDR were generated using EvoDiff-Seq and EvoDiff-MSA (Rand subsampling). The first 200 designs were generated by masking the N-

terminal IDR (amino acid positions 1 to 45, inclusive) and randomly decoding at each position in the sequence with EvoDiff-Seq. Another 200 sequences were generated by iteratively re-masking and generating at each position of the IDR in order from left to right. Both design strategies are equally represented in the final tested candidates. The last 200 designs were generated using EvoDiff-MSA, using the alignment created for Cox15 from Lu et al. (48). The yeast Cox15 sequence was set as the query sequence, and an alignment of 63 sequences was randomly subsampled. For each generation, a start token within the Cox15 query was randomly sampled; this results in varying N-terminal inpainting lengths for each independent generation. The masked N-terminal region was then decoded in random order. To nominate 8 designs for experimental validation, all 600 designs were scored for and then ranked by the probability of disorder by DR-BERT and the likelihood of mitochondrial localization by MitoFates (Fig. S14). For each of EvoDiff-Seq and EvoDiff-MSA, the top four designs with the largest edit distance to the native IDR N-terminal sequence of Cox15 were selected for experimental validation (Supplementary Table 1).

## Yeast experimental methods for design of synthetic IDRs

Synthetic IDRs designed to be mitochondrial targeting signals for the yeast protein Cox15 were tested *in vivo* via a subcellular localization assay, readout by confocal microscopy. A tagged Cox15-mScarlet serves to show wild-type localization of the native protein and provides an internal control for direct comparison to the synthetic IDRs tested in this study. This control strain was constructed by tagging the endogenous locus of Cox15 at the C-terminus with mScarlet followed by a selectable resistance marker (KANMX4) via homologous recombination in *S. cerevisiae* S288C derivative (BY4741: MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) (87), using the standard LiAc method (88), and isolated by marker selection on yeast synthetic minimal media (SD). The corresponding strain with the mScarlet-tagged Cox15 full-length ORF served as the parental strain for all subsequent strains in this study; the mScarlet-tagged Cox15 full-length ORF remained intact in all experimental strains.

The experimental Cox15 strains, containing wild type or synthetic IDR sequences, were made by inserting a second copy of the Cox15 ORF tagged with SuperFold-GFP (sfGFP) (89) at the HO locus of the Cox15-mScarlet control strain via recombinant homologous transformation as described above. Synthetic amino acid sequences were obtained via computational design, and corresponding DNA sequences were codon optimized for optimal expression in *S. cerevisiae* using a program available from Integrated DNA Technologies (*https://www.idtdna.com/CodonOpt*). The resultant nucleotide sequences were then used to design gene fragments (GeneArt Strings). These gene fragments served as PCR templates to generate sequences that were then inserted in plasmids via Gibson Assembly (90), replacing the native Cox15 IDR in the full length Cox15 ORF, transformed into *E. coli* (DH5α), and then isolated by marker selection (Ampicillin). All plasmids were purified from bacteria and confirmed by Sanger sequencing (Eurofins Scientific SE). Positive clones were subsequently used for the yeast genomic integration. The transformed linear DNA fragments contained the native Cox15 promoter, the Cox15 ORF (containing the synthetic IDRs) fused to super-fold GFP (sfGFP), followed by the tADH1 terminator and an auxotrophic yeast marker (URA3) for selection. For the IDR knockout strain, a previously constructed strain containing an endogenous Cox15 2-45 deletion was used as a PCR template to generate the IDR deletion. Following transformation and cloning, positive clones were selected for yeast genomic integration, and were integrated into the same locus (HO) as all other strains via direct homologous recombination in yeast cells. This was a two fragment integration, with the second fragment providing the remaining Cox15 ORF fused to sfGFP and the tAdh1 terminator sequence, followed by the same auxotrophic marker used in the other strains. The HO locus was amplified from all yeast strains via colony PCR, and strain identity was confirmed by Sanger sequencing (Eurofins Scientific SE).

## Confocal microscopy and quantification of subcellular localization

Confocal images of yeast strains were obtained as follows. Control and mutant strains were inoculated into yeast synthetic minimal media (SD) and grown overnight at 30C. Prior to imaging, stationary cultures were diluted 1/10 in fresh media and grown at 30C for 4 hours to ensure log-phase growth and proper expression of mScarlet/GFP fusion products. Live cells were concentrated 50-fold by centrifugation (30s at 3,500 x g) and imaged using a Leica TCS SP8 confocal microscope with a 63X oil-immersion objective and a 10X eye-piece objective. mScarlet excitation was at 587nm and GFP excitation was at 488nm.

To quantify co-localization, confocal images were first segmented using the YeastSpotter segmentation model (91) applied on the brightfield image. All segmentation outputs under 2000 pixels were filtered to remove dust particles and small budding cells (which are typically out of focus). For each cell, co-localization was quantified by computing the Pearson correlation coefficient between the mScarlet channel (587nm excitation) and the GFP channel (488nm), for all pixels assigned to the cell by YeastSpotter.

## Motif scaffolding

*In-silico* scaffolding performance was evaluated on a recently published benchmark (10) of 25 scaffolding problems across 17 unique proteins.

In our scaffolding benchmark, each unique protein was treated as 1 example, for a total of 17 unique scaffolding examples. 100 samples were generated for each unique scaffolding example. For proteins 6E6R, 6EXZ, 7MRX, and 5TRV, which were the 4 examples evaluated at 3 different scaffolding lengths in RFdiffusion (10), the number of successes across these three different scaffolding lengths were averaged to facilitate comparisons between RFDiffusion and EvoDiff.

To generate a scaffold with EvoDiff-Seq, a scaffold length between 50-100 residues (exclusive of the motif) was sampled uniformly; the motif was placed randomly within the length; and scaffold residues were generated from EvoDiff-Seq conditioned on the provided motif residues. In this approach, on average, protein sequences generated by EvoDiff-Seq were longer (between 45 and 194 residues in length) than those inverse-folded from structures generated by RFdiffusion, which range from 30-152 residues in total length inclusive of the length of the motif.

For scaffolding with EvoDiff-MSA, MSAs for each sequence corresponding to the original PDB structure were generated using the tools from AlphaFold (92) and then subsampled to 64 sequences, and a maximum of 150 residues in length, where the original sequence obtained from the PDB crystal structure was assigned as the query sequence. In cases where the scaffolding examples were shorter than 150 residues, sequences were padded with a < GAP > token, to allow EvoDiff to generate longer-scaffolds. Sequences generated by EvoDiff-MSA were between 56 and 150 residues in length, inclusive of the motif and scafffold. For each scaffolding example, a common set of 100 subsampled MSAs, where 50 were randomly subsampled and 50 were subsampled via MaxHamming, was used commonly across EvoDiff-MSA (Max), EvoDiff-MSA (Random), and ESM-MSA. That is, an individual generation trial for each model corresponded to a unique MSA from the common set of 100 MSAs constructed for a scaffolding example. At inference time, all non-motif residues in the query sequence were masked, and new residues in these locations were generated by EvoDiff-MSA.

OmegaFold was used to predict structures corresponding to sequences generated by EvoDiff. A generation was counted as 'successful' if its predicted structure had a pLDDT >70 and a motifRMSD <1.0 $\AA$ relative to the original motif crystal structure. Note that these success criteria are cutoffs proposed by structure-based models (10) and adopted here to facilitate comparison. The motifRMSD was computed as the RMSD between the alpha-carbons of the motif in the original crystal structure and the predicted structure for the scaffolded motif.

## Design of experimentally-tested calmodulin scaffolds

EvoDiff-Seq, EvoDiff-MSA (Rand subsampling), and EvoDiff-MSA (Max subsampling) were used to design 100 calmodulin scaffolds each, resulting in 300 designs total. Generations from EvoDiff-Seq began by isolating the residues of the functional motif from the sequence of bovine calmodulin from the PDB entry 1PRW. The subsequence of residues 16-70 (inclusive), which contains the functional $Ca^{2+}$-binding motif, was considered, and residues 35-50 (inclusive) were masked, such that only the residues of the functional motif remained. Then, for each EvoDiff-Seq generation, a total scaffold length between 10 and 100 residues was sampled and set to all mask tokens, with the start position of the motif positioned randomly within the scaffold. A sequence was generated from EvoDiff-Seq by randomly decoding at each position. For generation with EvoDiff-MSA, an alignment was constructed using the homology tools from AlphaFold. 63 sequences were subsampled from the alignment using Rand or Max subsampling, using a maximum sequence length of 150. All residues that did not belong to a motif were then masked, followed by decoding at each position randomly. To select designs for experimental validation, all 300 designs were compiled and filtered by the designability metrics of OmegaFold pLDDT > 70 and motifRMSD < 1.0Å. Designs that scored well by these metrics but seemed physically unfeasible (e.g., long unconstrained helices) were manually filtered out, and 13 designs were nominated for experimental validation (Supplementary Table 1 ).

## *E. coli* expression of calmodulin scaffolds

Except for the native bovine calmodulin (1PRW) and EvoDiff-Seq-1PRW-Design-5, all other calmodulin scaffolds successfully expressed in the scaled CHO expression system and were subsequently purified (see prior Methods section; Supplementary Table 1 ); these scaffolds also successfully expressed in *E. coli*, but expression was not scaled up, and thus, for these designs, purified samples derived from CHO expression were used in the downstream absorbance assay. Scaled expression of 1PRW and EvoDiff-Seq-1PRW-Design-5 occurred in *E. coli*.

For scaled expression of 1PRW and EvoDiff-Seq-1PRW-Design-5 in *E. coli*, assembly with NcoI and XhoI was used to insert constructs (Genewiz) into the pET28a vector containing 6xHis tags. Plasmids were transformed into *E. coli* BL21(DE3). Large-scale cultures were grown at 0.5L (EvoDiff-Seq-1PRW-Design-5) or 1L (1PRW) scale in auto-induction medium (84) for 2.5hrs at 37C pre-induction and 20hrs at 20C for induction.

Following growth, cultures were harvested via centrifugation at 8,000 x g for 20 min at 4C. Pellets were suspended at 5 mL / 1 g pellet in lysis buffer (for 1PRW and EvoDiff-Seq-1PRW-Design-5: 50 mM Tris, 150 mM NaCl, 1 mM TCEP, 10% v/v glycerol, with EDTA-free protease inhibitor cocktail, pH 8.0) and incubated at 4C for 1hr with stirring. High-pressure homogenization was used to lyse the cells (3x homogenization at 1000 bar), and the material was pelleted via centrifugation at 40,000 x g for 60 min at 4C. Following centrifugation, the clarified supernatant was collected.

## Purification of *E. coli*-expressed calmodulin scaffolds

Following centrifugal lysate clarification, proteins were purified by affinity chromatography at 4C. Clarified supernatant was equilibrated into equilibration buffer (for 1PRW, 1PRW 37: 20 mM Tris/HCl, 300 mM NaCl, 20 mM Imidazole, 0.5 mM TCEP, 10% v/v glycerol, pH 8.0), loaded onto Ni-NTA 3 mL affinity chromatography matrix (Ni-NTA Superflow, Qiagen), washed once with equilibration buffer, and eluted with Ni-NTA elution buffer (20 mM Tris/HCl, 300 mM NaCl, 250 mM Imidazole, 0.5 mM TCEP, 10% v/v glycerol, pH 8.0).

Following affinity chromatography, purification of *E. coli*-expressed 1PRW and EvoDiff-Seq-1PRW-Design-5 proceeded as follows. Throughout, intermediate samples were verified by SDS-PAGE. The purity and molecular weight of final samples were confirmed by liquid chromatography-mass spectrometry (LC-MS; Agilent).

For 1PRW, following affinity chromatography, the sample was subject to a digestion step with TEV protease (20:1 w/w ratio of 1PRW to TEV; TEV expressed and purified in-house by WuXi Biologics), in order to remove the His tag. Reverse Ni-NTA chromatography of the digested protein sample

was subsequently performed, with the sample equilibriated into equilibriation buffer (20 mM Tris/HCl, 300 mM NaCl, 20 mM Imidazole, 0.5 mM TCEP, 10% v/v glycerol, pH 8.0), loaded into the column, washed once with equilibriation buffer, and then eluted with Ni-NTA elution buffer. The sample was then subject to size exclusion chromatography (SEC) (HiLoad Superdex 75 Prep Grade; Cytiva; 50mM Tris/HCl, 0.3 M NaCl, 0.5 mM TCEP, 10% v/v glycerol, pH 7.5). Following SEC, the 1PRW sample was subject to Q-HP anion exchange chromatography (HiTrap Q HP anion exchange chromatography column, Cytiva) at 4C. Samples were diluted 6x in dilution buffer (20 mM Tris/HCl, 0.5 mM TCEP, 10% v/v glycerol, pH 7.5), equilibrated and loaded onto the column (20 mM Tris/HCl, 50 mM NaCl, 0.5 mM TCEP, 10% v/v glycerol, pH 7.5), washed with equilibration buffer, and eluted (20 mM Tris/HCl, 1 M NaCl, 0.5 mM TCEP, 10% v/v glycerol, pH 7.5) using a concentration gradient of 0-60%, 60-100% for elution. 1PRW monomeric fractions were pooled, and samples were buffer exchanged overnight at 4C into a final buffer (50 mM Tris/HCl, 300 mM NaCl, 0.5 mM TCEP, pH 7.5) and stored.

For EvoDiff-Seq-1PRW-Design-5, following affinity chromatography, the sample was purified via SEC (HiLoad Superdex 75 Prep Grade; Cytiva; 50mM Tris/HCl, 0.3 M NaCl, 0.5 mM TCEP, 10% v/v glycerol, pH 7.5). Monomeric fractions were pooled, and samples were stored.

## Spectroscopic analysis of $Ca^{2+}$ binding to calmodulin and designed scaffolds

Analysis of $Ca^{2+}$ binding to recombinantly-expressed bovine calmodulin (1PRW) and EvoDiff-designed scaffolds was performed via a spectroscopic assay adapted from previous reports (55, 56). Purified proteins were incubated at 50 $\mu$M final concentration with 32x molar excess $CaCl_2$ (1.6mM; MCE) for 16 hrs at room temperature in assay buffer (20mM HEPES, 100mM KCl, pH 7.8). Following the 16h incubation, absorbance spectra were collected at wavelengths from 450nm to 650nm on a SpectraMax M2e Microplate Reader (Molecular Devices). Measurements were background-corrected using readings from the buffer-only controls. Successful binders showed characteristic absorbance peaks at 500-550 nm relative to the absorbance spectra for no-protein and buffer controls.

## Design of experimentally-tested p53 scaffolds

To design scaffolds for the transactivation domain of p53, 900 designs from EvoDiff-Seq, 200 from EvoDiff-MSA (Rand subsampling), and 200 from EvoDiff-MSA (Max subsampling) were generated, resulting in 1300 total designs to select from. For EvoDiff-Seq, the p53 transactivation domain, residues 5-15 (inclusive) from the PDB entry 1YCR, was scaffolded at variable scaffold lengths between 50 and 100 residues. Generations were decoded by sampling the amino acid probabilities at each position with various temperatures ($T$). 500 examples were generated at $T$ = 1.0, 200 at $T$ = 0.8, and another 200 at $T$ = 0.7. For generation with EvoDiff-MSA, an alignment based on the full sequence of human p53 (Uniprot entry: P04637) was constructed using the homology tools from AlphaFold; the transactivation domain within 1YCR is from the human p53 sequence. 63 sequences containing the scaffolding motif were subsampled from the alignment using Rand or Max subsampling, using a maximum sequence length of 150. All residues that did not belong to the motif in the query sequence were then masked, followed by decoding at each position randomly without temperature sampling. The 900 generations from EvoDiff-Seq were rank ordered by the designability metrics of OmegaFold pLDDT and motifRMSD order and a set of 19 candidates was nominated (Supplementary Table 1). None of the designs from EvoDiff-MSA passed the *in-silico* pLDDT criterion (pLDDT >70), likely due to the native intrinsically disordered region within p53 (Fig. 6G). EvoDiff-MSA designs were thus filtered to those with motifRMSD <1.0 *Å*, and five sequences (3 Rand subsampling, 2 Max subsampling) were randomly selected such that three out of five sequences had shared homology with the native p53 sequence (Supplementary Table 1).

## Cell-free expression of p53 scaffolds

Designed sequences scaffolding the p53 transactivation domain were synthesized *in vitro* via a cell-free expression platform (Adaptyv Bio). DNA constructs encoding the ligands, each fused to a C-terminal assay tag, were designed by reverse-translating the target protein sequences. The sequences were optimized for manufacturability and yield using Adaptyv's proprietary algorithms to maximize expression efficiency. The optimized DNA constructs were ordered as gene fragments from Twist Bioscience.

Assembly of the constructs into an appropriate expression vector was performed using the NEBuilder HiFi DNA Assembly Kit (New England Biolabs) in 3 $\mu$L reactions. Protein expression was carried out in 8 $\mu$L reactions using an optimized prokaryotic in-vitro translation system. Reactions were incubated at 37C for 12 hours to ensure efficient protein synthesis. Post-expression, protein concentration and yield were normalized using an affinity-based quantification assay. All tested synthetic designs and all positive controls expressed successfully.

## Affinity characterization using biolayer interferometry

Designed p53 scaffolds and positive controls (i.e., the ligands) were tested for their *in vitro* binding affinity to MDM2 (i.e., the antigen). The p53 transactivation domain peptide from 1YCR (p53 control) and two synthetic binders from RFDiffusion (10) (p53 RF 53 and p53 RF 89) were used as positive controls. Sequences of all tested designs as well as of the target antigen are provided in Supplementary Table 1.

Multi-cycle kinetic assays were conducted in triplicate to evaluate ligand binding to the antigen using Biolayer Interferometry (BLI) (Gator Bio). Ligands, tagged with a twin strep tag, were immobilized on streptavidin-coated probes via tag-specific interactions. Antigen solutions at multiple concentrations (500 nM, 158.1 nM, 50 nM, 15.8 nM, and 0 nM) of full-size, recombinant human GST-MDM2 (Bio-Techne E3-202-050) were flowed over the probes, with signal acquisition at a sampling rate of 5 Hz. The assay sequence began with a 120-second baseline phase, followed by 120 seconds of ligand loading onto the probes. After loading, a secondary 200-second baseline was established to ensure probe stability. The association phase, during which the antigen was introduced, lasted 180 seconds, followed by a 240-second dissociation phase to monitor antigen unbinding from the immobilized ligand.

All experiments were conducted in kinetic buffer (PBS-T supplemented with 0.02% w/v BSA) at 25°C. Negative controls, including buffer-only samples, were included for baseline subtraction. Data were corrected based on the baseline values and fitted globally to a 1:1 binding model across all concentrations. Kinetic rate constants and dissociation constants ($K_D$) were derived from the fitted curves.

# Supplementary information

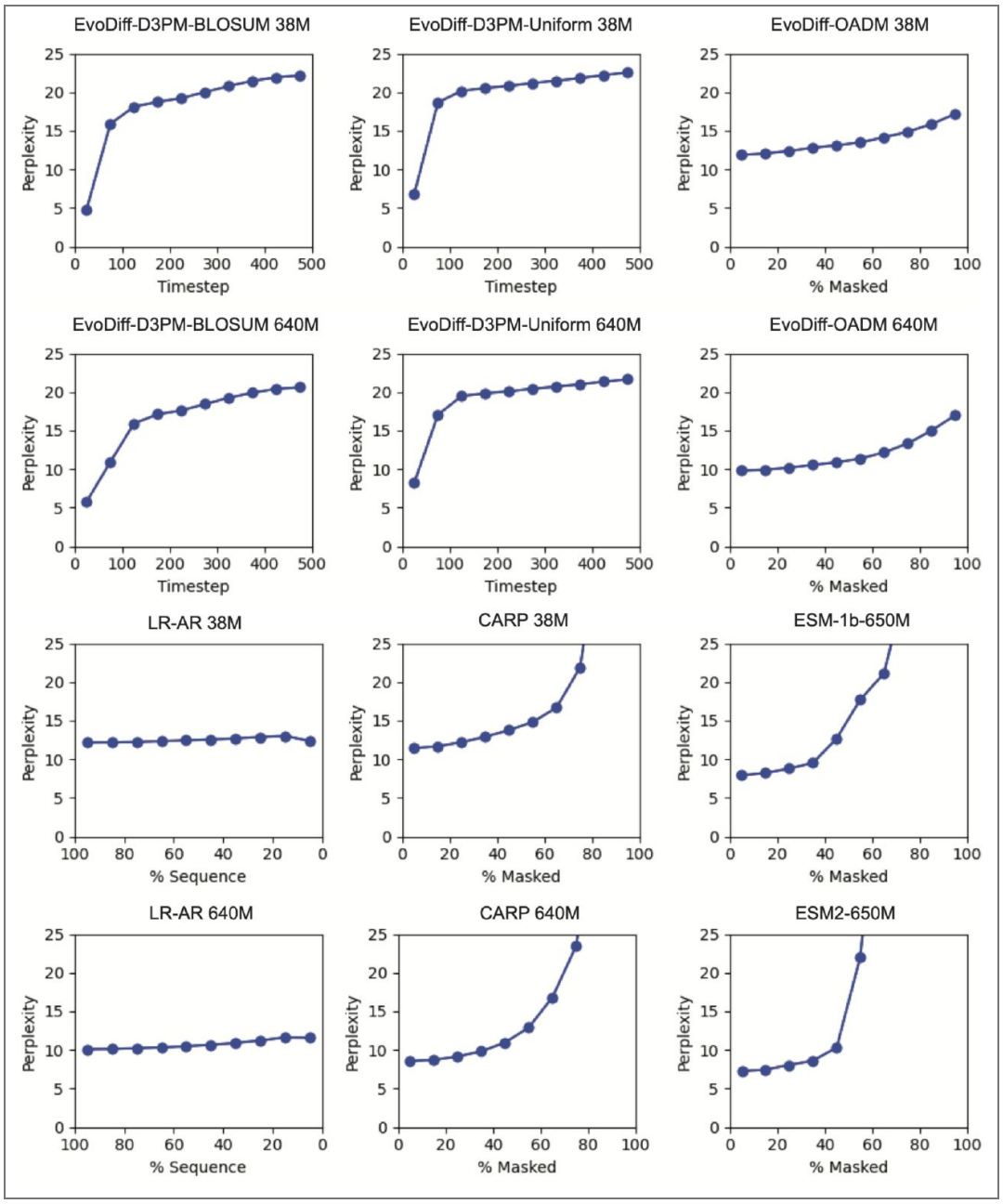



**Figure S1. Perplexity as a function of corruption step for EvoDiff sequence models.** Test-set perplexities at sampled intervals of the degree of corruption, specifically the diffusion timestep for D3PM models, the fraction of masked residues for OADM and masked language models, and the fraction of evaluated sequence for LRAR models. Intervals reflect evenly spaced windows of 50 timesteps for D3PM models or 10% masking for masked models.

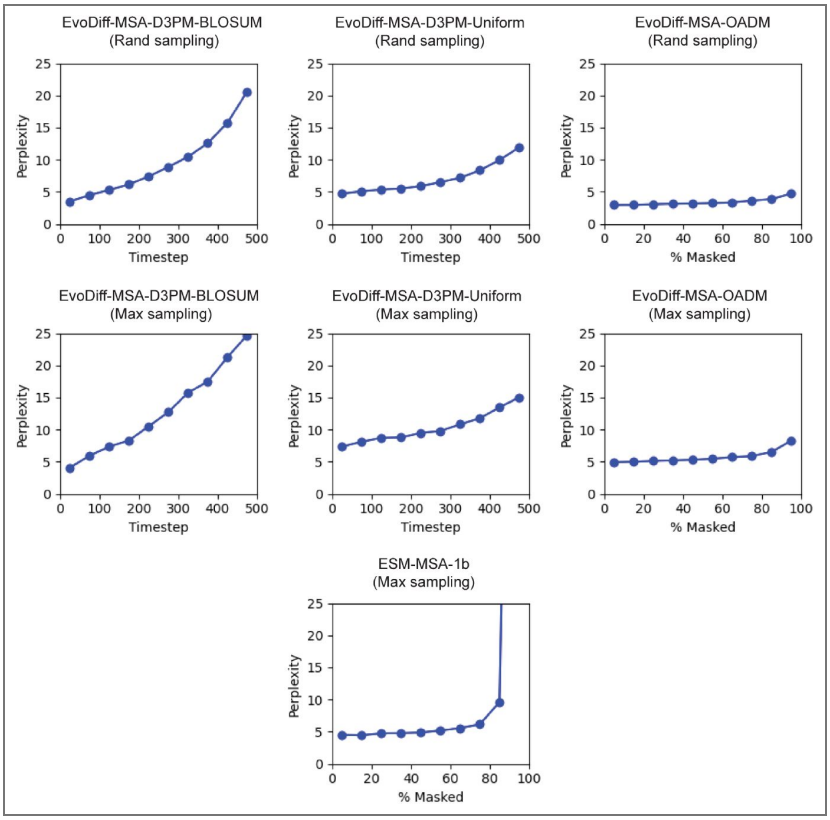



**Figure S2. Perplexity as a function of corruption step for EvoDiff MSA models.**

Test-set MSA perplexities at sampled intervals of the degree of corruption, specifically the diffusion timestep for D3PM models and the fraction of masked residues for OADM and ESM models. The test-set evaluated for each model was sampled using the same sampling scheme assigned during training. Intervals reflect evenly spaced windows of 50 timesteps for D3PM models or 10% masking for masked models.

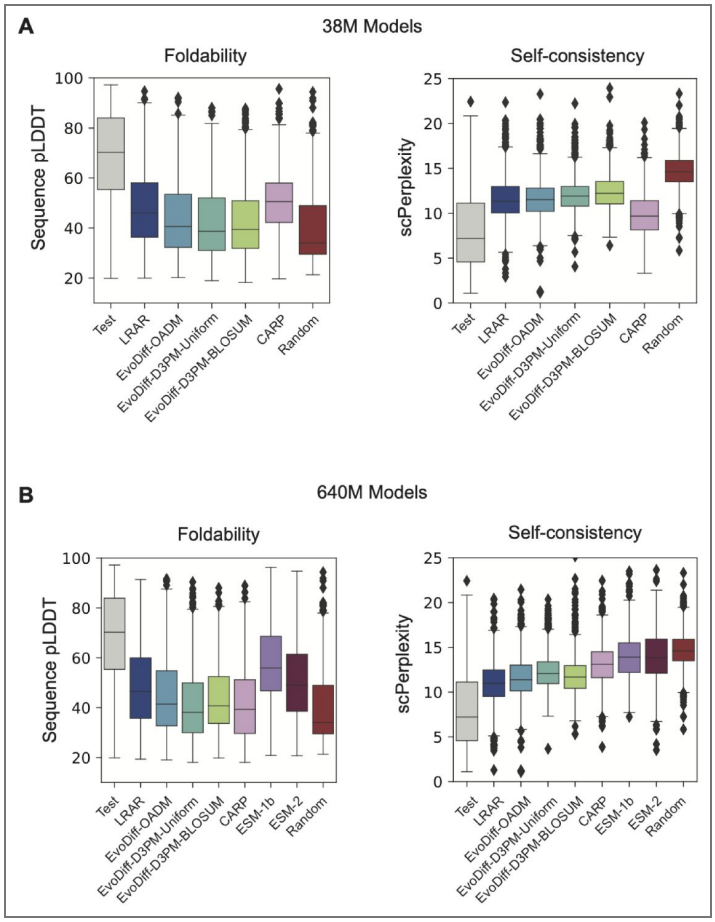



**Figure S3. Summary statistics for structural plausibility metrics for sequence models.**

**(A-B)** Distribution of pLDDT and scPerplexity metrics for sequences from the test set, 38M parameter EvoDiff and baseline models (A), and 640M parameter EvoDiff and baseline models (B) (n=l000 sequences per model). Test and Random baselines are reproduced in (A) and (B) for reference.

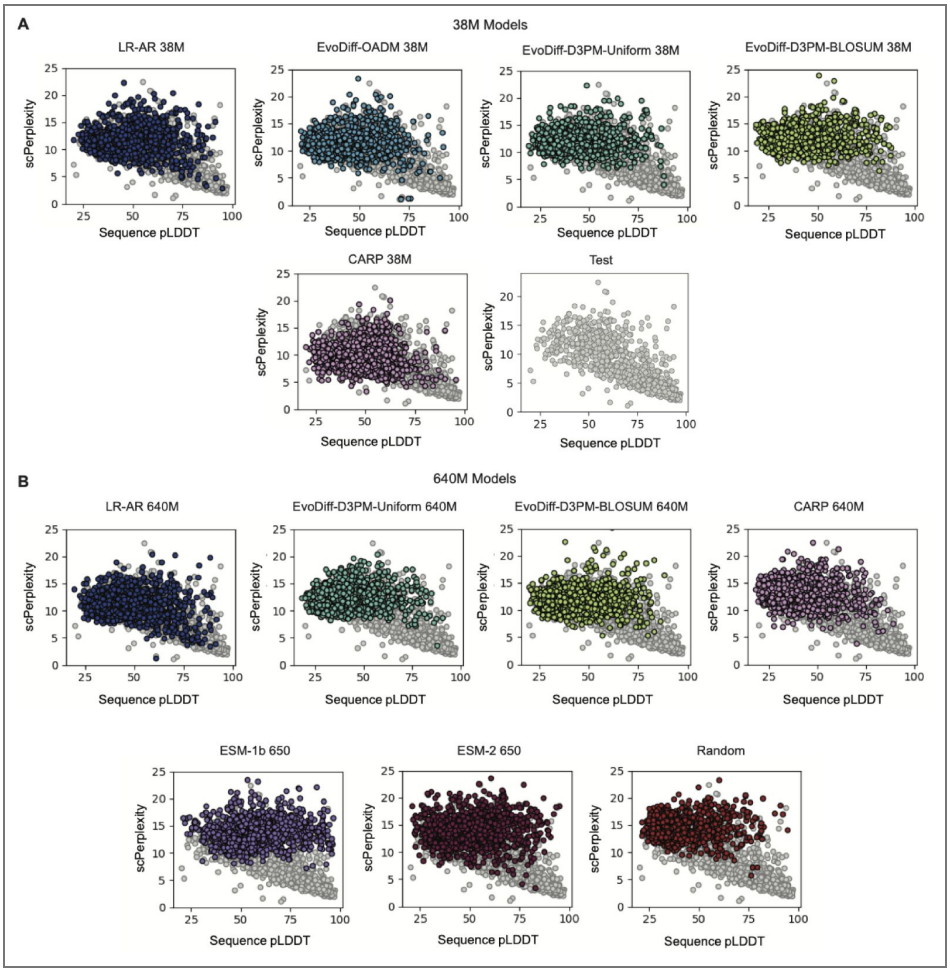



**Figure S4. Sequence pLDDT versus self-consistency perplexity for EvoDiff sequence models.**

**(A-B)** Results for sequences from 38M parameter EvoDiff, baseline models, and test data plotted alone (A), and 640M parameter EvoDiff and baseline models (B) (various colors, *n*=1000), except for EvoDiff-OADM-640M (EvoDiff-Seq, shown in Fig. 2C), relative to sequences from the test set (grey, *n*=1000). The self-consistency perplexity (ESM-IF Perplexity) is computed using sequences inverse-folded by ESM-IF.

**Figure S5. Representative proteins generated by EvoDiff-Seq.**

**(A)** Predicted structures and metrics for representative structurally plausible generations from EvoDiff-Seq, the 640M-parameter OADM model. **(B)** Corresponding per-residue pLDDT for pLDDT scores computed based on the OmegaFold predicted structures, for individual residues in representative high-fidelity generations from EvoDiff-Seq (Fig. 2D). Points are colored by pLDDT (0-100, red to blue).

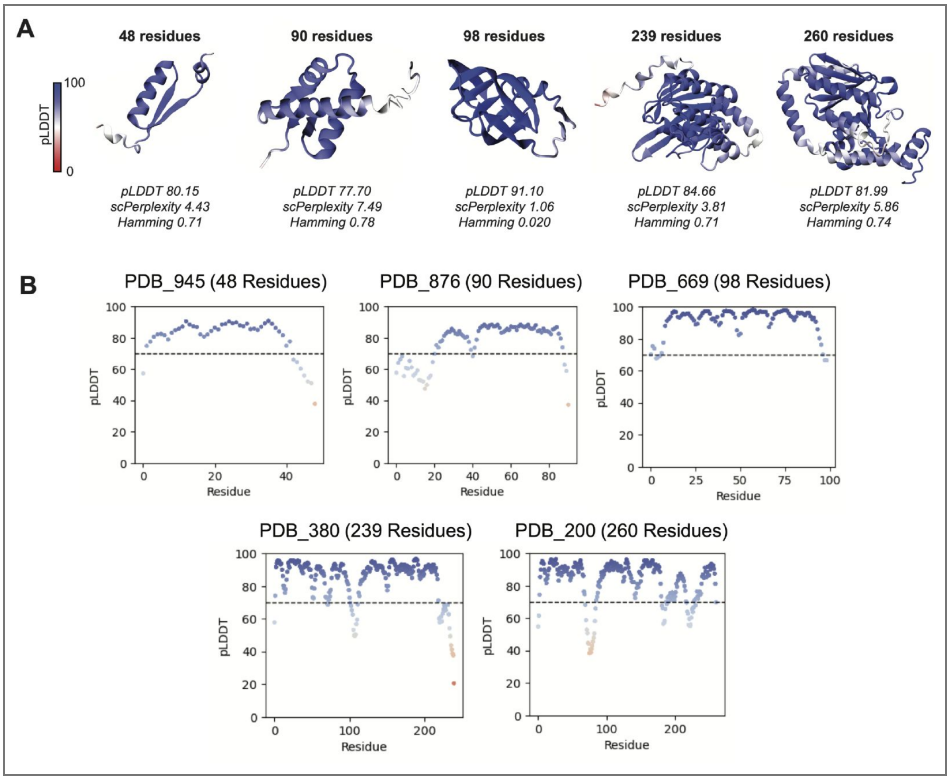
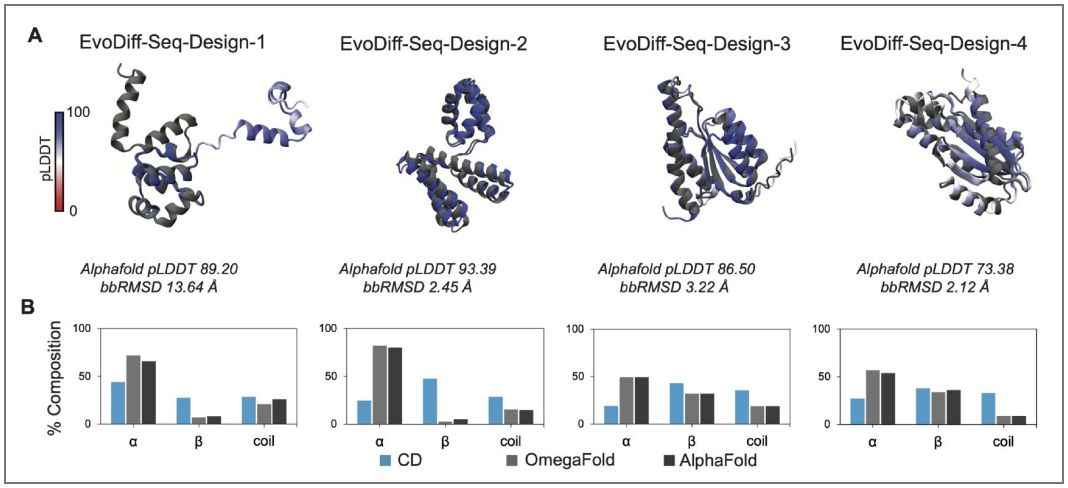



**Figure S6. Comparison of structure-prediction methods for unconditional EvoDiff-Seq designs**

**(A)** AlphaFold predictions (colored by pLDDT) overlaid on OmegaFold (grey). The AlphaFold pLDDT, and backbone RMSD (bbRMSD) taken between the two structures is denoted. **(B)** Percent composition extracted from CD spectra (blue) compared to structure prediction methods (AlphaFold and OmegaFold, grey).



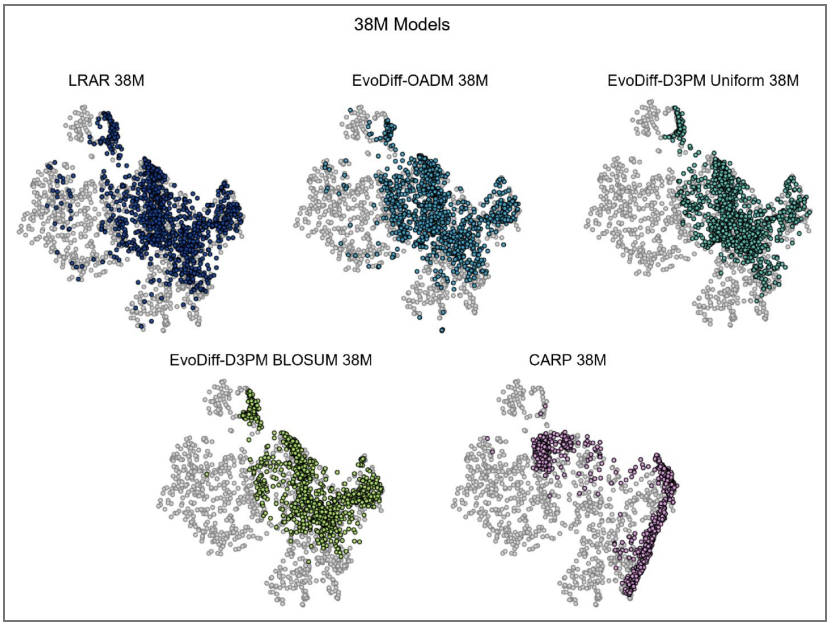



**Figure S7. Coverage of sequence and functional space for generated distributions from 38M parameter EvoDiff sequence models and baselines.**

UMAP of ProtT5 embeddings, annotated with FPD, of natural sequences from test set (grey, *n*= l000 plotted) and of generated sequences from EvoDiff 38M parameter models and baselines (various colors, *n*=l000).

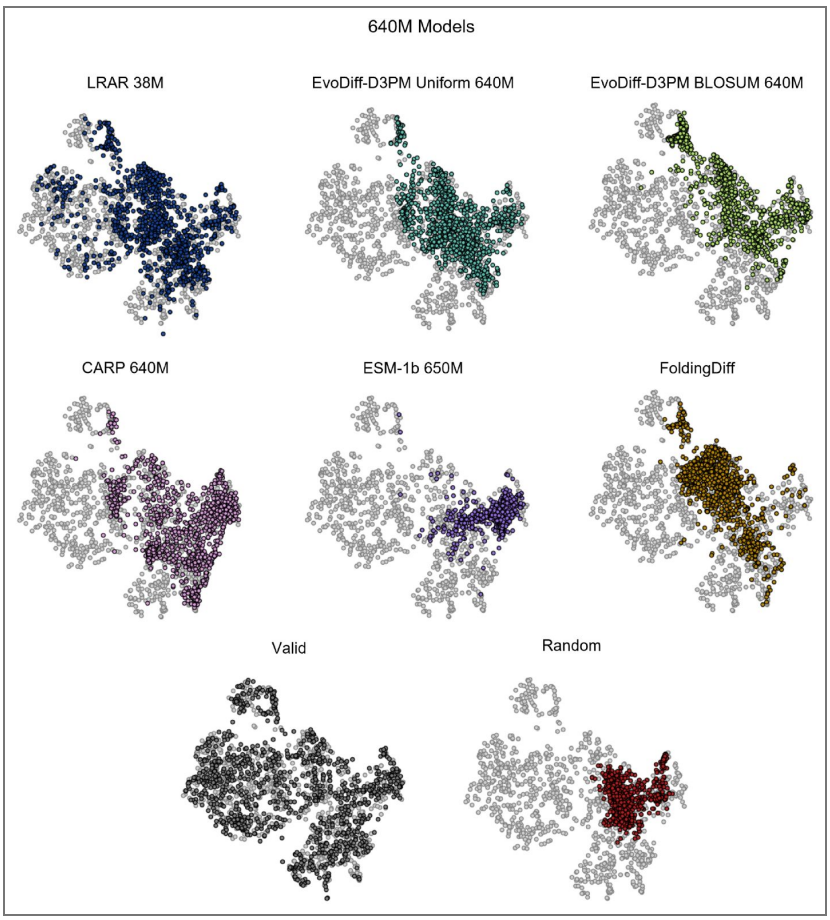



**Figure S8. Coverage of sequence and functional space for generated distributions from 640M parameter EvoDiff sequence models and baselines.**

UMAP of ProtT5 embeddings, annotated with FPD, of natural sequences from test set (grey, *n*= 1000) and of generated sequences from EvoDiff 640M parameter models and baselines (various colors, *n*=l000). A visualization of sequences from the validation set (dark grey, *n*= 1000) is included for reference. The visualization for the 640M OADM model is excluded due to inclusion in Fig. 3A.

**Figure S9. Structural features in generated sequences from all sequence models.**

**(A-B)** Multivariate distributions of helix and strand features in sequences from 38M (A) and 640M (B) parameter models, and baselines based on DSSP 3-state predictions and annotated with the KL divergence relative to the test set (*n*=1000 samples from each model). In (B), the distribution for the 640M OADM model is excluded (see Fig. 3B); the distribution for random sequences (*n*= 1000) is provided as reference.

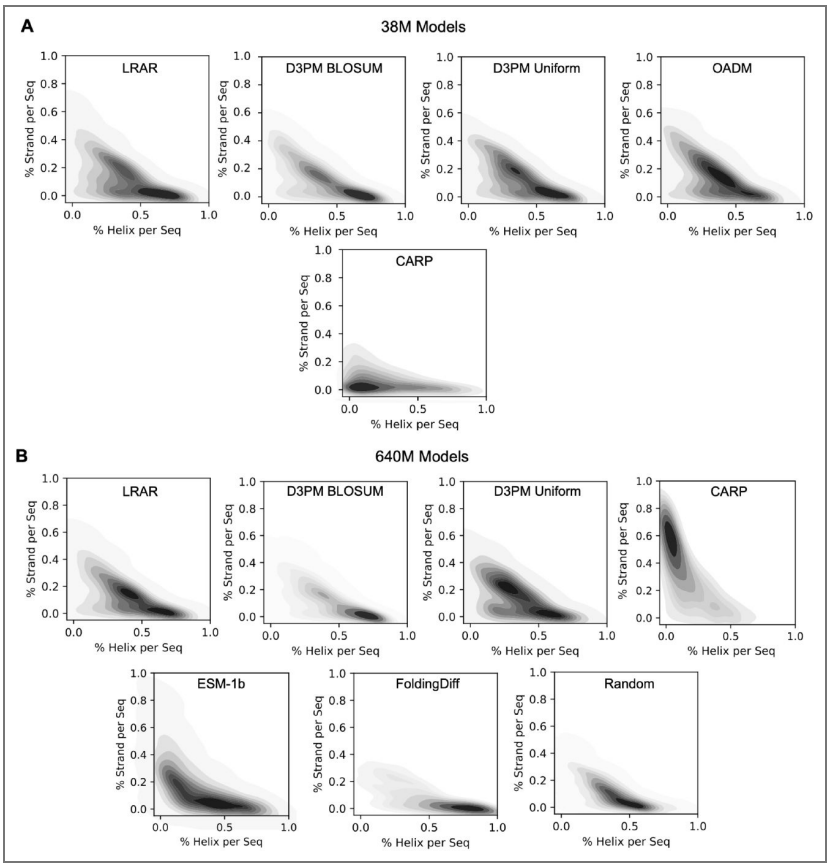



**Figure S10. Sequence pLDDT versus scPerplexity for EvoDiff MSA models,**

for sequences from the validation set (grey, *n*=250) and evaluated MSA models (various colors, *n*=250), except for EvoDiff-OADM-MSA-Max (EvoDiff-MSA, shown in Fig. 4F).

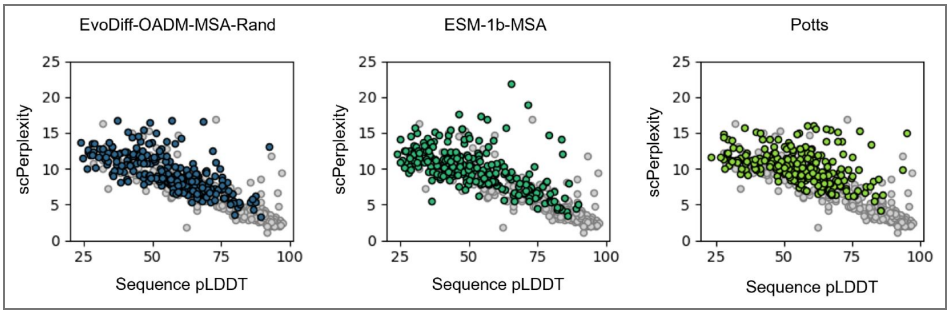


**Figure S11. Baseline performance of DR-BERT evaluator.**

**(A-C)** Distributions of DR-BERT predicted disorder scores across disordered and structured regions for sequences with true (A), scrambled (B), and randomly sampled (C) IDRs (*n*=l00).

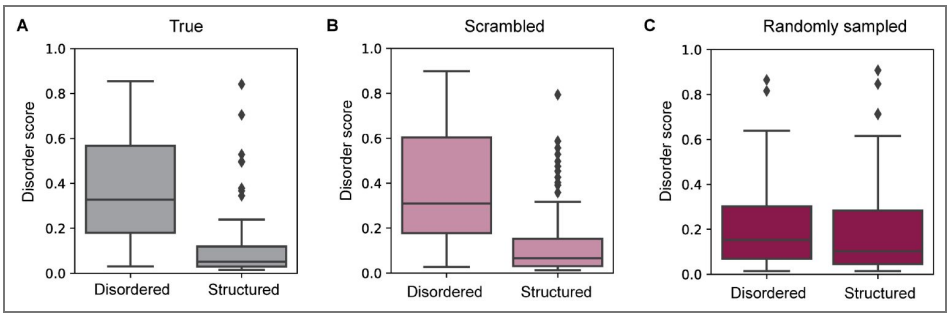



**Figure S12. Performance of DR-BERT evaluator on disorder regions.**

**(A-B)** Disorder scores predicted by DR-BERT for true (x-axis) vs. generated (y-axis) IDRs for the same given sequence, for generations from EvoDiff-Seq (A) and EvoDiff-MSA (B). Each dot represents an individual IDR (*n*=l00). The Pearson R is given for each of EvoDiff-Seq and EvoDiff-MSA.

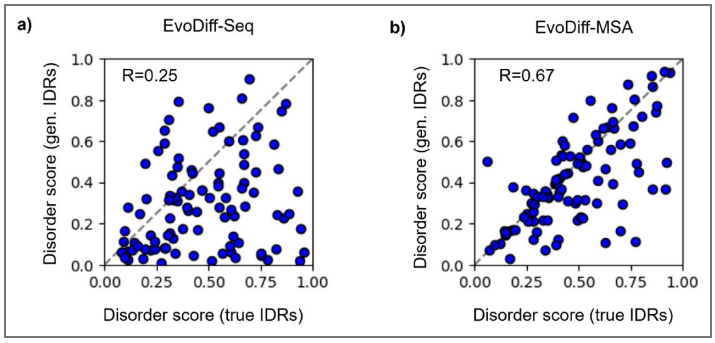
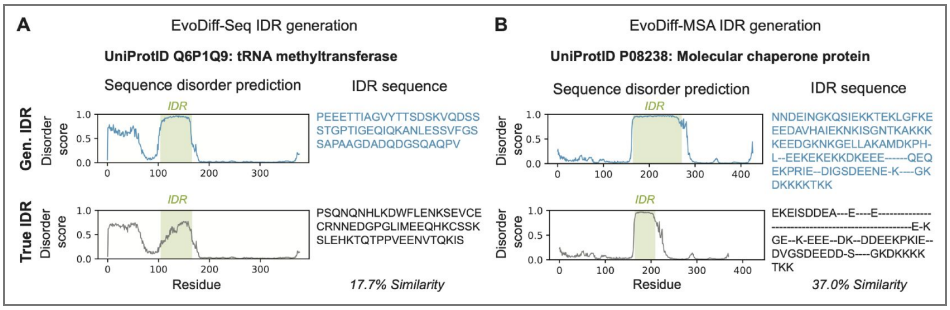



**Figure S13. Examples of predicted disorder scores**

and corresponding sequences for repre-sentative generated (top row) and native (bottom row) IDRs from EvoDiff-Seq **(A)** and EvoDiff-MSA **(B)**.



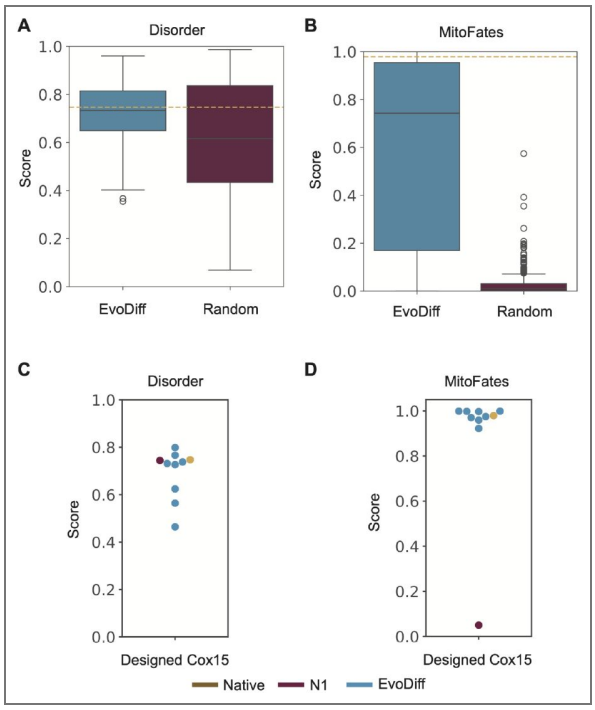



**Figure S14.** ***In-silico*** **predictions for inpainted Cox15 designs**

**(A)** Disorder scores predicted by DR-BERT and **(B)** MitoFates probability of presequences for 600 EvoDiff designs (blue) and 600 randomly sampled sequences (red). Scores for the native Cox15 IDR are denoted with a dashed yellow line. **(C)** Disorder and **(D)** Mitofates probability of presequence predictions for tested Cox15 IDRs, including EvoDiff designs (blue), native Cox15 IDR (yellow), and N1 (red).

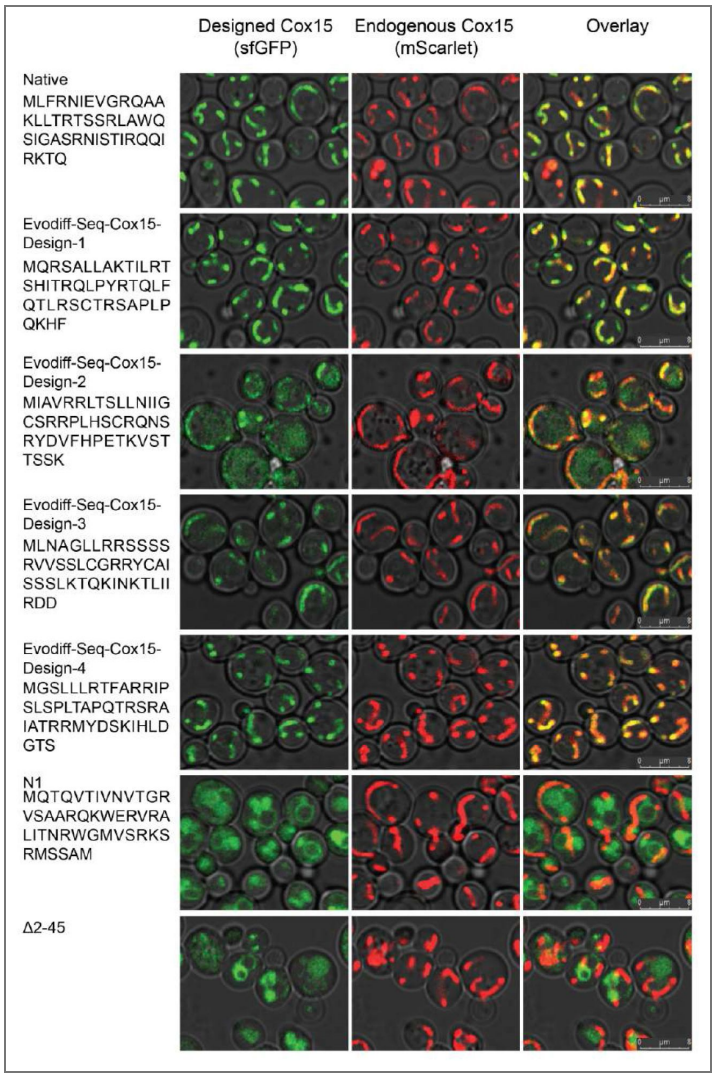



**Figure S15. Fluorescence microscopy imaging of yeast cells for EvoDiff-Seq generated Designs.**

Columns from left to right show: the N-terminal IDR sequence used for a given Cox15-sfGFP design (rows); signal from the designed Cox15 (sfGFP. colored in green); signal from the endogenous Cox 15 (mScarlet, colored in red); overlaid images. All images show brightfield in grey, to visualize cell boundaries. Images for the native IDR, N1, and Δ2-45 are included for reference.

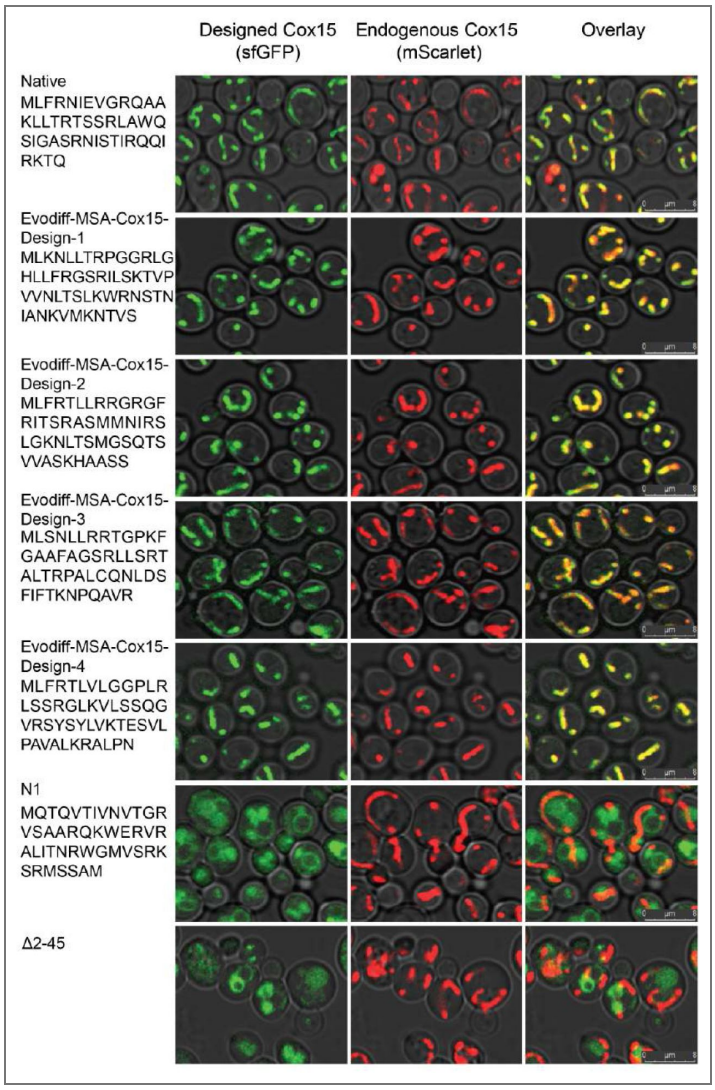



**Figure S16. Fluorescence microscopy imaging of yeast cells for EvoDiff-MSA generated Designs.**

Columns from left to right show: the N-terminal IDR sequence used for a given Cox15-sfGFP design (rows); signal from the designed Cox15 (sfGFP, colored in green); signal from the endogenous Cox15 (mScarlet, colored in red); overlaid images. All images show brightfield in grey, to visualize cell boundaries. Images for the native IDR, N1, and Δ2-45 are included for reference.

**Figure S17. Per-cell pixel-wise Pearson correlation coefficients of fluorescence microscopy signal intensities**

between the mScarlet channel and the sfGFP channel. Independent replicates are measured over *n*=110-264 cells segmented by YeastSpotter for each design. The native IDR is shown in yellow, N1 in red, and EvoDiff designs in blue.

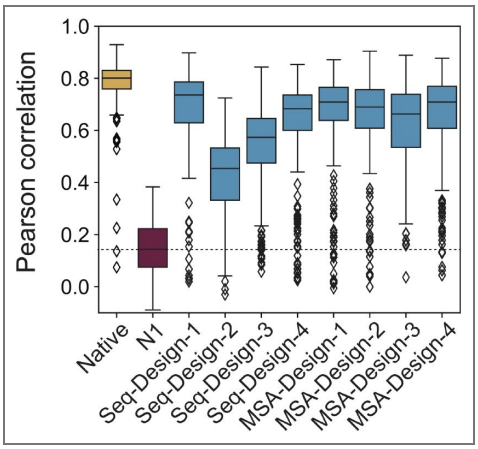



**Figure S18. EvoDiff-Seq and EvoDiff-MSA scaffold binding domains in sequence space**

**(A-D)** Generated sequences, predicted structures, and computed metrics for representative scaf-folding examples from EvoDiff-Seq **(A/B)** and EvoDiff-MSA **(C/D)**. Motif is shown in green, original scaffold in gray, and generated scaffold in blue.

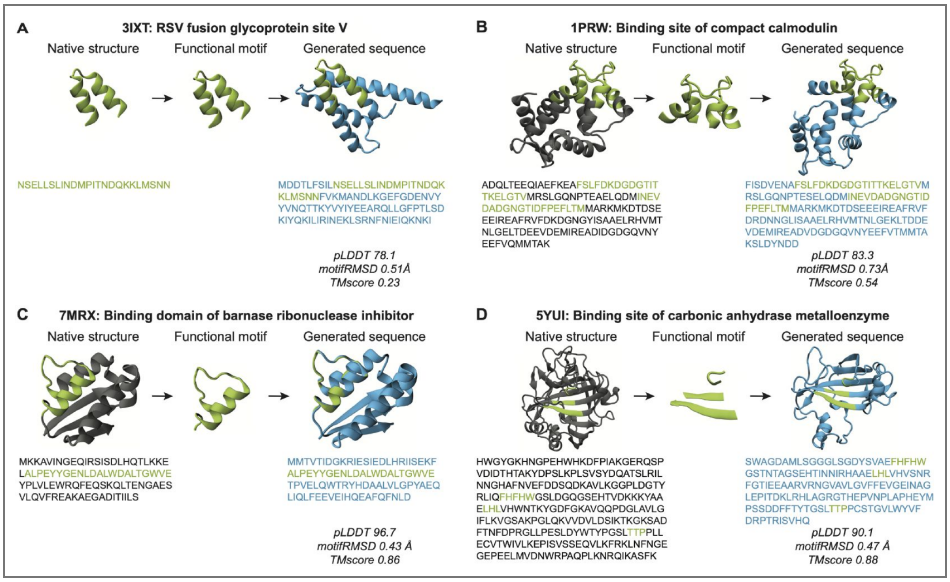


**Figure S19.** **Calmodulin binding assay.**

Included is the native IPRW protein (grey), successful EvoDiff design (EvoDiff-Seq-IPRW-Design-I, blue), an unsuccessful EvoDiff design (EvoDiff-Seq-IPRW-Design-2, orange), and buffer (red), with (solid-line) and without (dashed-line) $Ca^{2}$+. The mean value is reported, and error bars are represented by the std. dev taken from three independent trials

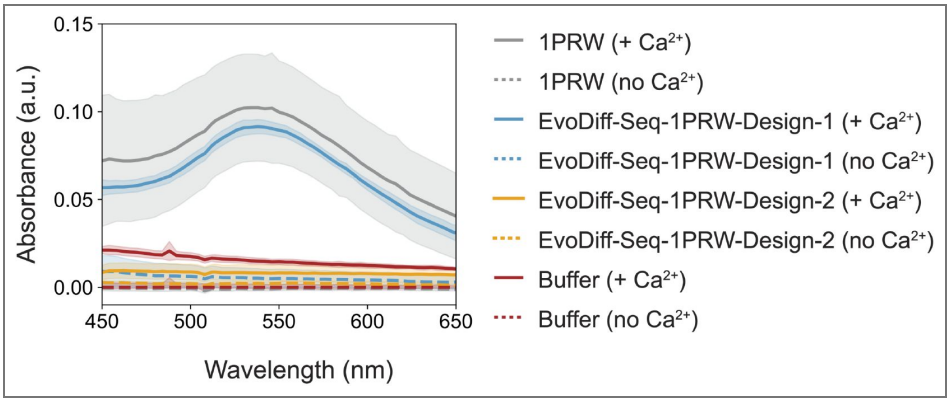



**Figure S20.** **BLI results for positive controls.**

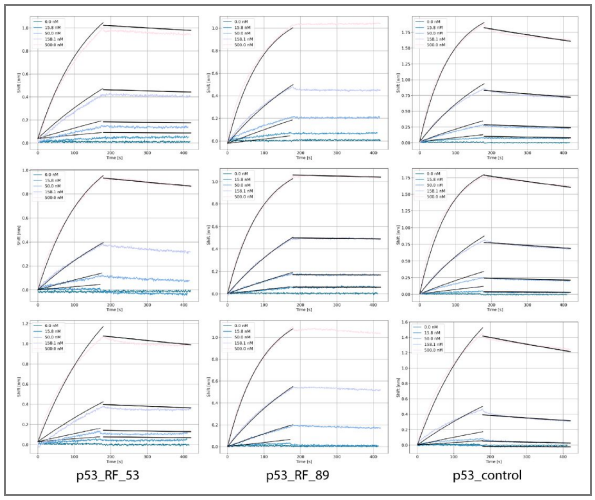



**Figure S21.** **BLI results for successful MDM2 binders generated by EvoDiff-Seq.**

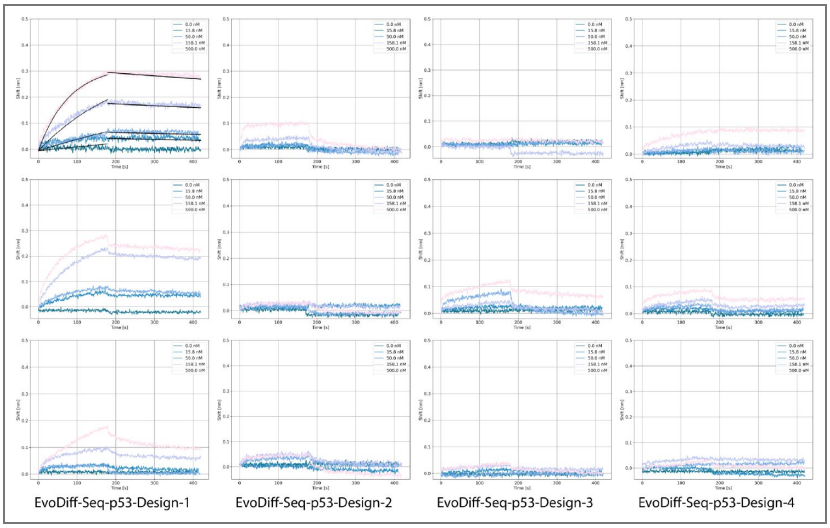


**Figure S22.** **BLI results for successful MDM2 binders generated by EvoDiff-MSA.**

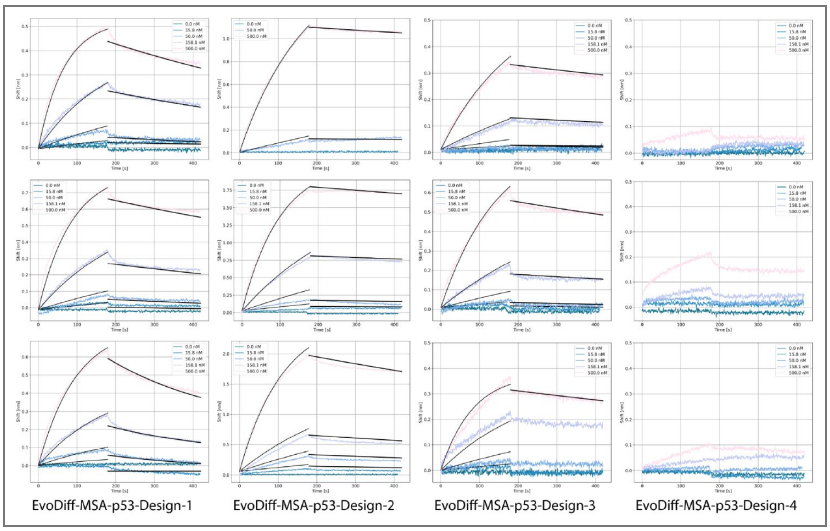



**Figure S23.** **BLI results for unsuccessful MDM2 binders EvoDiff-Seq-p53-Design-5 through EvoDiff-Seq-p53-Design-12**

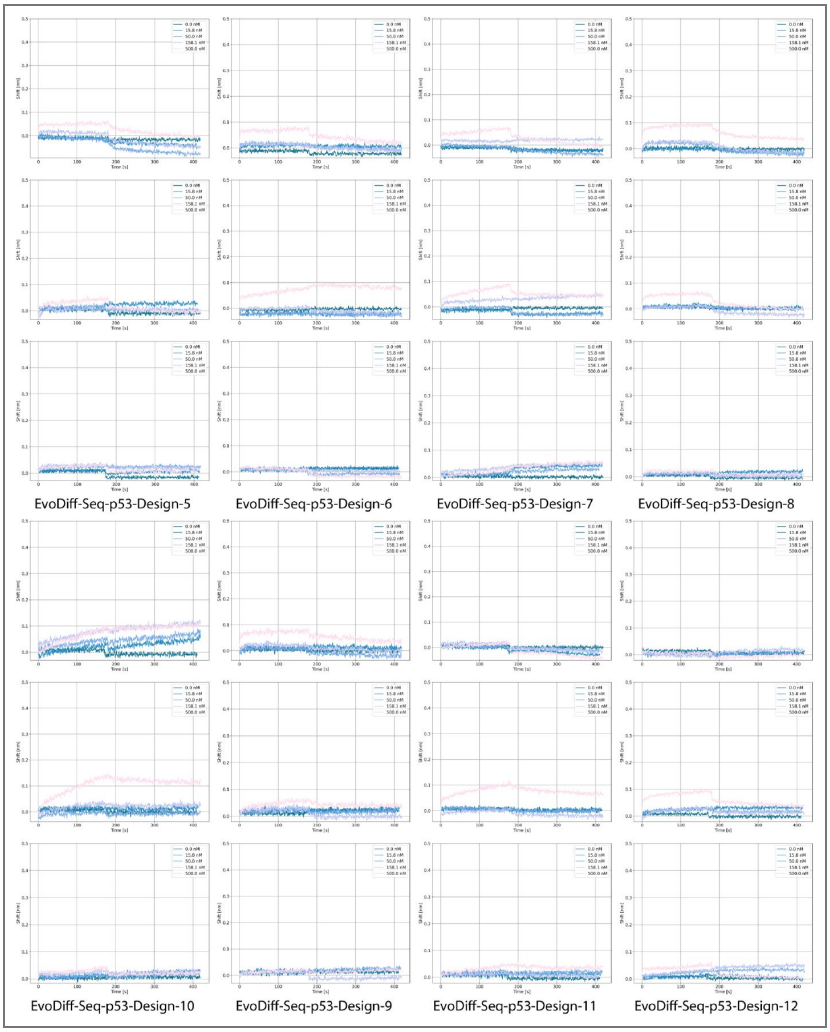


**Figure S24. BLI results for unsuccessful MDM2 binders EvoDiff-Seq-p53-Design-13 through EvoDiff-MSA-p53-Design-5**

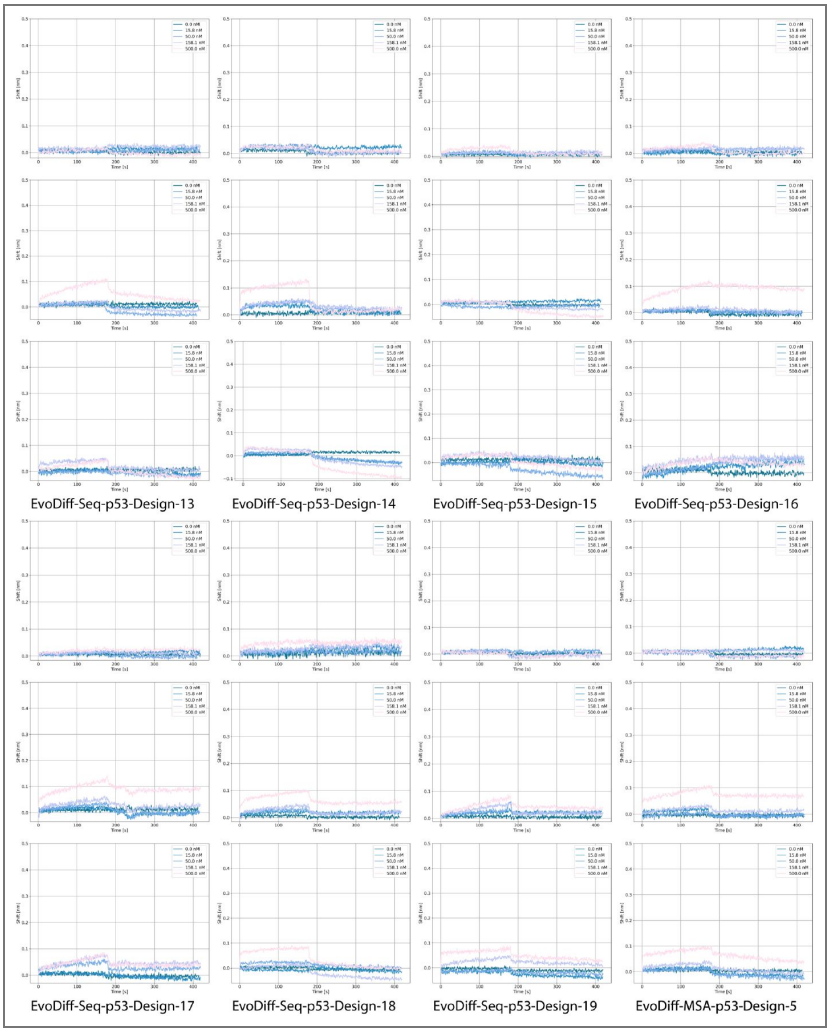
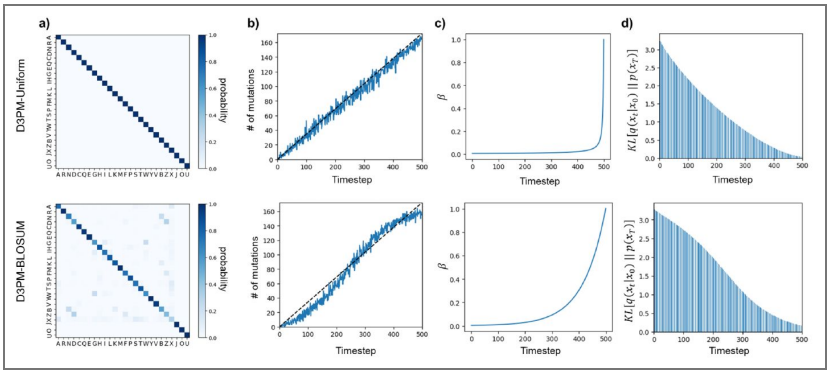



**Figure S25. Details of EvoDiff-D3PM corruption schemes.**

The top and bottom rows correspond to EvoDiff-D3PM-Uniform and EvoDiff-D3PM-BLOSUM, respectively. **(A)** Visualization of EvoDiff-D3PM transition matrices. **(B)** Evolution of the number of mutations accrued as a function of the diffusion timestep $t$ for a sample input. **(C)** Evolution of $\beta$ as a function of the diffusion timestep $t$. **(D)** Evolution of $D_{KL}$ $[q(x_t|x_0)||p(x_T)]$ as a function of the diffusion timestep $t$, indicating convergence to a uniform stationary distribution at $t = 500$ as $D_{KL}$ approaches zero.



| Model | parameters | Recon KL | perplexity | Hamming | FPD | SS KL |
|---|---|---|---|---|---|---|
| Test | - | 9.92e-4[1] | - | 0.0039[2] | 0.10[1] | 1.37e-5[1] |
| D3PM BLOSUM | 38M | 1.77e-2 | 17.16 | 0.83 ± 0.05 | 1.42 | 3.30e-5 |
| D3PM Uniform | 38M | 1.48e-3 | 18.82 | 0.83 ± 0.05 | 1.31 | 3.73e-5 |
| OADM | 38M | 1.11e-3 | 14.61 | 0.83 ± 0.07 | 0.92 | 1.61e-4 |
| D3PM BLOSUM | 640M | 3.73e-2 | 15.74 | 0.83 ± 0.05 | 1.53 | 4.96e-4 |
| D3PM Uniform | 640M | 2.90e-3 | 18.47 | 0.83 ± 0.05 | 1.35 | 2.13e-4 |
| OADM | 640M | 1.26e-3 | 13.05 | 0.83 ± 0.08 | 0.88 | 1.48e-4 |
| LRAR | 38M | 7.90e-4 | 12.38 | 0.82 ± 0.06 | 0.86 | 1.61e-4 |
| CARP | 38M | 5.71e-1 | 25.13 | 0.74 ± 0.07 | 6.30 | 2.72e-3 |
| LRAR | 640M | 7.01e-4 | 10.41 | 0.83 ± 0.06 | 0.63 | 1.76e-5 |
| CARP | 640M | 3.56e-1 | 31.77 | 0.84 ± 0.05 | 1.78 | 5.03e-3 |
| ESM-1b [3] | 650M | 4.91e-1 | 53.49 | 0.83 ± 0.06 | 6.67 | 5.48e-4 |
| ESM-2 [3] | 650M | 5.00e-1 | 68.39 | 0.84 ± 0.06 | 6.79 | 3.05e-3 |
| FoldingDiff [4] | 14M | 5.49e-2 | - | - | 1.64 | 1.76e-3 |
| RFdiffusion [5] | 60M | 7.19e-2 | - | - | 1.96 | 5.98e-3 |
| Random | - | 1.65e-1 | 20 | 0.85 ± 0.04 | 3.16 | 1.90e-4 |

[1] Calculated between the test set and validation set.
[2] Reported value is the minimum Hamming distance between any two natural sequences of the same length in UniRef50.
[3] Due to model constraints, the maximum sequence length sampled was 1022.
[4] For the FoldingDiff baseline, 1000 structures generated by FoldingDiff were randomly selected, and the corresponding 1000 inferred sequences were inverse-folded using ESM-IF. These sequences are between lengths of 50 and 128 residues.
[5] For the RFdiffusion baseline, 1000 structures were generated corresponding to the UniRef train distribution length, and 1000 corresponding sequences were inverse-folded using ESM-IF.

**Table S1. Performance of EvoDiff sequence models.**

The reconstruction KL (Recon KL) was calculated between the distribution of amino acids in the test set and in generated samples ($n$=1000). The perplexity was computed on 25k samples from the test set. The minimum Hamming distance to any train sequence of the same length (Hamming) is reported for each model as the mean ± standard deviation over the generated samples. Fréchet ProtT5 distance (FPD) was calculated between the test set and generated samples. The secondary structure KL (SS KL) was calculated between the means of the predicted secondary structures of the test and generated samples.

**Table S2. Validation-set perplexities for EvoDiff MSA models.**

The perplexity is calculated based on the ability of each model to reconstruct a subsampled MSA from the validation set. "Max Perplexity" and "Rand. Perplexity" indicate MaxHamming and Random subsampling, respectively, for construction of the validation MSA.

| Corruption | Subsampling | Params | Max Perplexity | Rand. Perplexity |
|---|---|---|---|---|
| D3PM BLOSUM | Random | 100M | 11.35 | 8.31 |
| D3PM BLOSUM | Max | 100M | 10.98 | 7.61 |
| D3PM Uniform | Random | 100M | 10.14 | 6.77 |
| D3PM Uniform | Max | 100M | 10.06 | 6.66 |
| OADM | Random | 100M | 6.05 | 3.64 |
| OADM | Max | 100M | 6.14 | 3.60 |
| ESM-MSA-1b | Max | 100M | 11.20 | 5.89 |

**Table S3. Structural plausibility metrics for EvoDiff sequence models and baselines.**

Metrics are reported as the mean ± standard deviation for 1000 generated samples for each model.

| Model | Params | ESM-IF scPerplexity | ProteinMPNN scPerplexity | OmegaFold pLDDT |
|---|---|---|---|---|
| Test | - | $8.04 \pm 4.04$ | $3.09 \pm 0.63$ | $68.25 \pm 17.85$ |
| D3PM Blosum | 38M | $12.38 \pm 2.06$ | $3.80 \pm 0.49$ | $42.76 \pm 14.55$ |
| D3PM Uniform | 38M | $12.03 \pm 2.04$ | $3.77 \pm 0.50$ | $42.37 \pm 14.39$ |
| OADM | 38M | $11.61 \pm 2.38$ | $3.72 \pm 0.50$ | $43.78 \pm 14.18$ |
| D3PM Blosum | 640M | $11.86 \pm 2.21$ | $3.73 \pm 0.48$ | $44.14 \pm 13.80$ |
| D3PM Uniform | 640M | $12.29 \pm 2.05$ | $3.78 \pm 0.49$ | $41.65 \pm 14.32$ |
| OADM | 640M | $11.53 \pm 2.50$ | $3.71 \pm 0.52$ | $44.46 \pm 14.62$ |
| LRAR | 38M | $11.61 \pm 2.38$ | $3.64 \pm 0.56$ | $48.26 \pm 14.87$ |
| CARP | 38M | $9.68 \pm 2.56$ | $3.66 \pm 0.62$ | $50.79 \pm 12.06$ |
| LRAR | 640M | $10.99 \pm 2.63$ | $3.59 \pm 0.54$ | $48.71 \pm 15.47$ |
| CARP | 640M | $14.13 \pm 2.42$ | $4.05 \pm 0.52$ | $41.56 \pm 14.35$ |
| ESM-1b | 650M | $13.90 \pm 2.44$ | $3.47 \pm 0.68$ | $58.07 \pm 15.64$ |
| ESM-2 | 650M | $14.02 \pm 2.87$ | $3.58 \pm 0.69$ | $50.70 \pm 15.67$ |
| Random | - | $14.68 \pm 1.97$ | $3.96 \pm 0.50$ | $39.97 \pm 14.05$ |

**Table S4. Performance of EvoDiff MSA models in generating query sequences conditioned on MSAs.**

Metrics are reported as the mean ± standard deviation over 250 generated samples for each model.

| Model | scPerplexity | pLDDT | Seq. similarity | TM score |
|---|---|---|---|---|
| Valid | 5.93 ± 3.19 | 73.99 ± 17.80 | 14.58 ± 21.64 [1] | - |
| OADM (Rand) - Rand MSA | 9.41 ± 2.61 | 55.99 ± 14.75 | 6.13 ± 9.88 | 0.49 ± 0.23 |
| OADM (Max) - Max MSA | 9.38 ± 2.57 | 57.08 ± 16.01 | 6.74 ± 11.00 | 0.50 ± 0.23 |
| OADM (Max) - Rand MSA | 9.59 ± 2.69 | 54.95 ± 16.83 | 6.55 ± 10.49 | 0.46 ± 0.23 |
| ESM-MSA-1b | 10.05 ± 2.92 | 51.64 ± 16.54 | 7.13 ± 11.60 | 0.40 ± 0.23 |
| Potts | 10.34 ± 2.26 | 55.46 ± 13.82 | 12.01 ± 17.19 | 0.17 ± 0.10 |

[1] Sequence similarity is calculated between the original query sequence and all the sequences in the MSA.

**Table S5. Scaffolding performance of EvoDiff-Seq.**

Number of scaffolding successes out of 100 generations for RFdiffusion, EvoDiff-Seq, the LRAR baseline, the CARP baseline, and randomly sampled scaffolds (Random), for each of 17 scaffolding problems. The bottom row contains the total number of successful scaffolds generated per model.

| PDB | RFdiffusion | EvoDiff-Seq | LRAR | CARP | Random |
|---|---|---|---|---|---|
| 1BCF | 100 | 24 | 0 | 4 | 0 |
| 6E6R | 71 | 16 | 7 | 3 | 1 |
| 2KL8 | 88 | 0 | 1 | 1 | 0 |
| 6EXZ | 42 | 0 | 0 | 0 | 0 |
| 1YCR | 74 | 13 | 12 | 10 | 7 |
| 6VW1 | 69 | 1 | 0 | 0 | 0 |
| 4JHW | 0 | 0 | 0 | 0 | 0 |
| 5TPN | 61 | 0 | 0 | 0 | 0 |
| 4ZYP | 40 | 0 | 0 | 0 | 0 |
| 3IXT | 25 | 23 | 22 | 13 | 7 |
| 7MRX | 7 | 0 | 0 | 0 | 0 |
| 1PRW | 8 | 68 | 70 | 54 | 5 |
| 5IUS | 2 | 0 | 0 | 0 | 0 |
| 5YUI | 0 | 4 | 0 | 0 | 0 |
| 5WN9 | 0 | 0 | 0 | 0 | 2 |
| 1QJG | 0 | 0 | 0 | 0 | 0 |
| 5TRV | 22 | 0 | 0 | 0 | 0 |
| Total | 610 | 149 | 112 | 85 | 22 |

| PDB | RFdiffusion | EvoDiff-MSA (Max) | EvoDiff-MSA (Random) | ESM-MSA |
|---|---|---|---|---|
| 1BCF | 100 | 100 | 98 | 99 |
| 6E6R | 71 | 87 | 63 | 96 |
| 2KL8 | 88 | 11 | 31 | 42 |
| 6EXZ | 42 | 86 | 87 | 73 |
| 1YCR | 74 | 3 | 0 | 0 |
| 6VW1 | 69 | 4 | 3 | 4 |
| 4JHW | 0 | 0 | 0 | 0 |
| 5TPN | 61 | 0 | 0 | 0 |
| 4ZYP | 40 | 0 | 0 | 0 |
| 3IXT | 25 | 1 | 0 | 5 |
| 7MRX | 7 | 72 | 68 | 66 |
| 1PRW | 8 | 48 | 46 | 92 |
| 5IUS | 2 | 3 | 1 | 7 |
| 5YUI | 0 | 58 | 44 | 70 |
| 5WN9 | 0 | 0 | 0 | 0 |
| 1QJG | 0 | 34 | 22 | 38 |
| 5TRV | 22 | 15 | 12 | 12 |
| Total | 610 | 522 | 475 | 604 |

**Table S6. Scaffolding performance of EvoDiff-MSA.**

Number of scaffolding successes out of 100 generations for RFdiffusion, EvoDiff-MSA (Max), EvoDiff-MSA (Random), and the ESM-MSA baseline, for each of 17 scaffolding problems. The bottom row contains the total number of successful scaffolds generated per model.

## Data availability

Code is available at *https://github.com/microsoft/evodiff*. Model weights, generated sequences, and computed metrics are available at *https://zenodo.org/record/8332830*.

## Acknowledgements

The authors thank Christian Dallago for helpful discussions about the methods, applications, and evaluations; Jonathan Carlson for valuable feedback on the manuscript; Kevin E. Wu for providing sequences generated from FoldingDiff; Joseph Watson and David Juergens for providing sequences generated from RFDiffusion; and Alessandro Sordoni, Hannes Schultz, Rémi Piché-Taillefer, Remi Tachet des Combes, and Sean Whitzell for assistance on using Microsoft's compute resources ; Philip Rosenfield for his immense effort in managing the project; Colby T. Ford for providing valuable input on the GitHub repository; and Hannah Wayment-Steele for identifying disordered protein annotations.

## Additional information

### Author contributions

Conceptualization: S.A., N.T., N.F., A.P.A., K.K.Y.; Methodology: S.A., N.T., R.vdB., A.P.A., K.K.Y.; Software Programming: S.A., N.T., K.K.Y.; Experimental Design: S.A., R.S., A.M.M, A.X.L., A.P.A., K.K.Y.; Investigation: S.A., N.T., R.S., A.X.L., A.P.A., K.K.Y.; Validation: S.A., N.T., R.S., A.X.L., A.P.A., K.K.Y.; Formal analysis: S.A., N.T., A.X.L., A.P.A., K.K.Y.; Resources Provision: N.F., K.K.Y.; Data Curation: S.A., N.T., A.X.L., K.K.Y.; Visualization: S.A., R.S., A.P.A., K.K.Y; Writing - Original Draft: S.A., A.P.A., K.K.Y.; Writing - Review & Editing: S.A., N.T., R.vdB., N.Ten., R.S., A.M.M., A.X.L., N.F., A.P.A., K.K.Y.; Supervision: A.M.M., N.F., A.P.A., K.K.Y.

### Author ORCID iDs

**Alan M Moses**: https://orcid.org/0000-0003-3118-3121
**Kevin K Yang**: https://orcid.org/0000-0001-9045-6826

## Additional files

***Supplemental Table 1***

## References

1. **Wu Z,** Johnston KE, Arnold FH, Yang KK (2021) Protein sequence design with deep generative models. *Current Opinion in Chemical Biology* **65**:18-27 https://doi.org/10.1016/j.cbpa.2021.04.004 | PubMed
2. **Lovelock SL,** et al. (2022) The road to fully programmable protein catalysis. *Nature* **606**:49-58 https://doi.org/10.1038/s41586-022-04456-z | PubMed
3. **Sohl-Dickstein J,** Weiss E, Maheswaranathan N, Ganguli S (2015) Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256-2265
4. **Dhariwal P,** Nichol A (2021) Diffusion models beat GANs on image synthesis. In: Advances in Neural Information Processing Systems. **34** pp. 8780-8794
5. **Ho J,** Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. **33** pp. 6840-6851
6. **Rombach R,** Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684-10695 https://doi.org/10.1109/cvpr52688.2022.01042
7. **Anand N,** Achim T (2022) Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv* https://doi.org/10.48550/arxiv.2205.15019

8. **Wu KE,** et al. (2022) Protein structure generation via folding diffusion. *arXiv* https://doi.org/10.48550/arXiv.2209.15611
9. **Trippe BL,** et al. (2023) Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. In: The Eleventh International Conference on Learning Representations. **11**
10. **Watson JL,** et al. (2023) De novo design of protein structure and function with RFdiffusion. *Nature* **620**:1089-1100 https://doi.org/10.1038/s41586-023-06415-8 | PubMed
11. **Ingraham J,** et al. (2022) Illuminating protein space with a programmable generative model. *bioRxiv* https://doi.org/10.1101/2022.12.01.518682
12. **Lin Y,** AlQuraishi M (2023) Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. In: Proceedings of the 40th International Conference on Machine Learning.
13. **Yim J,** et al. (2023) SE (3) diffusion model with application to protein backbone generation. *arXiv* https://doi.org/10.48550/arxiv.2302.02277
14. **Lee JS,** Kim J, Kim PM (2023) Score-based generative modeling for de novo protein design. *Nature Computational Science* **3**:382-392 https://doi.org/10.1038/s43588-023-00440-3 | PubMed
15. **Chu AE,** Cheng L, El Nesr G, Xu M, Huang P.-S (2023) An all-atom protein generative model. *bioRxiv* https://doi.org/10.1101/2023.05.24.542194 | PubMed
16. **Tokuriki N,** Tawfik DS (2009) Protein dynamism and evolvability. *Science* **324**:203-207 https://doi.org/10.1126/science.1169375 | PubMed
17. **Romero PA,** Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology* **10**:866-876 https://doi.org/10.1038/nrm2805 | PubMed
18. **Gao J,** Truhlar DG (2002) Quantum mechanical methods for enzyme kinetics. *Annual Review of Physical Chemistry* **53**:467-505 https://doi.org/10.1146/annurev.physchem.53.091301.150114 | PubMed
19. **Davey NE** (2019) The functional importance of structure in unstructured protein regions. *Current Opinion in Structural Biology* **56**:155-163 https://doi.org/10.1016/j.sbi.2019.03.009 | PubMed
20. **Frauenfelder H,** Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. *Science* **254**:1598-1603 https://doi.org/10.1126/science.1749933 | PubMed
21. **McCammon J** (1984) Protein dynamics. *Reports on Progress in Physics* **47**:1 https://doi.org/10.1088/0034-4885/47/1/001
22. **Henzler-Wildman K,** Kern D (2007) Dynamic personalities of proteins. *Nature* **450**:964-972 https://doi.org/10.1038/nature06522 | PubMed
23. **Agarwal PK** (2005) Role of protein dynamics in reaction rate enhancement by enzymes. *Journal of the American Chemical Society* **127**:15248-15256 https://doi.org/10.1021/ja055251s | PubMed
24. **Hoogeboom E,** et al. (2022) Autoregressive diffusion models. In: The Eleventh International Conference on Learning Representations. **11**
25. **Austin J,** Johnson DD, Ho J, Tarlow D, van den Berg R (2021) Structured denoising diffusion models in discrete state-spaces. In: Advances in Neural Information Processing Systems. **34**
26. **Suzek BE,** et al. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**:926-932 https://doi.org/10.1093/bioinformatics/btu739 | PubMed
27. **Yang KK,** Fusi N, Lu AX (2022) Convolutions are competitive with transformers for protein sequence pretraining. *bioRxiv* https://doi.org/10.1101/2022.05.19.492714
28. **Rao RM,** et al. (2021) MSA Transformer. In: Proceedings of the 38th International Conference on Machine Learning. **139** pp. 8844-8856
29. **Ahdritz G,** et al. (2022) OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv* https://doi.org/10.1101/2022.11.20.517210

30. **Verkuil R,** et al. (2022) Language models generalize beyond natural proteins. *bioRxiv* https://doi.org/10.1101/2022.12.21.521521

31. **Wu R,** et al. (2022) High-resolution de novo structure prediction from primary sequence. *bioRxiv* https://doi.org/10.1101/2022.07.21.500999

32. **Ruff KM,** Pappu RV (2021) AlphaFold and implications for intrinsically disordered proteins. *Journal of Molecular Biology* **433**:167208 https://doi.org/10.1016/j.jmb.2021.167208 | PubMed

33. **Hsu C,** et al. (2022) Learning inverse folding from millions of predicted structures. In: Proceedings of the 39th International Conference on Machine Learning. **162** pp. 8946-8970

34. **Dauparas J,** et al. (2022) Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**:49-56 https://doi.org/10.1126/science.add2187 | PubMed

35. **Littmann M,** Heinzinger M, Dallago C, Olenyi T, Rost B (2021) Embeddings from deep learning transfer GO annotations beyond homology. *Scientific Reports* **11**:1160 https://doi.org/10.1038/s41598-020-80786-0 | PubMed

36. **Gene Ontology Consortium** (2019) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Research* **47**:D330-D338 https://doi.org/10.1093/nar/gky1055 | PubMed

37. **Elnaggar A,** et al. (2021) ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**:7112-7127 https://doi.org/10.1109/TPAMI.2021.3095381 | PubMed

38. **Kabsch W,** Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* **22**:2577-2637 https://doi.org/10.1002/bip.360221211 | PubMed

39. **Vorberg S,** Seemayer S, Söding J (2018) Synthetic protein alignments by CCMgen quantify noise in residue-residue contact prediction. *PLOS Computational Biology* **14**:e1006526 https://doi.org/10.1371/journal.pcbi.1006526 | PubMed

40. **Vögtle F.-N,** et al. (2009) Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. *Cell* **139**:428-439 https://doi.org/10.1016/j.cell.2009.07.045 | PubMed

41. **Zarin T,** et al. (2019) Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *eLife* **8**:e46883 https://doi.org/10.7554/eLife.46883 | PubMed

42. **Uversky VN** (2018) Intrinsic disorder, protein–protein interactions, and disease. *Advances in Protein Chemistry and Structural Biology* **110**:85-121 https://doi.org/10.1016/bs.apcsb.2017.06.005 | PubMed

43. **Mészáros B,** Simon I, Dosztányi Z (2011) The expanding view of protein–protein interactions: complexes involving intrinsically disordered proteins. *Physical Biology* **8**:035003 https://doi.org/10.1088/1478-3975/8/3/035003 | PubMed

44. **Wright PE,** Dyson HJ (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology* **16**:18-29 https://doi.org/10.1038/nrm3920 | PubMed

45. **Vacic V,** et al. (2012) Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLOS Computational Biology* **8**:1-14 https://doi.org/10.1371/journal.pcbi.1002709 | PubMed

46. **Coskuner-Weber O,** Mirzanli O, Uversky VN (2022) Intrinsically disordered proteins and proteins with intrinsically disordered regions in neurodegenerative diseases. *Biophysical Reviews* **14**:679-707 https://doi.org/10.1007/s12551-022-00968-0 | PubMed

47. **Iakoucheva LM,** Brown CJ, Lawson JD, Obradović Z, Dunker AK (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology* **323**:573-584 https://doi.org/10.1016/s0022-2836(02)00969-5 | PubMed

48. **Lu AX,** et al. (2022) Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning. *PLOS Computational Biology* **18**:e1010238 https://doi.org/10.1371/journal.pcbi.1010238 | PubMed

49. **Nambiar A,** Forsyth JM, Liu S, Maslov S (2023) DR-BERT: A protein language model to annotate disordered regions. *bioRxiv* https://doi.org/10.1101/2023.02.22.529574
50. **Maccecchini M,** Rudin Y, Schatz G (1979) Transport of proteins across the mitochondrial outer membrane. A precursor form of the cytoplasmically made intermembrane enzyme cytochrome c peroxidase. *Journal of Biological Chemistry* **254**:7468-7471 https://doi.org/10.1016/s0021-9258(18)35963-5 | PubMed
51. **Saitoh T,** et al. (2007) Tom20 recognizes mitochon-drial presequences through dynamic equilibrium among multiple bound states. *The EMBO journal* **26**:4777-4787 https://doi.org/10.1038/sj.emboj.7601888 | PubMed
52. **Saitoh T,** et al. (2011) Crystallographic snapshots of tom20– mitochondrial presequence interactions with disulfide-stabilized peptides. *Biochemistry* **50**:5487-5496 https://doi.org/10.1021/bi200470x | PubMed
53. **Fukasawa Y,** et al. (2015) MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Molecular & Cellular Proteomics* **14**:1113-1126 https://doi.org/10.1074/mcp.m114.043083 | PubMed
54. **Strome B,** Elemam K, Pritisanac I, Forman-Kay JD, Moses AM (2023) Computational design of intrinsically disordered protein regions by matching bulk molecular properties. *bioRxiv* https://doi.org/10.1101/2023.04.28.538739
55. **Wang J,** et al. (2022) Scaffolding protein functional sites using deep learning. *Science* **377**:387-394 https://doi.org/10.1126/science.abn2100 | PubMed
56. **Keech AM,** et al. (1997) Spectroscopic studies of cobalt (II) binding to Escherichia coli bacterioferritin. *Journal of Biological Chemistry* **272**:422-429 https://doi.org/10.1074/jbc.272.1.422 | PubMed
57. **Chène P** (2003) Inhibiting the p53–mdm2 interaction: an important target for cancer therapy. *Nature reviews cancer* **3**:102-109 https://doi.org/10.1038/nrc991 | PubMed
58. **Wells M,** et al. (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proceedings of the National academy of Sciences* **105**:5762-5767 https://doi.org/10.1073/pnas.0801353105 | PubMed
59. **Jiang Z,** et al. (2023) PRO-LDM: Protein sequence generation with conditional latent diffusion models. *bioRxiv* https://doi.org/10.1101/2023.08.22.554145
60. **Zhou B,** et al. (2023) Conditional protein denoising diffusion generates pro-grammable endonucleases. *bioRxiv* https://doi.org/10.1101/2023.08.10.552783
61. **Gruver N,** et al. (2023) Protein design with guided discrete diffusion. *arXiv* https://doi.org/10.48550/arxiv.2305.20009
62. **Lisanza SL,** et al. (2023) Joint generation of protein sequence and structure with RoseTTAFold sequence space diffusion. *bioRxiv* https://doi.org/10.1101/2023.05.08.539766
63. **Shi C,** Wang C, Lu J, Zhong B, Tang J (2022) Protein sequence and structure co-design with equivariant translation. In: The Eleventh International Conference on Learning Representations.
64. **Madani A,** et al. (2023) Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* **41**:1099-1106 https://doi.org/10.1038/s41587-022-01618-2 | PubMed
65. **Ferruz N,** Schmidt S, Höcker B (2022) ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications* **13**:4348 https://doi.org/10.1038/s41467-022-32007-7 | PubMed
66. **Truong TF,** Bepler T (2023) PoET: A generative model of protein families as sequences-of-sequences. *arXiv* https://doi.org/10.48550/arxiv.2306.06156
67. **Zhang L,** Chen J, Shen T, Li Y, Sun S (2023) Enhancing the protein tertiary structure prediction by multiple sequence alignment generation. *arXiv* https://doi.org/10.48550/arxiv.2306.01824
68. **Rives A,** et al. (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**:e2016239118 https://doi.org/10.1073/pnas.2016239118 | PubMed

69. **Nisonoff H,** Xiong J, Allenspach S, Listgarten J (2024) Unlocking guidance for discrete state-space diffusion and flow models. *arXiv* https://doi.org/10.48550/arxiv.2406.01572

70. **Gruver N,** et al. (2024) Protein design with guided discrete diffusion. In: Advances in neural information processing systems. **36**

71. **Monzon V,** Haft DH, Bateman A (2022) Folding the unfoldable: using AlphaFold to explore spurious proteins. *Bioinformatics Advances* **2**:vbab043 https://doi.org/10.1093/bioadv/vbab043 | PubMed

72. **Pak MA,** et al. (2023) Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLOS One* **18**:e0282689 https://doi.org/10.1371/journal.pone.0282689 | PubMed

73. **Liu S,** et al. (2023) A text-guided protein design framework. *arXiv* https://doi.org/10.48550/arxiv.2302.04611

74. **Hoogeboom E,** Nielsen D, Jaini P, Forré P, Welling M (2021) Argmax flows and multinomial diffusion: Learning categorical distributions. *arXiv* https://doi.org/10.48550/arxiv.2102.05379

75. **Song J,** Meng C, Ermon S (2020) Denoising diffusion implicit models. *arXiv* https://doi.org/10.48550/arxiv.2010.02502

76. **Henikoff S,** Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**:10915-10919 https://doi.org/10.1073/pnas.89.22.10915 | PubMed

77. **Kalchbrenner N,** et al. (2017) Neural machine translation in linear time. *arXiv* https://doi.org/10.48550/arXiv.1610.10099

78. **Paszke A,** et al. (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems. **32** pp. 8024-8035

79. **Vaswani A,** et al. (2017) Attention is all you need. *arXiv* https://doi.org/10.48550/arxiv.1706.03762

80. **Kingma DP,** Ba J (2017) Adam: A method for stochastic optimization. *arXiv* https://doi.org/10.48550/arXiv.1412.6980

81. **Johnson M,** et al. (2008) Ncbi blast: a better web interface. *Nucleic acids research* **36**:W5-W9 https://doi.org/10.1093/nar/gkn201 | PubMed

82. **Ferruz N,** et al. (2023) From sequence to function through structure: deep learning for protein design. *Computational and Structural Biotechnology Journal* **21**:238-250 https://doi.org/10.1016/j.csbj.2022.11.014 | PubMed

83. **Zhang Y,** Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57**:702-710 https://doi.org/10.1002/prot.20264 | PubMed

84. **Studier FW** (2005) Protein production by auto-induction in high-density shaking cultures. *Protein expression and purification* **41**:207-234 https://doi.org/10.1016/j.pep.2005.01.016 | PubMed

85. **Hanson J,** Yang Y, Paliwal K, Zhou Y (2017) Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **33**:685-692 https://doi.org/10.1093/bioinformatics/btw678 | PubMed

86. **Altenhoff AM,** et al. (2021) OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Research* **49**:D373-D379 https://doi.org/10.1093/nar/gkaa1007 | PubMed

87. **Brachmann C Baker,** et al. (1998) Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14**:115-132 https://doi.org/10.1002/(sici)1097-0061(19980130)14:2<115::aid-yea204>3.0.co;2-2 | PubMed

88. **Gietz RD,** Schiestl RH (2007) High-efficiency yeast trans-formation using the LiAc/SS carrier DNA/PEG method. *Nature protocols* **2**:31-34 https://doi.org/10.1038/nprot.2007.13 | PubMed

89. **Pédelacq J.-D,** Cabantous S, Tran T, Terwilliger TC, Waldo GS (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nature biotechnology* **24**:79-88 https://doi.org/10.1038/nbt1172 | PubMed

90. **Gibson DG,** et al. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods* **6**:343-345 https://doi.org/10.1038/nmeth.1318 | PubMed

91. **Lu AX,** Zarin T, Hsu IS, Moses AM (2019) YeastSpotter: accurate and parameter-free web segmentation for microscopy images of yeast cells. *Bioinformatics* **35**:4525-4527 https://doi.org/10.1093/bioinformatics/btz402 | PubMed

92. **Jumper J,** et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**:583-589 https://doi.org/10.1038/s41586-021-03819-2 | PubMed

# Peer reviews

## Reviewer #1 (Public review):

Summary:

This manuscript presents a new foundation model, EvoDiff, for designing primary protein sequences. By leveraging an evolutionary-scale dataset, the model can be applied to evolution-guided sequence generation, sequence inpainting, and functional scaffolding.

Strengths:

The model provides an efficient approach for designing protein sequences and could be useful for developing protein therapeutics, engineering enzymes, designing biomaterials, and many other applications. The manuscript presents solid results showing that proteins designed by EvoDiff can achieve the same biological functions as their wild-type counterparts.

Weaknesses:

Compared with other sequence-generation models, EvoDiff does not substantially improve the success rate, suggesting that significant experimental effort is still required to screen and identify successful hits.

https://doi.org/10.7554/eLife.112029.1.sa1

## Reviewer #2 (Public review):

In this work, Alamdari et al. present EvoDiff, which provides the capability to generate protein sequences directly in sequence space, using a discrete diffusion model. There are several versions. EvoDiff-seq is trained on UniRef50 sequences (~42 million), and EvoDiff-MSDA operates instead by using sequence alignment methods to generate new members of protein families. The authors demonstrate many modes of sequence generation, including unconditional and conditional, inpainting of disordered regions, and also generating scaffolding of functional motifs. Their evaluation is also multifaceted, covering foldability, folding self-consistency, language embeddings, secondary structure distributions, and experiments for a set of different scenarios.

The paper has many notable strengths. It is comprehensive in breadth, and the experimental component is distinctive, although I am not personally suited to review the rigor of that element.

I would suggest that the paper's results certainly support the conclusion that order-agnostic sequence generation can yield useful candidates for multiple conditional design tasks. I am

not totally convinced that it necessarily establishes diffusion as a generally superior approach to other competitors, like the conventional protein language models- EvoDiff is certainly competitive, and I think that the demonstration of diffusion is nice. I also am not sure that it is fair to say that sequence alone is sufficient for the broad design capabilities claimed (other than the "in principle" statement).

I am overall quite supportive of the work and its demonstration, but I have a few comments for consideration in any revision.

(1) I did not work through all dates of everything, but it appears to me that there are several recent conceptual and methodological competitors. These include DPLM and ProtBFN - both of these seem to be after the first preprint of EvoDiff, but given the time gap, there probably deserves to be some additional discussion or comparison. I would say, ideally, they should offer direct benchmarking. If the authors are disinclined, then I would think they should just temper their claims of contemporary SOTA performance or general superiority. Instead, the paper would still remain valuable as an early and experimentally demonstrated sequence diffusion framework. I don't think it needs to be more than that.

(2) Related to the above, the manuscript should more carefully distinguish the demonstrated advantage of order-agnostic generation from the quality of unconditional generation. Regarding Figure 3, the authors argue that Evodiff's diffusion objective is necessary, but this does not seem to account for or address the LRAR baselines in Tables S1 and S3. Unless I am misunderstanding, the 640M LRAR model exhibits several better scores. The authors later suggest that EvoDiff's principal advantage is conditioning on arbitrary positions, which is valid, but that's a little different that what is being claimed. I suggest that the LRAR results should be shown or discussed alongside Figure 3, and then the authors revise to say that they have flexible conditional generation as the principal empirical benefit.

(3) I really like the IDR experiment, but I'm not sure it demonstrates that Evodiff can design functional IDRS generally. Cox15 is a favorable target because its mature sequence strongly identifies a conserved mitochondrial protein, and EvoDiff-MSA is supplied directly with its orthologous family. The eight tested sequences were also selected from hundreds of candidates using both DR-BERT and MitoFates, making the experiment a test of the full generation-and-prediction pipeline rather than of EvoDiff alone. Moreover, mitochondrial targeting is tested, but the disordered character of the generated sequences is not experimentally established. In any case, I think it would certainly be more convincing if there were other examples, with unrelated proteins or IDR functions. I would appreciate that this is again a step beyond what the authors might be compelled to do, but their claim could be simply more calibrated.

(4) The authors might benefit from explaining the advantages or complementarity of EvoDiff to other property-directed approaches for exploring sequence space. This has been done, for example, by using Bayesian optimization and genetic algorithms to tune properties of IDP condensates (DOI: 10.1021/acs.jpcb.8b03822). In my understanding, these are addressing a different problem from EvoDiff by optimizing sequences explicitly towards physical targets, while EvoDiff is a generative framework that can be used for sampling/inpainting/ etc. Is it clear how these strategies might be plausibly integrated? If so, that would be a relevant point of discussion and a potential advantage for EvoDiff.

https://doi.org/10.7554/eLife.112029.1.sa0