# Conditional Generation of Protein Sequence and Structure

Sidney Lyayuga Lisanza

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

David Baker, Chair

Neil King

Phil Bradley

Program Authorized to Offer Degree:

Biochemistry

University of Washington

## Abstract

Conditional Generation of Protein Sequence and Structure

Sidney Lyayuga Lisanza

Chair of the Supervisory Committee:
Dr. David Baker
Biochemistry

The advent of atomic accuracy protein sequence structure prediction with deep learning networks spurred by AlphaFold has had a remarkable impact on the field of biochemistry. It has resulted in rapid progress in protein design because it allows for the quick interrogation of structural hypotheses without the need to acquire experimental data which is expensive and time consuming. However, it remains elusive how to properly use these deep learning models for the generation of protein sequences and structures with user defined functional and biochemical properties. This was the focus of my dissertation work. I first interrogated this question by taking pre trained structure prediction networks, namely RoseTTAFold, and applying techniques from image processing to make them generative in a method termed “constrained Hallucination”. I apply the technique to optimize sequences such that their predicted structures contain desired functional sites on a slew of design problems ranging from epitope scaffolding, metal binding, to protein binding. Experimentally characterization of these designs demonstrate the have the

desired activities. In follow up work, I improve upon joint sequence-structure generation by employing the denoising diffusion probabilistic framework popularized in image generation. I developed ProteinGenerator, a sequence space diffusion model based on RoseTTAfold that simultaneously generates protein sequences and structures. Beginning from random amino acid sequences, the model generates sequence and structure pairs by iterative denoising, guided by any desired sequence and structural protein attributes. To explore the versatility of this approach, I designed and tested proteins enriched for specific amino acids, with internal sequence repeats, with masked bioactive peptides, with state dependent structures, and with key sequence features of specific protein families. And lastly looking to the future, particularly difficult protein design problems such as the design of highly active enzymes, experimental data feedback is necessary to improve functionality with minimal design iterations. Active learning (AL) and bayesian optimization (BO) approaches provide a principled way to incorporate experimental feedback into the design process, and subsequently minimize the number of iterations cycling between computation and experimental testing to optimize the desired function. However, these approaches do not incorporate strong generative priors to bias exploration/exploitations to valid regions of protein space. Therefore to improve upon current BO and AL methods, I hypothesize that coupling a joint sequence and structure diffusion model with bayesian optimization methods will allow for the more efficient search of the sequence activity landscape to find highly active variants. To this end I developed a joint sequence and structure denoising generative model, ProteinGenerator2 (PG2), to which I bias generation with both zero shot predictors to yield predicted highly active and diverse sequence pools for testing.

# TABLE OF CONTENTS

# List of Figures

# List of Tables

# ACKNOWLEDGEMENTS

I would like to thank everyone who has supported me during the course of my graduate work. First and foremost thank you to my family: my mother, Esther Lisanza, my father, Leonard Muaka, and my sisters Vivian and Lilly Lisanza for allowing me to grow in a nurturing and loving environment. Thank you to my undergraduate mentors Mihai Azoitei and Una Natterman for giving me the confidence and encouragement to pursue my Ph.D. Thank you to my excellent scientific collaborators, Jake Gershon, Jue Wang, Doug Tischer, Sergey Ovchinnikov, Sam Tipps, Lucas Arnoldt, and all of my colleagues at the Institute for Protein Design. Thank you to my advisor David Baker for allowing me to work in his group and providing crucial feedback when necessary. Thank you to my dearest friends Florence Dou, Jordan Drew, Barbara Reynolds, Jeremiah Sims, Sanaa Monsoor, Raphael Williams, Michael Kiflezghi, Tafari Clarke-James, and Harley Pyles. I would like to thank the Graduate Opportunities & Minority Achievement Program, and the Black Graduate Student Association for making my time at a predominately white institution as welcoming as possible.

# DEDICATION

I dedicate this work to my sisters Vivian and Lilly, thank you for being so loving and supporting.

I’m excited to watch you both grow and accomplish great things.

# Chapter 1. Introduction

Proteins are molecular machines that perform a myriad of functions within living organisms. Their function is determined by their three-dimensional structure and is entirely encoded by their amino acid sequence. A long standing problem in biology is predicting a protein's structure just from its amino acid sequence alone. Being able to do this will not only allow for direct prediction of function from sequence alone, but would also present unparalleled opportunities to be able to design de novo proteins with novel functions. Deep learning models, within the past few years spurred by the introduction of AlphaFold 2[1] have made remarkable strides in structure prediction. But the question remains how to properly leverage these structure prediction networks for guided de novo protein generation, which is the focus of this dissertation.

## 1.1 Structure Prediction

Classical structure prediction methods relied on two main approaches: template and template free modeling[2]. Template modeling requires a search for homologous structures, which function as initial guesses, to which further refinement is done to account for mutations, insertion, and deletions in the target-template alignment. Template free modeling requires the construction of a multiple sequence alignment (MSA), from which pairwise contacts can be inferred from correlated mutations. A rough structure of the protein emerges that can then be used as a starting point for gradient based minimization. Template free modeling lends itself well to more obscure structures without close homologs.

State of the art structure prediction networks, inspired by AlphaFold, leverage these two approaches in their predictions. For most proteins they require structural templates or MSAs. They use this initial information to develop a coarse grained picture of the protein's globular

structure through pairwise contact prediction between residues. In the case of MSA based predictions, this is done through the attention mechanism borrowed from natural language processing (NLP) work[3] that learns to attend sequence positions to other positions that are in contact in the three-dimensional structure. From there, they have equivariant structural modules that do iterative refinement of the structure to get the fine atomic details[1].

There are currently two main types of structure prediction networks: MSA and protein language model (pLM) based networks. MSA based networks, as the name suggests, require MSAs to be pre calculated prior to prediction[1]. Whereas pLM methods require the training of a pLM on massive amounts of sequence only data through the mask language modeling task to generate an embedding space for protein sequences which should hold some general notions of the "semantics" of proteins in an organized way[4]. Although the hope was that pLM structure prediction networks would pave the way for true single sequence structure prediction, it was found that they work the best for sequences with strong evolutionary sequences, and not as well for orphan proteins[4].

## 1.2 Classical De Novo Protein Design

De novo protein design has followed one basic paradigm. First the generation of a structure, followed by the design of a sequence that fits that structure. In classical protein design the structure was initially generated with fragment based approaches, where ideal fragments (well defined secondary structures and short loop regions with energetically favorable dihedrals) are assembled following some blueprint[5]. After a satisfactory solution was found for that backbone, it is fed as input to a design algorithm to search for satisfactory sequences with Markov Chain Monte Carlo (MCMC), minimizing an energy function, with minimal perturbation to the backbone[6]. This approach has been applied to the design of de novo binders[7], enzymes[8],

vaccines[9], and nanomaterials[10]. However these classical approaches are compute intensive and require thousands of trajectories to find satisfactory solutions.

## 1.3 Generative De Novo Design

Generative protein design allows for the more efficient search of either structural or sequence space to find optima by first learning to approximate a target distribution and then sampling it based on some priors. Significant progress was first made within the inverse folding problem, where the given prior is the backbone coordinates of a protein, and the model is tasked with generating a sequence that encodes for it. The first models used convolutional neural networks (CNN) with autoregressive decoding[11], whereas more recent state of the art models now use graph neural networks[12] (GNNs), while maintaining autoregressive decoding. There has been further work on extending these GNNs with large language models, but there seems to be minor boosts if any in performance[13].

Initial attempts at machine learning assisted backbone generation, relied on sequence sampling with methods like MCMC and subsequent prediction with structure prediction networks in a manner akin to activation maximization in images[14] where some loss criterion is optimized. In the first iteration of this approach the loss was simply increasing the kullback-leibler (KL) divergence between the sampled sequence outputs and randomly generated sequence outputs. This was first done with trRosetta[15], but the method was quickly improved with more powerful structure prediction networks, namely AlphaFold[16], RosettaFold[17], and ESMfold[18]. In addition gradient based optimization replaced MCMC for more efficient searches of sequence space. Furthermore, more problem specific losses were applied to generate sequence-structure pairs with tailored attributes for motif scaffolding or binder design[17]. This line of work is the focus of the second chapter of this thesis.

Interestingly, applying these activation maximization techniques to MSA based structure prediction networks often resulted in adversarial sequences[16]. The given structure prediction network used to design the proteins would predict the designs to have high confidence, but come experimental testing the designs would often not express or aggregate. This could be circumvented by filtering with an orthogonal structure prediction network, or redesigning the generated backbone with an inverse folding model. However this was a mode of failure exclusive to MSA based structure prediction models, for pLM based methods, model confidence did in fact correlate with experimental success[19].

A likely reason for this is that MSA based structure prediction models are trained to approximate the distribution of structure given sequence, p(structure | sequence). They do not have an explicit understanding of p(sequence), likewise they are prone to giving adversarial sequences, while still providing valid structures. Where as in the pLM case, it is a more explicit sequence-structure generative model with the model pretraining actually learning: p(sequence, structure ) = p(structure |sequence)p(sequence), where p(sequence) arises from masked language model training, and p(structure | sequence) arises from training a small linear layer to take a projection of the attention between two positions in the protein sequence and outputting a distribution over pairwise distances[19]. The third chapter of this thesis focuses on giving MSA based structure prediction methods an explicit understanding of p(sequence) through the diffusion generative paradigm[20] improving upon previous co-sequence structure generation methods[21].

# Chapter 2. Scaffolding Protein Functional Sites Using Deep Learning

This section contains content previously published as: Wang, J., **Lisanza, S**., *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).

## 2.1 Abstract

The binding and catalytic functions of proteins are generally mediated by a small number of functional residues held in place by the overall protein structure. Here, we describe deep learning approaches for scaffolding such functional sites without needing to prespecify the fold or secondary structure of the scaffold. The first approach, "constrained hallucination," optimizes sequences such that their predicted structures contain the desired functional site. The second approach, "inpainting," starts from the functional site and fills in additional sequence and structure to create a viable protein scaffold in a single forward pass through a specifically trained RoseTTAFold network. We use these two methods to design candidate immunogens, receptor traps, metalloproteins, enzymes, and protein-binding proteins and validate the designs using a combination of in silico and experimental tests.

## 2.2 Introduction

The biochemical functions of proteins are often carried out by a subset of residues that constitute a functional site—for example, an enzyme active site or a protein or small-molecule binding site—and hence the design of proteins with new functions can be divided into two steps. The first step is to identify functional site geometries and amino acid identities that produce the desired activity—for enzymes, this can be done using quantum chemistry calculations[22–24], and

for protein binders, by fragment docking calculations[7,25]. Alternatively, functional sites can be extracted from a native protein having the desired activity[26,27]. Here, we focus on the second step: Given a functional site description from any source, design an amino acid sequence that folds up to a three-dimensional (3D) structure containing the site. Previous methods can scaffold functional sites made up of one or two contiguous chain segments [26–30], but, with the exception of helical bundles[28], these do not extend readily to more complex sites composed of three or more chain segments, and the generated backbones are not guaranteed to be designable (i.e., encodable by some amino acid sequence).

An ideal method for functional de novo protein design would (i) embed the functional site with minimal distortion in a designable scaffold protein; (ii) be applicable to arbitrary site geometries, searching over all possible scaffold topologies and secondary structure compositions for those optimal for harboring the specified site; and (iii) jointly generate backbone structure and amino acid sequence. We previously demonstrated that the trRosetta structure-prediction neural network[31] can be used to generate new proteins by maximizing the trRosetta output probability that a sequence folds to some (unspecified) 3D structure during Monte Carlo sampling in sequence space[15]. We refer to this process as "hallucination," as it produces solutions that the network considers to be ideal proteins but that do not correspond to any known natural protein; crystal and nuclear magnetic resonance structures confirm that the hallucinated sequences fold to the hallucinated structures[15]. trRosetta can also be used to design sequences that fold into a target backbone structure by carrying out sequence optimization using a structure recapitulation loss function that rewards similarity of the predicted structure to the target structure[32]. Given this ability to design both sequence and structure, we reasoned that trRosetta could be adapted to tackle the functional site scaffolding problem.

## 2.3 Partition Constrained Hallucination Using a Multiobjective Loss Function

To extend existing trRosetta-based design methods to scaffold functional sites (Fig. 2.1A), we optimized amino acid sequences for folding to a structure containing the desired functional site using a composite loss function that combines the previously used hallucination loss with a motif reconstruction loss over the functional motif [rather than the entire structure, as in ([32])] (Fig. 2.1B; see materials and methods in the supplementary materials). Although we succeeded in generating structures with segments closely recapitulating functional sites, Rosetta structure predictions suggested that the sequences poorly encoded the structures (fig. 2.S1A), and hence we used Rosetta design calculations to generate more-optimal sequences[33]. Several designs targeting programmed cell death ligand 1 (PD-L1) generated by constrained hallucination with binding motifs derived from programmed cell death protein 1 (PD-1) (table S1)[34], followed by Rosetta design, were found to have binding affinities in the mid-nanomolar range (fig. 2.S1, B to E). Although this experimental validation is encouraging, the requirement for sequence design using Rosetta is inconsistent with the aim of jointly designing sequence and structure.

Following the development of RoseTTAFold (RF)[35], we found that it performed better than trRosetta in guiding protein design by functional site–constrained hallucination (fig. 2.S1G), likely reflecting the better overall modeling of protein sequence-structure relationships. Constrained hallucination with RoseTTAFold has the further advantages that, because 3D coordinates are explicitly modeled (trRosetta only generates inter-residue distances and orientations), site recapitulation can be assessed at the coordinate level and additional problem-specific loss terms can be implemented in coordinate space that assess interactions with a target (fig. 2.S2; materials and methods).

## 2.4 Generalized Functional Motif Scaffolding by Missing Information Recovery

While powerful and general, the constrained hallucination approach is compute-intensive, as a forward and backward pass through the network is required for each gradient descent step during sequence optimization. In the training of recent versions of RoseTTAFold, a subset of positions in the input multiple sequence alignment are masked, and the network is trained to recover this missing sequence information in addition to predicting structure. This ability to recover both sequence and structural information provides a second solution to the functional site scaffolding problem: Given a functional site description, a forward pass through the network can be used to complete, or "inpaint," both protein sequence and structure in a masked region of protein (Fig. 2.1C; materials and methods). Here, the design challenge is formulated as an information recovery problem, analogous to the completion of a sentence given its first few words using language models[36] or the completion of corrupted images using inpainting[37]. A wide variety of protein structure prediction and design challenges can be similarly formulated as missing information recovery problems (Fig. 2.1D). Although protein inpainting has been explored before[38,39], in this study we approach it using the power of a pretrained structure-prediction network.

We began from a RoseTTAFold (RF) model trained for structure prediction and carried out further training on fixed-backbone sequence design in addition to the standard fixed-sequence structure prediction task to avoid model degradation (fig. 2.S3; materials and methods). This model, denoted RFimplicit, was able to recover small, contiguous regions missing both sequence and structure (fig. 2.S3). Encouraged by this result, we trained a model explicitly on inpainting segments with missing sequence and structure given the surrounding

protein context, in addition to sequence design and structure prediction tasks (fig. 2.S4A; materials and methods and algorithm 2.S1). The resulting model was able to inpaint missing regions with high fidelity (Fig. 2.1E and fig. 2.S4) and performed well at sequence design (32% native sequence recovery during training) and structure prediction (fig. 2.S4C). We call this network RFjoint and use it to generate all inpainted designs below unless otherwise noted.

To evaluate in silico the quality of designs generated by our methods, we use the AlphaFold (AF) protein structure prediction network[1], which has high accuracy on de novo designed proteins[40] (fig. 2.S7A). RF and AF have different architectures and were trained independently, and hence AF predictions can be regarded as a partially orthogonal in silico test of whether RF-designed sequences fold into the intended structures, analogous to traditional ab initio folding[32,41]. We used AF to compare the ability of hallucination and inpainting to rebuild missing protein regions (Fig. 2.1, F and G, and fig. 2.S5). Inpainting yielded solutions with more accurately predicted fixed regions ("AF-RMSD"; Fig. 2.1G and fig. 2.S5B) and structures overall more confidently predicted from their amino acid sequences ("AF pLDDT"; Fig. 1F and fig. 2.S5A) and required only 1 to 10 s per design on an NVIDIA RTX 2080 graphics processing unit (hallucination requires 5 to 20 min per design). However, hallucination gave better results when the missing region was large (fig. 2.S5) and generated greater structural diversity (fig. 2.S8; and see below).

In the following sections, we highlight the power of the constrained hallucination and inpainting methods by designing proteins containing a wide range of functional motifs (Figs. 2.2 to 5 and table 2.S1). For almost all problems, we obtained designs that are closely recapitulated by AF with overall and motif (functional site) root mean square deviation (RMSD) of typically $<2$ and $<1$ Å, respectively, with high model confidence [predicted local distance difference test

(pLDDT) > 80; table 2.S2]; such recapitulation suggests that the designed sequences encode the designed structures [although it should be noted that AF has limited ability to predict protein stability[42] or mutational effects[43,44]]. More critically, we assessed the activities of the designs experimentally (with the exception of those labeled "in silico" in Figs. 2.2 to 2.5).

## 2.5 Designing Immunogen Candidates and Receptor Traps

The goal of immunogen design is to scaffold a native epitope recognized by a neutralizing antibody as accurately as possible in order to elicit antibodies binding the native protein upon immunization. Additional interactions with the antibody are undesirable because the aim is to elicit antibodies recognizing only the original antigen, and hence for hallucination, we add a repulsive loss term to penalize interactions with the antibody beyond those present in the scaffolded epitope (fig. 2.S2; supplementary text). As a test case, we focused on respiratory syncytial virus F protein (RSV-F), which has several antigenic epitopes for which structures with neutralizing antibodies have been determined[26,29,30]. We scaffolded RSV-F site II, a 24-residue helix-loop-helix motif that had previously been grafted successfully onto a three-helix bundle[26], as well as RSV-F site V, a 19-residue helix-loop-strand motif that has not yet been scaffolded successfully[45]. We were able to hallucinate designs recapitulating both epitopes to sub-angstrom backbone RMSD in a variety of folds [Fig. 2.2A and fig. 2.S9; structures and sequences for all designs below are given in data S1 and S2 and differ considerably from native proteins (table S2); RF hallucinated models and AF structure predictions are shown in figs. 2.S9, 2.S11, and 2.S17; only the AF model is shown in the main figures]. Inpainting also generated scaffolds for RSV-F site V, with comparable quality but less diversity than the hallucinations (fig. 2.S8).

We expressed 37 hallucinated RSV-F site V scaffolds with high AF pLDDT and low motif AF-RMSD in *Escherichia coli* and found that three bound the neutralizing antibody

hRSV90[45] with a dissociation constant (*K*d) of 0.9 to 1.3 μM (Fig. 2.2C and fig. 2.S11; materials and methods and supplementary text). The *K*d for the RSVF trimer is lower (23 nM), but the interface is larger, encompassing both sites II and V[45]. Mutation of either of two key epitope residues reduced or abolished binding of the designs, suggesting that they bind the target through the scaffolded motif (Fig. 2.2C and fig. 2.S11A), and circular dichroism (CD) spectra were consistent with the designed scaffold structures for both the original hallucinations (Fig. 2.2D) and the epitope mutants (fig. 2.S11C). Four of the inpainted designs bound hRSV90 by yeast display but were poorly expressed in *E. coli* (fig. 2.S11, C to E). Overall, the designs provide a diverse set of promising starting points for further RSV-F epitope-based vaccine development.

We next applied hallucination to the in silico design of receptor traps that neutralize viruses by mimicking their natural binding targets and thus are inherently robust against mutational escape. We again augmented the loss function with a penalty on interactions beyond those in the native receptor to avoid opportunities for viral escape. As a test case, we scaffolded the helix of human angiotensin-converting enzyme 2 (hACE2) interacting with the receptor binding domain of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike protein[46]. The hallucinated hACE2 mimetics have a diverse set of helical topologies, and AF structure predictions recapitulate the binding interface with sub-angstrom accuracy (Fig. 2.2B and fig. 2.S9C).

## 2.6 Designing Metal-coordinating Proteins

Di-iron sites are important in biological systems for iron storage[47] and can mediate catalysis[48,49]. We were able to recapitulate the di-iron site from *E. coli* bacterioferritin, composed of four parallel helical segments, to sub-angstrom AF-RMSD using both inpainting (Fig. 2.3, A to E, and fig. 2.S13) and hallucination (fig. 2.S12; the hallucinations were not tested owing to

buried polar residues; supplementary text). The designs had diverse helix connectivities and low structural similarity to the parent [figs. 2.S13B and 2.S12; template modeling (TM)–score 0.55 to 0.71 to PDB ID 1BCF_A]. We chose 96 inpainted designs to test experimentally and found that 76 had soluble expression, at least eight (see supplementary text) had a spectroscopic shift indicative of Co2+ binding (a proxy for iron binding)[50,51], and three (dife_inp_1, dife_inp_2, and dife_inp_3; Fig. 2.3B and fig. 2.S13E) had CD spectra consistent with the designed fold (Fig. 2.3D and fig. 2.S13F) and were stabilized by metal binding (Fig. 2.3E and fig. 2.S13G). Mutation of the metal binding residues abolished binding (Fig. 2.3B and fig. 2.S13E), and titration analysis of dife_inp_1 suggested that both metal binding sites were successfully scaffolded (Fig. 2.3C).

We next scaffolded the calcium-binding EF-hand motif[52], a 12-residue loop flanked by helices. Both constrained hallucination and inpainting readily generated scaffolds recapitulating either one or two EF-hand motifs to within 1.0 Å AF-RMSD of the native motif (Fig. 2.3F; fig. 2.S14, A and B; and table S2). We chose 20 hallucinations and 55 inpaints to display on yeast and screen for calcium binding using tryptophan-enhanced terbium fluorescence[53]. Six hallucinations and four inpaintings had fluorescence consistent with ion binding [fig. 2.S14A; materials and methods; one of these proteins (EFhand_inp_2) was designed using RFimplicit (supplementary text)]. The top hit from yeast, the inpainted EFhand_inp_1, purified from *E. coli* as a monomer (fig. 2.S14C), had the expected CD spectrum (Fig. 2.3G) and a clear terbium binding signal (Fig. 2.3H) that was eliminated by CaCl2 competition (Fig. 2.3H).

## 2.7 In Silico Design of Enzyme Active Sites

We next sought to scaffold the active site of carbonic anhydrase II, which catalyzes the interconversion of carbon dioxide and bicarbonate and has recently been of interest for carbon

sequestration[49–51]. The active site consists of three Zn2+-coordinating histidines on two strands and a threonine on a loop, which orients the CO2 (table S1). Despite the complexity of the irregular, discontinuous three-segment site, hallucination was able to generate designs with sub-angstrom motif AF-RMSDs with correct His placement for Zn2+ coordination (Fig. 2.4A and fig. 2.S9D); these are less than 100 residues in size, considerably smaller than the 261-residue native protein.

We next scaffolded the catalytic side chains of Δ5-3-ketosteroid isomerase (KSI) (table S1) involved in steroid hormone biosynthesis[54]. We attempted to use gradient descent by backpropagation through AF (materials and methods; a side chain–predicting version of RF was not available at the time) but found it difficult to obtain accurate side-chain placement; the landscape may be too rugged with the high-resolution side chain–based loss (supplementary text). Better results were obtained with a two-stage approach using, first, both AF and trRosetta (to smoothen the loss landscape) and a description of the active site at the backbone level, followed by a second all-atom AF-only stage once the overall backbone was roughly in place. This yielded multiple plausible solutions with nearly exact matches to the catalytic side-chain geometry (Fig. 2.4, C and D, and fig. 2.S9E). In silico validation with a held-out AF model (materials and methods) recapitulated the designed active sites. The use of stage-specific loss functions illustrates the ready customizability of the hallucination approach to specific design challen

## 2.8 Designing Protein-Binding Proteins

To design binders to the cancer checkpoint protein PD-L1, we scaffolded two discontiguous segments of the interfacial β sheet from a high-affinity mutant of PD-1 (Fig. 2.5A; materials and methods)[34]. Inpainting yielded designs with not only good AF predictions of the

binder monomer (AF pLDDT > 80, motif AF-RMSD < 1.4 Å) but also of the complex between the binder and PD-L1, with an interchain predicted alignment error (inter-PAE) of <10 Å (materials and methods). In contrast to our initial efforts with trRosetta hallucination (fig. 2.S1; supplementary text), it was not necessary to redesign the inpainted sequences using Rosetta. Of 31 designs selected for experimental testing, one design, pdl1_inp_1, bound PD-L1 with a *K*d of 326 nM (Fig. 5, B and C), worse than high-affinity consensus (HAC) PD-1 (*K*d = 110 pM)[55] but better than wild-type PD-1 (*K*d = 3.9 μM)[55]. The pdl1_inp_1 design expressed as a monomer (fig. 2.S15E), was thermostable, and had a CD spectrum consistent with that of a mixed α-β fold (fig. 2.S15F). Unlike native PD-1, which has an immunoglobulin family β-sandwich fold, pdl1_inp_1 has two helices buttressing the interfacial β sheet, as well as an additional fifth inpainted strand extending the interface (fig. 2.S15, A and B). The closest Protein Data Bank (PDB)[56] hit had a TM-score of 0.61, and the closest Basic Local Alignment Search Tool (BLAST) NR hit had a sequence identity of 25.4%.

We next used our methods to design ligands engaging multiple receptor binding sites. The nerve growth factor (NGF) receptor TrkA dimerizes upon ligand binding[57], and starting from the TrkA-NGF crystal structure, we positioned helical segments derived from two copies of a previously designed TrkA binding protein[7] and used hallucination followed by inpainting (materials and methods) to scaffold them on a single chain (Fig. 2.5, D and E). A design predicted to be well structured (AF pLDDT > 80) and interact with TrkA (inter-PAE < 10 Å) was expressed, purified, and found to bind TrkA, as assessed by biolayer interferometry (BLI) (Fig. 2.5F). A double mutant that knocked out both designed binding sites abolished TrkA binding, whereas single mutants knocking out either one of the binding sites maintained partial binding (Fig. 2.5F and fig. 2.S16), suggesting that the protein binds two molecules of TrkA, as designed.

RoseTTAFold is able to predict the structures of protein complexes[58], and we hypothesized that it could generate additional binding interactions between hallucinated or inpainted binder and a target beyond the scaffolded motif. We used a "two-chain" hallucination protocol (fig. 2.S17; materials and methods) to design binders to the Mdm2 oncogene by scaffolding the native N-terminal helix of the tumor suppressor protein p53 and obtained diverse designs with AF inter-PAE < 7 Å, target-aligned binder RMSD < 5 Å, binder pLDDT > 85, and spatial aggregation propensity (SAP) score < 35 (fig. 2.S17, D and E); three examples are shown in Fig. 2.5G.

The above approaches to protein-binder design require starting from a previously known binding motif, but hallucination should in principle be able to generate de novo interfaces as well. To test this, we used two-chain hallucination to optimize 12-residue peptides for binding to 12 targets starting from random sequences, minimizing an interchain entropy loss (fig. 2.S17H). Most of the hallucinated peptides bound at native protein interaction sites (fig. 2.S18A); the remainder bound in hydrophobic grooves resembling protein binding sites (fig. 2.S18B). We used the same procedure to generate 55- to 80-residue binders against TrkA and PDL-1 without starting motif information and obtained designs predicted by AF to complex with the target, at the native ligand binding site, with a target-aligned binder RMSD < 5 Å and an inter-PAE < 10 Å (fig. 2.S17, F and G).

Unlike classical protein design pipelines, which treat backbone generation and sequence design as two separate problems, our methods simultaneously generate both sequence and structure, taking advantage of the ability of RoseTTAFold to reason over and jointly optimize both data types. This results in excellent performance in both generating protein backbones with a geometry capable of hosting a desired site and sequences that strongly encode these backbones.

Our hallucinated and inpainted backbones accommodate all of the tested functional sites much more accurately than any naturally occurring protein in the PDB or AF predictions database (fig. 2.S20 and table S3; supplementary text)[59], and our designed structures are predicted more confidently from their (single) sequences than most native proteins with known crystal structures and are on par with structurally validated de novo designed proteins (fig. 2.S7, A and B). The hallucination and inpainting approaches are complementary: Hallucination can generate diverse scaffolds for minimalist functional sites but is computationally expensive because it requires a forward and backward pass through the neural network to calculate gradients for each optimization step (materials and methods), whereas inpainting usually requires larger input motifs but is much less compute-intensive and outperforms the hallucination method when more starting information is provided. This difference in performance can be understood by considering the manifold in sequence-structure space corresponding to folded proteins. The inpainting approach can be viewed as projecting an incomplete input sequence-structure pair onto the subset of the manifold of folded proteins (as represented by RoseTTAFold) containing the functional site—if insufficient starting information is provided, this projection is not well determined, but with sufficient information, it produces protein-like solutions, updating sequence and structure information simultaneously. The loss function used in the hallucination approach is constructed with the goal that minima lie in the protein manifold, but there will likely not be a perfect correspondence, and hence stochastic optimization of the loss function in sequence space may not produce solutions that are as protein-like as those from the inpainting approach.

## 2.9 Conclusion

The approaches for scaffolding functional sites presented here require no inputs other than the structure and sequence of the desired functional site and, unlike previous methods, do

not require specifying the secondary structure or topology of the scaffold and can simultaneously generate both sequence and structure. Despite a recent surge of interest in using machine learning to design protein sequences[11,12,60–65], the design of protein structure is relatively underexplored, likely because of the difficulty of efficiently representing and learning structure[66]. Generative adversarial networks and variational autoencoders have been used to generate protein backbones for specific fold families[67–69], whereas our approach leverages the training of RoseTTAFold on the entire PDB to generate an almost unlimited diversity of new structures and enable the scaffolding of any desired constellation of functional residues. Our "activation maximization" hallucination approach extends related work in this area[70–72] by leveraging its key strength, the ability to use arbitrary loss functions tailored to specific problems and design any length sequence without retraining. The ability of our inpainting approach to expand from a given functional site to generate a coherent sequence-structure pair should find wide application in protein design because of its speed and generality. The two approaches individually, and the combination of the two, should increase in power as more-accurate protein structure, interface, and small-molecule binding prediction networks are developed.

## 2.10 Acknowledgements

We thank L. Goldschmidt and K. VanWormer, respectively, for maintaining the computational and wet lab resources at the Institute for Protein Design; C. Norn for general discussions about trRosetta; B. Coventry for advice on interface design; C. Goverde for advice on RSV-F epitopes and motif grafting methods; T. Yu, G. R. Lee, L. An, and X. Wang for advice on flow cytometry; R. Dong and V. Muhunthan for exploratory analyses; N. Hiranuma for exploratory RoseTTAFold training sessions; B. Trippe for feedback on the manuscript; S.

Pellock for expertise on enzyme design; A. Fitzgibbon for conceptual discussions on training RoseTTAFold; and C. Garcia for providing biotinylated TrkA.

### 2.10.1 Funding

We thank Microsoft for support and for providing Azure computing resources. This work was supported with funds provided by the Audacious Project at the Institute for Protein Design (D.B. and A.S.); a Microsoft gift (M.B. and J.D.); Eric and Wendy Schmidt by recommendation of the Schmidt Futures (D.J.); the DARPA Synergistic Discovery and Design project HR001117S0003 contract FA8750-17-C-0219 (D.B. and W.Y.); the DARPA Harnessing Enzymatic Activity for Lifesaving Remedies project HR001120S0052 contract HR0011-21-2-0012 (N.B.); the Washington Research Foundation (J.W.); the Open Philanthropy Project Improving Protein Design Fund (D.B. and D.T.); Amgen (S.L.); the Human Frontier Science Program Cross Disciplinary Fellowship (LT000395/2020-C) and EMBO Non-Stipendiary Fellowship (ALTF 1047-2019) (L.F.M.); the EMBO Fellowship (ALTF 191-2021) (T.S.); European Molecular Biology Organization Grant (ALTF 139-2018) (B.I.M.W.); the “la Caixa” Foundation (M.E.); the National Institute of Allergy and Infectious Diseases (NIAID) Federal Contract HHSN272201700059C (I.A.), NIH grant DP5OD026389 (S.O.); the National Science Foundation MCB 2032259 (S.O.); the Howard Hughes Medical Institute (D.B., R.R., and K.M.C.), the National Institute on Aging grant 5U19AG065156 (D.B., J.L.W., D.R.H., and M.E.); the National Cancer Institute grant R01CA240339 (D.B. and J.-H.C.); Swiss National Science Foundation (K.M.C. and B.C.); Swiss National Center of Competence for Molecular Systems Engineering (K.M.C. and B.C.); Swiss National Center of Competence in Chemical Biology (K.M.C. and B.C.); and European Research Council grant 716058 (K.M.C. and B.C.).

## 2.11 Main Figures

Figure 2.1: **Methods for protein function design.**

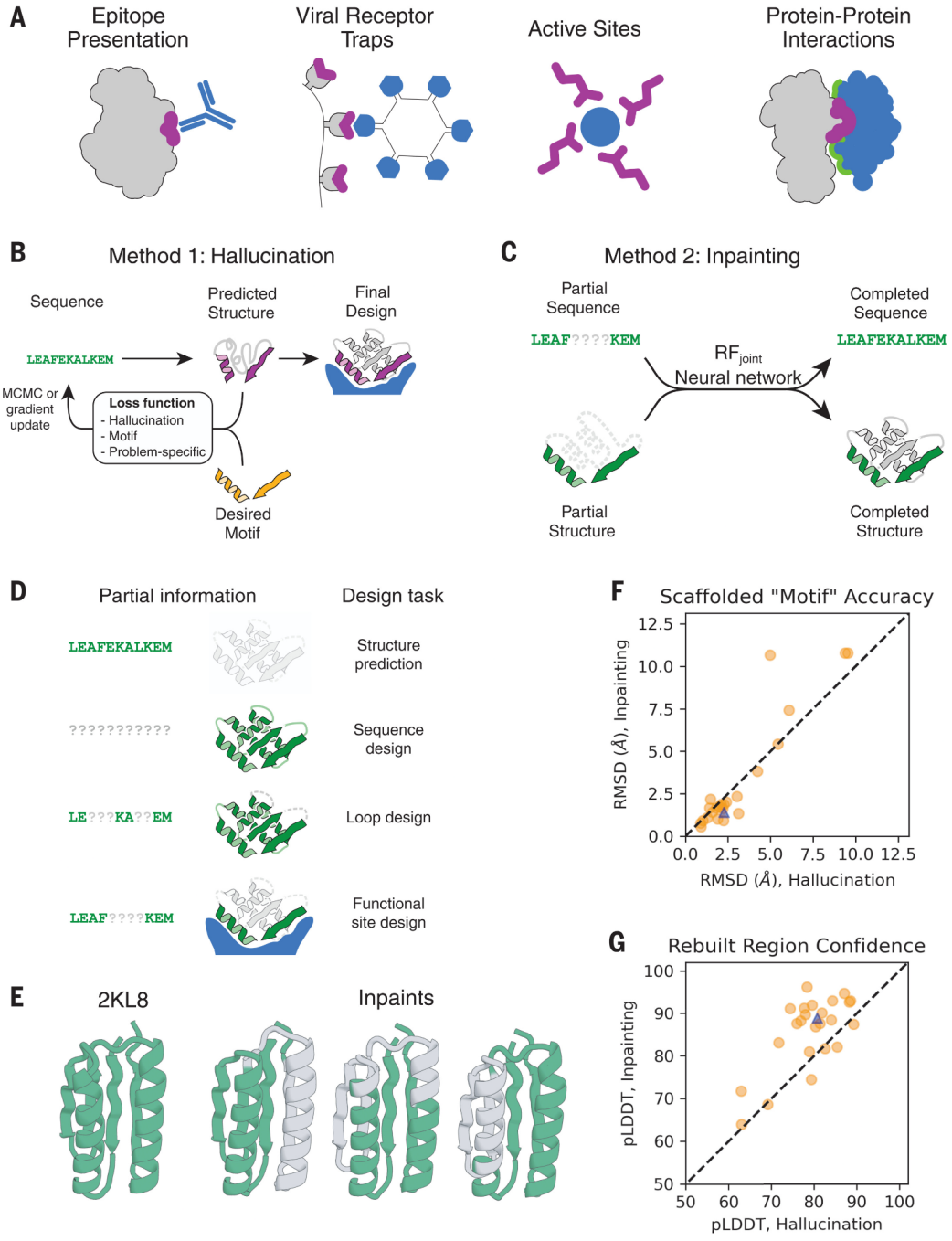



**Fig. 2.1. Methods for protein function design.** (**A**) Applications of functional-site scaffolding. (**B** and **C**) Design methods. (B) Constrained hallucination. At each iteration, a sequence is passed to the trRosetta or RoseTTAFold neural network, which predicts 3D coordinates and

inter-residue distances and orientations (fig. S2). The predictions are scored by a loss function that rewards certainty of the predicted structure along with motif recapitulation and other task-specific functions. MCMC, Markov chain Monte Carlo. (C) Missing information recovery ("inpainting"). Partial sequence and/or structural information is input into a modified RoseTTAFold network (called RFjoint), and complete sequence and structure are output. (**D**) Protein design challenges formulated as missing information recovery problems. Question marks in column 1 indicate missing sequence information; gray cartoons in column 2, missing structural information. (**E**) RFjoint can simultaneously recover structure and sequence of a masked protein region. 2KL8 was fed into RFjoint with a continuous (length 30) window of sequence and structure masked out, with the network tasked with predicting the missing region of protein. Outputs (inpainted region in gray) closely resemble the original protein (2KL8, left) and are confidently predicted by AlphaFold (pLDDT/motif RMSD of models shown, from left to right: 91.6/0.91, 92.0/0.69, and 90.4/0.82). (**F** and **G**) Motif scaffolding benchmarking data comparing RFjoint with constrained hallucination. A set of 28 de novo designed proteins, published since RoseTTAFold was trained, were used. For each protein, 20 random masks of length 30 were generated, and RFjoint and hallucination were tasked with filling in the missing sequence and structure to "scaffold" the unmasked "motif." For this mask length, RFjoint typically modestly outperforms hallucination, both in terms of the RMSD of the unmasked protein (the "motif") to the original structure (F) and in AlphaFold confidence (pLDDT in the replaced region) (G). Circles represent average of 20 outputs for each of the benchmarking proteins. Triangle represents 2KL8. Colors in all panels: native functional motif, orange; hallucinated/inpainted scaffold, gray; constrained motif, purple; binding partner, blue; nonmasked region, green; and masked region, light-gray dotted lines.

Figure 2.2: **Design of epitope scaffolds and receptor traps.**

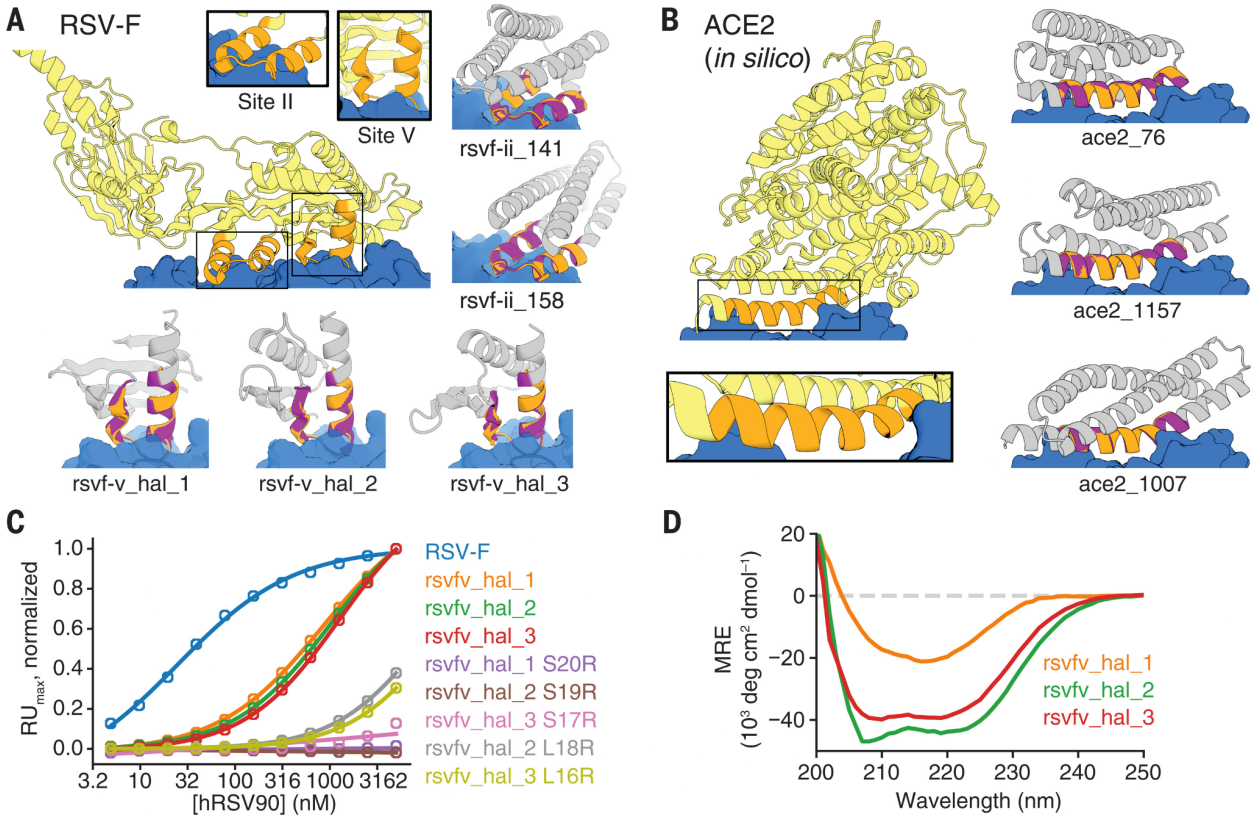



**Fig. 2.2. Design of epitope scaffolds and receptor traps.** (**A**) Design of proteins scaffolding immunogenic epitopes on RSV protein F (site II: PDB ID 3IXT chain P residues 254 to 277; site V: PDB ID 5TPN chain A residues 163 to 181). Comparisons of the RF hallucinated models to AF2 structure predictions from the design sequence are in fig. S9; here, because of space constraints, we show only the AF2 model (the two are very close in all cases). Here and in the following figures, we assess the extent of success in designing sequences that fold to structures harboring the desired motif through two metrics computed on the AF2 predictions: prediction confidence (AF pLDDT) and the accuracy of recapitulation of the original scaffolded motif (motif AF-RMSD). For RSV-F designs, these metrics are rsvf_ii_141 (85.0, 0.53 Å), rsvf_ii_158 (82.9, 0.51 Å), rsvf_ii_171 (88.4, 0.69 Å), rsvfv_hal_1 (82, 0.7 Å), rsvfv_hal_2 (88, 0.64 Å), and rsvfv_hal_3 (86, 0.65 Å). (**B**) Design of COVID-19 receptor trap based on ACE2 interface helix (PDB ID 6VW1 chain A residues 24 to 42). Design metrics: ace2_76 (89.1, 0.55 Å), ace2_1157 (80.4, 0.47 Å), and ace2_1007 (83.3, 0.57 Å). Colors: native protein scaffold, light yellow; native functional motif, orange; hallucinated scaffold, gray; hallucinated motif, purple; and binding partner, blue. See table S2 for additional metrics on each design. (**C**) Normalized maximum surface plasmon resonance signal (response units) of purified RSV-F epitope scaffolds and point mutants at various concentrations of hRSV90 antibody, with sigmoid fits. RSV-F refers to purified trimeric native F protein. *K*d values are as follows: RSV-F: 24 nM; rsvfv_hal_1: 0.9 μM; rsvfv_hal_2: 1.0 μM; rsvfv_hal_3: 1.3 μM. (**D**) Mean residue ellipticity (MRE) versus wavelength, from CD spectroscopy, for the three RSV-F site V hallucinations with binding activity.

Figure 2.3: **Design of metal binding.**

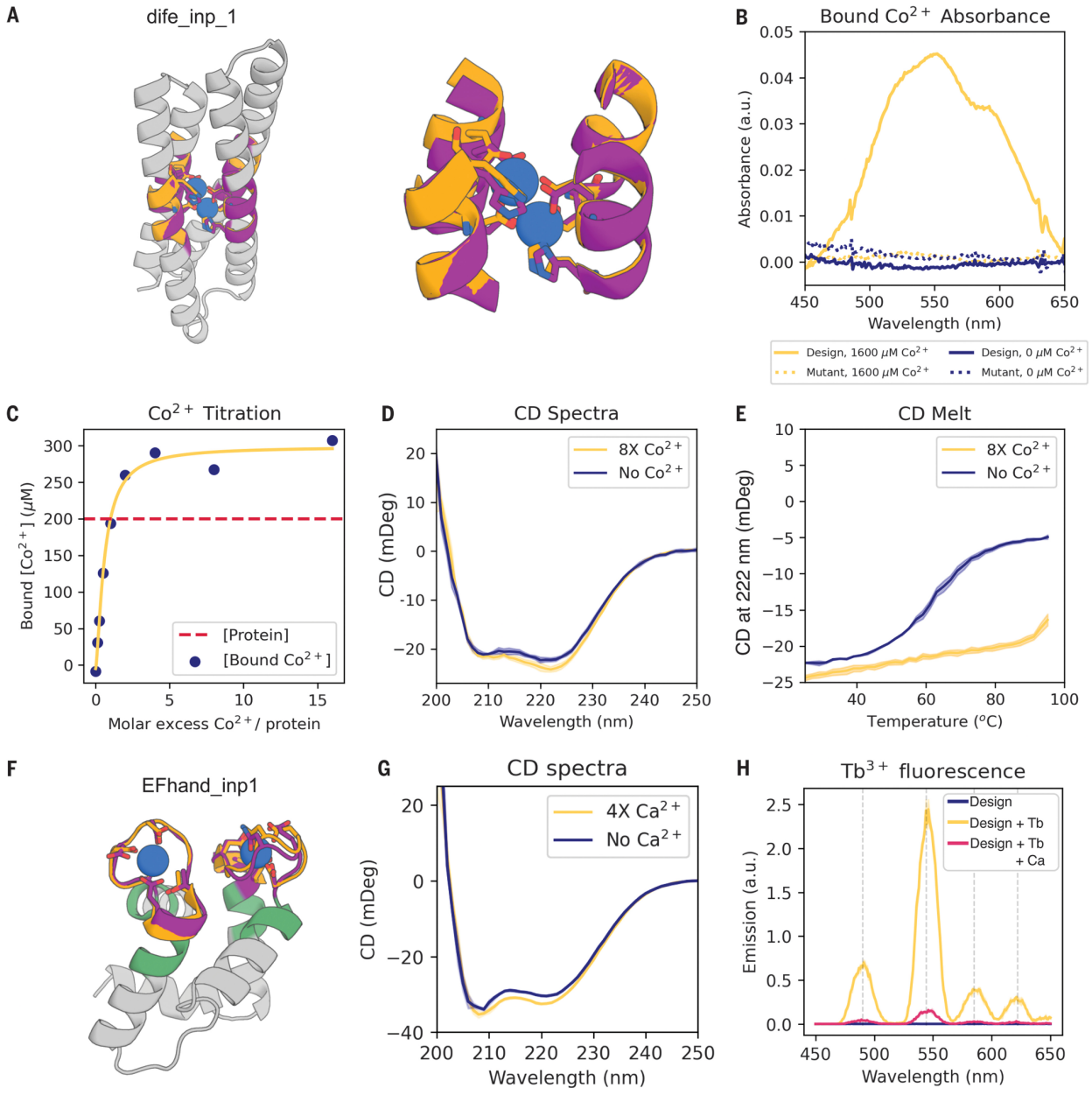



**Fig. 2.3. Design of metal binding.** (**A**) Scaffolding of di-iron binding site from *E. coli* cytochrome b1 (PDB ID 1BCF chain A residues 18 to 25, 27 to 54, 94 to 97, and 123 to 130) using inpainting. Colors: native protein scaffold, light yellow; native functional motif, orange; hallucinated scaffold, gray; hallucinated motif, purple; and bound metal, blue. (**B**) Absorbance spectra of dife_inp_1 (or mutant) in the presence (or absence) of an eight-fold molar excess of Co2+. Peaks at 520, 555, and 600 nm, consistent with Co2+ binding to the scaffolded motif (*32*).

In the mutant, the six coordinating residues [side chains shown in (A)] are mutated to alanine (E16A, E55A, H58A, E89A, H92A, E115A). Protein concentration: 200 μM. (**C**) dife_inp_1 Co2+ titration (protein concentration: 200 μM). Quantification of the absorbance at 550 nm, using a predicted extinction coefficient of 155 for Co2+ binding the motif (*32*), is consistent with both binding sites being recapitulated. (**D**) CD spectra of dife_inp_1 in the presence and absence of Co2+ are both consistent with the predicted helical structure. (**E**) Temperature dependence of dife_inp_1 CD signal in the presence and absence of Co2+. Coordination of Co2+ in the core stabilizes the protein. Protein concentration: 6.7 μM; Co2+ concentration: 53.3 μM. (**F**) Inpainted design EFhand_inp_1 scaffolding the double EF-hand motif with input motif residues in purple, input nonmotif residues in green, and overlaid with the native motif from PDB ID 1PRW (orange). (**G**) CD spectra of EFhand_inp_1 incubated with and without CaCl2 suggest stabilization of the protein upon binding calcium. (**H**) Tryptophan-enhanced terbium fluorescence spectra of EFhand_inp_1 suggests that the design binds terbium (*57*). Terbium binding signal is competed by 1 mM CaCl2 (red). Design metrics (AF pLDDT, motif AF-RMSD): dife_inp_1 (92, 0.65 Å) and EFhand_inp1 (84, 0.7 Å).

Figure 2.4: **In silico design of enzyme active sites.**

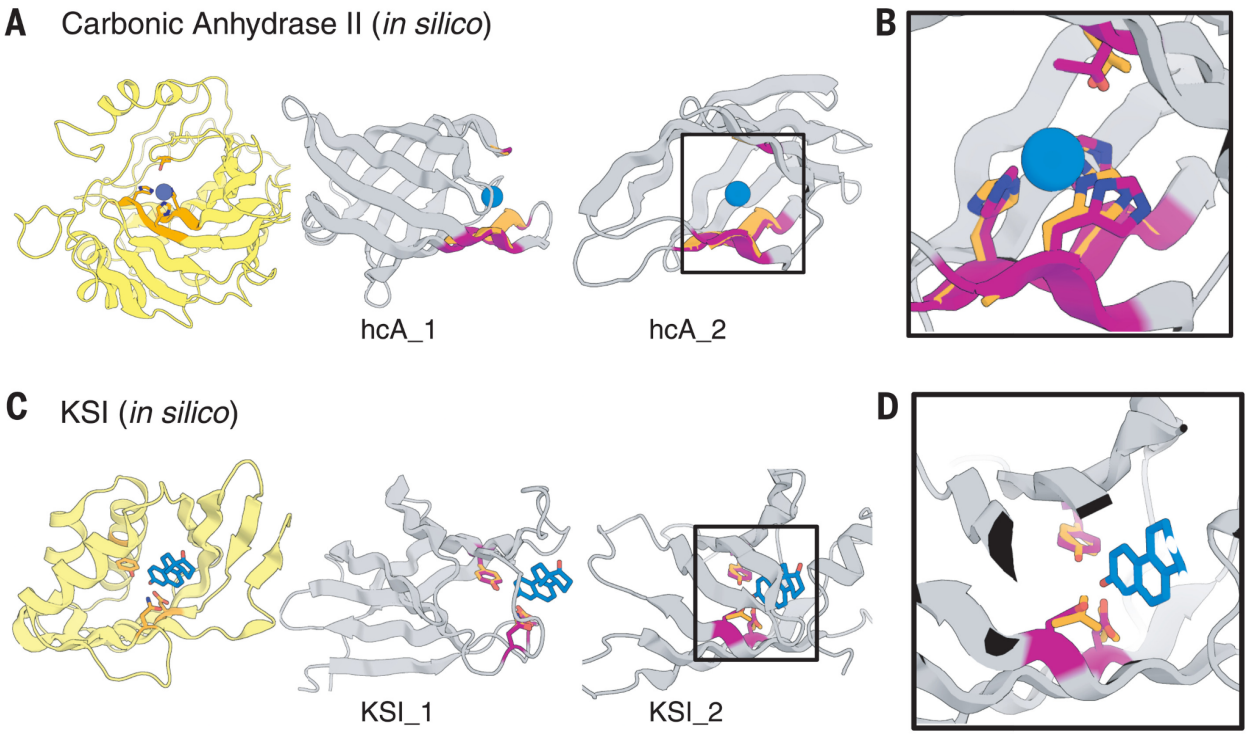



**Fig. 2.4. In silico design of enzyme active sites.** (**A** and **B**) Hallucinations using backbone description of site using RF. (**C** and **D**) Hallucination using side-chain description of site using AF2 augmented with trRosetta (materials and methods). (A) Carbonic anhydrase II active site (PDB ID 5YUI chain A residues 62 to 65, 93 to 97, and 118 to 120). (B) Δ5-3-ketosteroid isomerase active site (PDB ID 1QJG chain A residues 14, 38, and 99). Colors: native protein scaffold, light yellow; native functional motif, orange; hallucinated scaffold, gray; hallucinated

motif, purple; and bound metal, blue. [(B) and (D)] Zoomed-in view of designed active sites. Design metrics (AF pLDDT, motif AF-RMSD): hcA_1 (73, 1.04 Å), hcA_2 (71, 0.62 Å), KSI_1 (84, 0.30 Å Cβ), and KSI_2 (72, 0.53 Å Cβ).

Figure 2.5: **Design of protein-binding proteins.**

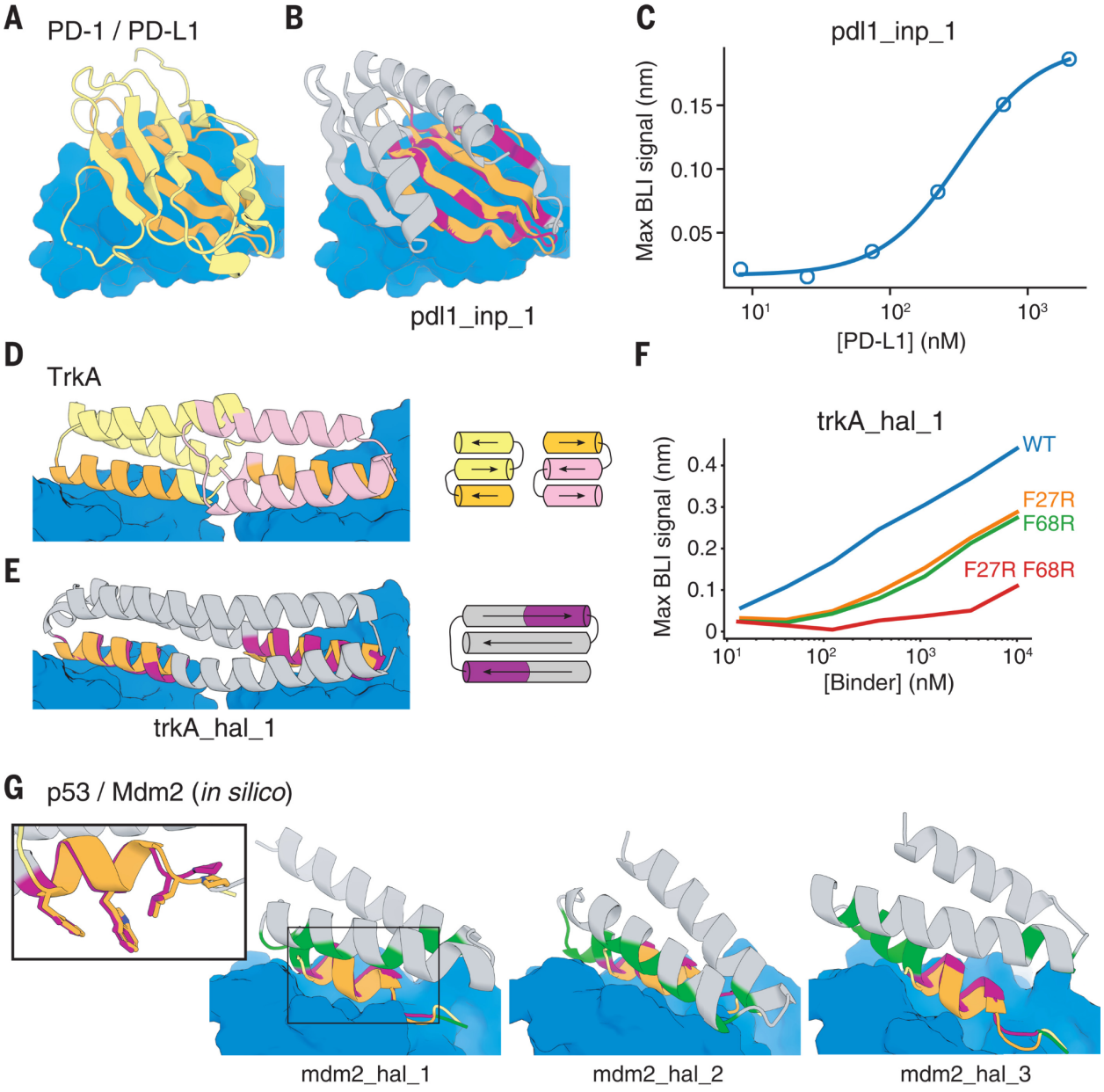



**Fig. 2.5. Design of protein-binding proteins.** Designs containing target-binding interfaces built around native-complex–derived binding motifs. Targets are in blue, native scaffolds in yellow or pink, native motifs in orange, designed scaffolds in gray, and designed motifs in purple. (**A**) Crystal structure of HAC PD-1 in complex with PD-L1. (**B**) Inpainted PD-L1 binder

superimposed on PD-1 interface motif. (**C**) BLI binding signal versus PD-L1 concentration. *K*d = 326 nM. (**D**) Crystal structure of previously designed TrkA minibinder in complex with TrkA, superimposed on TrkA receptor dimer. (**E**) Hallucinated bivalent TrkA binder. Protein topology diagrams are on the right. (**F**) BLI binding signal versus TrkA concentration; mutations at both scaffolded binding sites reduce TrkA binding. (**G**) Hallucinated Mdm2 binder designs superimposed on native p53 helix in complex with Mdm2 (see also fig. S17, D and E). New binding interactions (hallucinated residues within 5 Å of the target) are in green. (Inset) Overlay of mdm2_hal_1 and native p53 helix showing key side chains for binding.

## 2.12 Materials and Methods

### 2.12.1 Sequence Representation

For structure prediction, the input to trRosetta and RosettaFold is a tensor $X \in \mathbb{R}^{N \times L \times A}$ representing a one-hot-encoded multiple sequence alignment (MSA), where L is the sequence length, N is the number of aligned sequences, and $A = 21$ is the alphabet size (20 amino acids plus gap character, although gaps are never used during design). For design with RosettaFold, which was used for most of the designs in this paper, we optimized a single sequence ($N = 1$) and applied a 20% dropout, which is implemented at a variety of layers within the network. The first set of PD-1 mimetics (Fig. S1) were hallucinated with trRosetta and optimized a 1000-sequence MSA ($N = 1000$) with 0-20% dropout on input 2D features (14). Designing an MSA improves motif accuracy with trRosetta (13) but is not necessary when using RosettaFold. When residues on the functional motif are known to form desirable interactions with the binding partner or a ligand, we constrained these positions to stay the same (native) amino acid during optimization. Conversely, we also included the ability to avoid certain amino acids at all positions (e.g. cysteine). Both capabilities are implemented as adding or subtracting a large number ($10^8$ ) to the sequence logits at the beginning of optimization.

### 2.12.2 Loss Function

We optimize a loss function

$$\mathcal{L} = wM\mathcal{L}M + wH\mathcal{L}H + \mathcal{L}aux$$

consisting of the motif loss $\mathcal{L}M$, which scores the accuracy of the functional site in the design, and a hallucination loss $\mathcal{L}H$, which scores how strongly the sequence encodes a backbone geometry (Fig. 1B), as well as optional auxiliary losses $\mathcal{L}aux$ for specific tasks (Fig. S2; Supplementary Text). For all the designs in this paper we used $wM = wH = 1$. For a protein of length L, the motif loss is defined as a negative cross-entropy between reference (one-hot-encoded) and predicted residue-residue geometric feature distributions $p(y)$:

$$\mathcal{L}M = -\sum [(\sum\sum mij \; log \; p(yij = yij \; 0 \;) \; L \; j{\neq}i \; L \; i = 1 \;)/(\sum\sum mij \; L \; j{\neq}i \; L \; i{=}1 \;)]$$

where

$$mij = \{ \; 1, 0, \text{otherwise} \; \| C\beta i - C\beta j \| \leq 20 \text{ and } i,j \in \text{motif}\}$$

$$y \in \{d, \omega, \theta, \varphi, \theta \; T \; , \varphi \; T\}$$

represents residue-residue distances and orientation angles and $y \; 0$ is the value of the distance or angle in the reference motif. The features $d$ and $\omega$ are symmetric while the angles $\theta,\varphi$ are asymmetric, so $\theta \; T$ and $\varphi \; T$ are included to match the double-counting of $d$ and $\omega$ across the diagonal. This cross-entropy is averaged over all residue pairs in the motif, represented as a binary mask m. We restrict this loss to residue pairs within 20 Å because RosettaFold and trRosetta do not make quantitative predictions beyond this distance. In some cases we supplemented this cross-entropy motif loss with a backbone coordinate RMSD loss (Supplementary Text).

The hallucination loss is defined as the entropy of renormalized network predictions:

$$\mathcal{L}H = \sum [(\sum\sum(1 - mij)H(\hat{p}(yij)) \; L \; j{\neq}i \; L \; i{=}1 \;)/ \; (\sum\sum(1 - mij) \; L \; j{\neq}i \; L \; i{=}1 \;)] \; y \in \{d,\omega,\theta,\varphi,\theta T,\varphi T\}$$

where the entropy is defined as

$$H(p) = \sum pk \log pk\, k \text{ and } \tilde{p}(y) = \exp(\beta \log p(y)) / \sum \exp(\beta \log p(y)).$$

The last of the K distance or orientation bins (>20 Å pairwise distance or "no contact") is excluded to avoid the trivial minimum-entropy solution of an extended chain where most residues are not in contact. Empirically, we found that performing this renormalization with $\beta$ = 10, and only using bins up to 5 Å for the pairwise distance distributions $p(d)$, gave more realistic structures. In an earlier version of our method we defined the hallucination loss using a KL divergence rather than entropy, which gave similar results (Fig. S2D; Supplementary Text) (14).

### 2.12.3 Optimization Methods

In early tests, we used an MCMC method based on our previous work on unconstrained hallucination (12). Starting from a random sequence, single mutations were proposed and the loss function evaluated. The mutation was either accepted or rejected according to the standard Metropolis criterion. Acceptance temperature was 0.002 and annealed by exponential decay with a 500-step half-life; design quality was not sensitive to these parameters. For proteins around 120 residues long, we found this approach converged in about 30,000 steps and took about 90 minutes on Nvidia GeForce RTX2080 GPUs, which we used for all hallucination runs. Although slow, this approach has the advantage that mutations can include insertions and deletions, which is useful when redesigning loops.

For most design problems, we used a gradient-descent method based on our previous fixed backbone sequence design study (13). Starting with randomly initialized input logits $X \sim NN \times L \times A(0,\ 0.01)$, we apply a softmax followed by an argmax operation to obtain a one-hotencoding *Xoh*. To backpropagate the gradient of the loss $\nabla \mathscr{L}$ through the discrete one-hot sequence to the continuous logits, we employed a reparameterization trick(13, 59) where

gradients were passed through the one-hot sequence as if it had the softmax values of the logits (60, 61). For a protein of length L, on optimization step t, we update the input logits with normalized gradients and a constant learning rate $\alpha$:

$$X(t+1) \leftarrow X(t) - \alpha \sqrt{L}\, \nabla\mathscr{L} \| \nabla\mathscr{L} \|$$

Typically we used $\alpha = 0.05$, although results are reasonable for any $0.01 < \alpha < 0.2$ (Fig. S19A). We also tested decaying the learning rate over time, but this did not outperform constant learning rate, as seen previously for fixed backbone hallucination (13). With trRosetta, we found that sampling from the softmax distribution over sequence logits (59) yielded higher DAN-lDDT and lower motif RMSD than simply taking the most probable sequence (argmax), but argmax was better when using RosettaFold.

Gradient-based optimization with trRosetta converged in 200 steps for a 120-residue protein, taking approximately 5 minutes on our GPUs, while RosettaFold took 400 steps or 10 minutes per design. A hybrid procedure of gradient descent followed by MCMC yielded improved designs but required much more GPU time, while MCMC-only or MCMC followed by gradient descent yielded inferior results (Fig. S19B-C). In practice, we found that the most efficient use of GPU time was to first generate designs using gradient descent to sample a diverse structural space (and explore hyperparameters such as motif placements and sequence length), then use the best resulting designs to "seed" many short MCMC trajectories (300-1000 steps) to obtain further-refined and diversified final designs.

### 2.12.4 Motif Placement

At the beginning of optimization, each discontinuous segment of the motif is mapped to a random block of residue positions on the designed sequence. The motif loss is applied to these "constrained" regions, while the hallucination loss is applied to the remaining residue positions.

The positions corresponding to the motif stay fixed during optimization. For each new problem, we start by specifying a range for the total protein length L and generate many designs with randomly sampled L from the range and randomly placed motif segments. We then identify the values of L and inter-segment gap lengths that yielded the best designs and run followup design trajectories with these parameters in order to deeply sample productive regions of the search space. In early testing, we developed algorithms which adaptively place motifs during optimization either by minimizing motif loss over all possible placements or performing a greedy search (Supplementary Text). While potentially useful for certain problems, these were not consistently better than the simpler fixed-placement strategy (Fig. S19D-E).

### 2.12.5 Scaffolding Enzyme Active Sites Using AlphaFold

To design de novo scaffolds for the active site of Δ5 -3-ketosteroid isomerase (KSI) (36), we used AF in a two-stage method, the first stage focusing on backbone generation and the second on sidechain geometry optimization. In stage 1, we perform 200 steps of gradient descent to optimize a real-valued tensor $X \in \mathbb{R}1\times L\times A$ representing sequence logits. The argmax of the softmax of the logits is used as input to AF and trRosetta. To allow backpropagation through the argmax function, we use the gradient straight-through trick as described previously (13). Gradients are obtained from both AF and trRosetta, weighted equally, and used to update the logits X. Losses used for AF are the predicted LDDT and aligned error (for hallucination) and Cb distogram CCE (for motif recapitulation, defined similarly as the CCE used with RosettaFold above), sidechain FAPE (21) and RMSD (root-mean-squared-deviation); losses for trRosetta are KL divergence (Supplementary Text) and CCE, but excluding the theta dihedral. Stage 1 is run using the ADAM optimizer (62) with a learning rate of 5e-3. The gradients are normalized by the norm at each iteration. We found that if we do not use trRosetta as part of the loss, it is very

unstable and the motif RMSD rarely goes below 2 Å (see further discussion in Supplementary Text). In stage 2, the sequence from stage 1 is subjected to 400 steps of semi-greedy optimization using AF: at each step a random position is mutated, if the loss decreases, the mutation is accepted, if not, up to 20 independent random mutations are attempted. If none of the 20 mutations decreased loss, the mutation with best loss is accepted. For the first stage, 400 independent designs were generated. Each design had 3 random indices between 0 and 99 selected to define the positions of the active site. The top 4 designs were selected for stage 2. The loss for stage 2 is the weighted sum of predicted LDDT and aligned error, and sidechain FAPE and RMSD. The confidence loss was scaled by 0.01 and sidechain loss by 1.0. To attempt to avoid false local optima in a particular set of AF weights, during stage 2 we evaluated the loss using a randomly chosen one of 4 AF models (model_1_ptm, model_2_ptm, model_3_ptm, and model_5_ptm) (sets of weights) on each step. This is similar to averaging the 4 models (54, 55) but is more compute efficient. We withhold model_4_ptm for validation -- the designs shown in the figures come from this model.

### 2.12.6 Protein Binder "Two-Chain" Hallucination

To design Mdm2 binders, we first used standard hallucination to scaffold the p53 helix, with the repulsive loss on. These designs are roughly shape-complementary to Mdm2 but do not make biochemical interactions. We then refined a small number of high-scoring designs by 100-1000 steps of MCMC with RosettaFold predicting the entire binder/target complex but only optimizing the binder sequence. We predicted complexes by concatenating the binder and target sequences with a 200 amino-acid gap between them in the residue index input to RosettaFold (16). RosettaFold has limited accuracy predicting native protein structures and complexes from single sequences. To ensure that the target is accurately predicted (as this is a prerequisite for

accurately hallucinating interactions to it) we input the structure of the target plus the stub as homology templates to RosettaFold (Fig. S17A). As expected, this usually yielded predictions of the target (and target-stub relative position) extremely close to the crystal structure. During 2-chain refinement, we applied the motif loss to preserve the structure of the binder and its relative position to the target and the hallucination loss to the rest of the binder to encourage formation of interactions with the target; no repulsive or attractive losses were used. For this task, gradient descent did not give good results, and MCMC refinement of a previously hallucinated monomer was the most efficient and robust approach.

To generate the 12-residue stubs against various targets, as well as binder designs against TrkA and PD-L1 without using a pre-specified motif, we initialized a completely random sequence of a pre-defined length (12 AAs or 55-80 AAs) and concatenated it to the sequence of the target (Fig. S17A). On each iteration we predicted the structure of the complex using template input for the target, as described above. To promote binder-target contacts, we used an "inter-chain" entropy loss which was computed only on the inter-chain residue pairs and given a weight of 1 to 5 (Supplementary Text); the usual (intra-chain) entropy loss (with weight 1) was also used, to promote hallucination of a well-packed binder monomer. The entropy calculation was modified in some cases to improve handling of the "no-contact" bin (see "Leaky entropy" in Supplementary Text). For the stub design problem (Fig. S18), we ran 600 steps of MCMC (gradient descent was not possible for these targets due to GPU memory limitations); for TrkA and PD-L1 (Fig. S17F-G), we ran 200-400 gradient descent steps followed by 200-300 MCMC steps. Multiple rounds of filtering and design refinement/diversification were performed (Supplementary Text).

### 2.12.7 Training RosettaFold to jointly model sequence and structure (RFjoint)

Standard RosettaFold (16) (RF) has been trained on structure prediction (sequence inputs, structure outputs) using homolog templates (structure input). In the newer versions, we mask a portion of the input MSA and apply a loss to predictions of the masked amino acids (sequence output) to encourage the network to extract more meaning from the MSA (21, 63). RFjoint was fine tuned from a pre-trained RosettaFold model (RF-Nov05-2021, see Supplementary Text, "RosettaFold variants" section for details on the architectural details of this model). The training regime for this model, which was initially trained solely on structure prediction, is below: Training set: 25% of examples came from the PDB (published before February 17th, 2020), which is the same training set used in the original RosettaFold model (16). The other 75% of examples included a distillation set of AlphaFold predicted structures (64). This distillation set was clustered at 30% sequence identity cutoff, and sequences sharing greater than 30% similarity to any protein in the PDB were excluded. Only proteins greater than 200 residues in length, with mean AlphaFold pLDDT > 85 were included in training, and only residues with per residue pLDDT > 70 were included from these models. The AdamW Optimizer was used throughout training, with default pytorch parameters. The epoch size was 25600 training examples, with a batch size of 64. The learning rate for the initial round of training (200 epochs) was 0.001, with a linear warm-up for the first 1000 optimization steps. The learning rate was then decayed by a factor of 0.95 after every 10000 optimization steps. A crop size of 256 residues was used, with cropping following the same strategy as described previously (16). The number of MSA seed sequences was 128, and the number of extra MSA sequences was 1024. For the second stage of training (100 epochs), the learning rate was set of 0.0005 (no warmup), with learning rate decay

by a factor of 0.95 every 10000 optimization steps. A larger crop size (350 residues), and more MSA sequences (256 seed sequences, 2048 extra sequences) were used in this second phase of training.

Starting with this pre-trained RosettaFold, we fine-tuned this model for inpainting, for an additional 27 epochs on three tasks (Fig. S4), training only on the PDB training set. For tasks 1 and 2 (fixed backbone sequence design, and inpainting respectively, chosen 33% of the time each) were masked in essentially the same manner. Contiguous regions of 10-35 amino acids comprising at least one full secondary structure element (helix, loop or strand) were masked out (Task 1: only sequence masked; Task 2: sequence and structure masked). The sequence and structure of a further 3-6 'flanking' residues were masked out either side of this contiguous region (Fig. S4A, red). The distograms (but not angle maps or amino acid identity) were provided for the residue immediately N- and C-terminal to the central contiguous masked region (Fig. S4A, asterisks). Noise was also applied to these two positions, by randomly translating them following a normal distribution ($\mu = 0$ Å, $\sigma = 1$ Å), such that at inference time, coordinates would be provided to the network as a "guide" rather than as absolute positions. Losses were not applied to the flanking regions either side of these two coordinates. The masking of flanking sequence and structure modestly improved the performance of the network in the benchmarking test, compared to just masking a 10-35 residue window (Fig. S4D). The final task (structure prediction from MSA information) was the original task the pre-trained RosettaFold was trained on, which differs slightly from the original RosettaFold network (15). Specifically, in this task, 15% of the MSA (excluding the input sequence) was randomly masked or corrupted (following the strategy used by AlphaFold (21), of this 15% of residues, 70% of residues were replaced with a 'mask' token, 10% were mutated to a random amino acid, 10% were mutated to

another amino acid in the MSA column, and 10% were not replaced). Homologous template structural inputs were unchanged from the original network (15). The applied loss function was the same for all three tasks:

The loss function formulation for RFjoint is as follows.

$$\mathcal{L}total = 1.0\mathcal{L}dist + 3.0\mathcal{L}aa + 1.0\mathcal{L}tors + 5.0\mathcal{L}FAPE + 0.1\mathcal{L}lddt$$

Where $\mathcal{L}dist$ is a cross entropy loss over the distogram and anglegram as described in (15), predictions $\mathcal{L}aa$ is a cross entropy loss over any masked positions in the input MSA, $\mathcal{L}tors$ is a cross entropy loss on binned backbone dihedral angle predictions, $\mathcal{L}FAPE$ is a backbone level frame aligned point error, as described in (21), with a relu cutoff of 20. $\mathcal{L}lddt$ is the lDDT loss as calculated in (15). Note that structure related losses are applied over the entire predicted protein, and the sequence cross entropy loss is only applied at masked (Tasks 1 and 2) and/or corrupted (Task 3) regions. For the fixed-backbone sequence design task (Fig. S4A, Task 1) and for the inpainting task (Fig. S4A, Task 2), no loss was applied on the 'flanking' region of protein N- and C-terminal to the central masked region. The learning rate was set to 0.0003 throughout the training of these three tasks, with a batch size of 512. We refer to this fine-tuned RosettaFold inpainting model as RFjoint, and selected training curves from this model are shown in Fig. S4B,C. Details of a different training strategy used to train an earlier version of the inpainting network, which implicitly learned to inpaint, are provided in the supplementary methods.

### 2.12.8 Joint Sequence-Structure Inpainting with a Jointly Trained RosettaFold

To apply RFjoint to protein design, we input a sequence and structure, masking certain residues in the sequence by replacing them with mask tokens and masking corresponding residues in the structure by setting their template embeddings to zero (16). We then predict the

structure and sequence logits for the entire protein. The output structure, including regions that were originally both masked and unmasked, is used as the design model, and the most probable predicted amino acid at each masked position (argmax) is taken to complete the sequence. Note that in the RF-Nov05-2021 version of RosettaFold used to train RFjoint, as in AlphaFold, latent representations of the output structure are 'recycled' back through the network to refine the final structure. During inpainting, we utilize this 'recycling' to refine our inpainted sequence and structure, typically recycling information 5-15 times (similar to the number of times used for structure prediction with RosettaFold, which is typically 10). A single design of 100 amino acids in length, using 10 iterations of inpainting, takes 5.3 seconds on a GeForce RTX 2080 GPU. We refer to this prediction, with recycling, as a 'forward pass' through the network. The iterative inpainting method described above is approximately deterministic. To sample ensembles of outputs with small variations in sequence and structure using RFjoint, we either vary the exact boundaries of masked regions, the length of regions to replace a masked region or by varying specific input coordinates (for example, in Fig. S6C, the coordinates of two Cα- coordinates were randomly translated up to a specified distance from their original positions, and the network was tasked with inpainting the masked region given the unmasked positions of the two translated residues). For each of the design cases presented in the paper, the precise strategy used to generate and filter the designs is described in the supplementary methods.

### 2.12.9 Motif Selection

Because RosettaFold predicts helices and sheets more accurately than loops, we selected functional motifs composed of as much secondary structure as possible. In initial exploratory design runs, our methods performed poorly if the motif to scaffold contained too many loops or depended on networks of tertiary polar contacts (e.g. antibody H3 CDR regions). For antigenic

epitopes, viral receptor traps, and enzyme active sites, we chose the functional motifs based on previous structural literature. For binding interfaces, we identified interface residues as those with any atom within 5 Å of the binding partner and scaffolded motifs consisting of 2-4 contiguous blocks manually chosen to contain as many of the interface residues as possible. Table S1 lists the design targets, their PDB accessions, the residue numbers of constrained regions, and references.

### 2.12.10 Design Filtering and Selection

For each experimentally tested design case shown in this paper, we generated between 4000 and 30,000 designs, and filtered these based on the AF pLDDT, motif RMSD of AF predictions to native, (see supplementary text for exact cutoffs). Broadly, these included 'confident/accurate' AF pLDDT (> 80), sub-angstrom (< 1 Å) AF-RMSD. Orthogonal filters were determined on a per-problem basis (fully outlined in the supplementary text), but broadly comprised features such as radius of gyration, Rosetta per-residue spatial aggregation propensity (SAP) score (65), net charge (# Arg + # Lys - # Asp - # Glu) and structural diversity. The cutoffs were typically chosen to give an experimentally tractable final number of designs. In some cases, in preparation of the final set of proteins to be ordered, and after design filtering, we performed a final visual inspection to look qualitatively at aspects such as poor core packing, presence of cavities, buried polar groups, or surface hydrophobics, which typically reduced the set of proteins by around 0-50%. For designs that were only validated in silico, that are represented in the figures, we filtered designs predominantly on AlphaFold pLDDT and AF-RMSD, as well as radius of gyration. The AlphaFold metrics are presented in Table S2. The "model 4" weights were used for all AF predictions for filtering. The pLDDT was taken as the average of the residue-wise confidence values output by the network. Using AF to filter our designs has the risk

of designing “adversarial examples”, or sequence-structure pairs that score well by AF that do not fold or function in reality, due to the presence of artifactual minima in the loss landscape of the structure-prediction model (66, 67). However, because we design using RosettaFold, which is trained independently of AF (although both use the PDB as training data), any final designs must be well-predicted by two partially orthogonal networks, which is expected to provide some (although not total (68)) robustness to adversarial examples. This is supported by our finding that a high fraction of our designs are solubly expressed. Additionally, if we redesign the sequence of our highest-pLDDT designs by Rosetta, pLDDT continues to be high, indicating that the original hallucination had a designable backbone (and isn’t purely an artifact of RF or AF’s loss landscape) (Fig. S7C). Finally, we find that AF pLDDT of our RF-generated designs correlate well with physics-based metrics such as Rosetta energy and ab initio folding (Fig. S7D, F; Supplementary Text). To score protein binder designs, we used a modified AlphaFold prediction script that took as input the design model of the target-binder complex (from RF hallucination or inpainting) and the concatenated binder-target sequence (with a residue number gap to denote different chains). AF was asked to predict the complex structure from single-sequence, given the target protein structure as template information and its structural representation (atom coordinates) of the binder-target complex initialized to the target-binder complex design model. The confidence in AF2’s prediction of the interface was assessed by the inter-chain predicted aligned error (interPAE), or the average value of interchain positions in the predicted aligned error matrix. We found that inter-PAE < 10 Å corresponded to predicted complexes that were docked roughly correctly, while predictions with inter-PAE above this threshold usually had binder and target far apart in space. In addition to inter-PAE, we also filtered on: binder pLDDT (average residuewise confidence over the binder from complex prediction); AF-Rosetta ddG

(Rosetta ddG calculated on the AF model after minimizing interface side chains); target-aligned binder RMSD (RMSD of the binder, after aligning AF and RF models on the target).

### 2.12.11 Protein Purification

All designs tested in E. Coli were cloned, expressed and purified using standard methods. Briefly, Golden Gate assembly with BsaI-HF (New England Biolabs) was used to insert designs into a modified pET29b+ vector containing C-terminal SNAC (69) and 6xHis tags (or, in the case of EFhand_inp_1, into a modified pET29b+ vector with a C-terminal TEV cleavage site and a 6xHis tag). Plasmids were transformed into BL21 bacteria. For small-scale expression tests, bacteria were cultured overnight at 37oC in 2 ml cultures of lysogeny broth (LB) supplemented with 50 μg/mL of kanamycin. Cells were then grown in 2 ml cultures of Terrific Broth (TB) for one hour, before induction with 1 mM of IPTG for 4 hours. Cells were then lysed with B-PER supplemented with 1 mM PMSF, 0.1 mg/mL Lysozyme, 25 U/ml Benzonase, before lysate clarification by centrifugation. Lysate was incubated with 75 μl Ni-NTA resin, before washing thrice with wash buffer (25 mM Tris, 300 mM NaCl, 20 mM Imidazole, pH 7.8) and elution in 25 mM Tris, 300 mM NaCl, 250 mM Imidazole. Expression was assessed by SDS-PAGE. For larger scale cultures, cultures were grown overnight at 37oC in autoinduction medium (70), before sonication-based lysis in wash buffer supplemented with 1mM PMSF, 0.1 mg/mL Lysozyme, 0.01 mg/ml DNase I. After centrifugal lysate clarification, lysates were incubated with an appropriate volume of Ni-NTA resin and subsequently washed thrice with wash buffer. For purification of di-iron binding proteins, the His-tag was cleaved off by cleavage of the SNAC-tag. Briefly, after binding to the Ni-NTA resin, the protein was washed in SNAC cleavage buffer (100 mM CHES, 100 mM Acetone oxime, 100 mM NaCl, 500mM GuHCl, pH 8.6) before addition of 2 mM NiCl2. After overnight cleavage, proteins were further purified by size

exclusion chromatography on a Superose 75 column in 20 mM Hepes, 100 mM KCl, pH 7.8, and monomeric fractions pooled.

### 2.12.12 Spectroscopic Analysis of Cobalt Binding to Di-Iron Binding Proteins

Analysis of cobalt binding to inpainted di-iron binders was performed essentially as described previously (32). Proteins (200 μM in 20 mM Hepes, 100 mM KCl, pH 7.8) were incubated overnight with (or not) an 8x molar excess (1600 μM) CoCl2. Absorbance spectra were collected in a Jason V-750 spectrophotometer. Mean background absorbance (measured between 700 and 800 nm) were subtracted from all spectra. Successful designs showed absorbance peaks characteristic of cobalt coordinated in a tetra/penta-coordinate state.

### 2.12.13 Fluorescence Analysis of Terbium Binding to EF-Hand Designs

Yeast-displayed designs: Transformed yeast were cultured in TRP(-), URA(-) media for two days followed by expression culture. Samples containing ~8.5e7 cells were incubated in TBS (pH 8.0) containing 1mM Ca2+ and washed twice with TBS only. Yeast cells were resuspended in TBS containing 50 μM Tb3+ For 3 hours and then washed twice in TBS + 1mM Ca2+. Washed samples were moved to a black bottom, plate-reader 96 plates for fluorescence spectra measurement. Fluorescence signals were collected using a flash plate reader in time-resolved fluorescence mode (TRF, delay time: 100us , integration time: 1000us, gain: 130).

Purified designs: Designs harboring the EF-hand motif , were purified by His-purification as described above. After size exclusion chromatography in 20 mM Hepes, 150 mM KCl, pH 7.8, the His tag was cleaved by TEV-cleavage, with the addition of 40 μM Super-TEV protease,

1 mM DTT and 0.5 mM EDTA (overnight at room temperature). To ensure the EF-hands were not bound to any residual calcium in buffers, after passing through a NiNTA-column after TEV-cleavage, protein were run on a size exclusion column equilibrated in 20 mM Hepes, 150 mM KCl, pH 7.8 buffer, which had been Chelex treated overnight to remove any residual calcium. Proteins were incubated (or not) with terbium (40 μM terbium in 5 μM protein) for 3 hours, before analysis of terbium fluorescence on a NEO2 plate reader. Samples were excited at 250 nm (to excite the tryptophan residue near the EF-hand motif), and fluorescence was measured between 450 and 650 nm, 100-1000 μs after excitation.

### 2.12.14 Circular Dichroism Spectroscopy

All circular dichroism (CD) analyses except those for RSV-F site V immunogens were performed on a JASCO J-1500 CD Spectrophotometer. Di-iron binding proteins were analyzed at 6.7 μM in 20 mM Hepes, 10 mM KCl, pH 7.8, with or without an 8x molar excess of CoCl2. Analysis of the EF-hand inpaint was performed at 20 μM in chelex100-treated 20 mM Hepes, 150 mM KF, pH 7.6, in the presence or absence of 200 μM CaCl2. Analysis of the PDL-1 binder was performed at 5 μM in 20 mM Hepes, 10 mM KCl, pH 7.8. Thermal melt analyses were performed between 25 oC and 95 oC, measuring CD at 222 nm. All reported measurements were measured within the linear range of the instrument.

For RSV-F designs, CD spectra were measured using a ChirascanTm V100 spectrometer in a 1-mm path-length cuvette. The protein samples were diluted to 30 μM in PBS. Wavelengths between 195 nm and 250 nm were recorded. Thermal melt analyses were performed between 20 °C and 95 °C with an increment of 2 °C/min, measuring CD at 222 nm. All spectra were corrected for buffer absorption.

### 2.12.15 Measuring Protein Binding

Yeast surface display: As an initial screen for protein binding, linear DNA were synthesized as “e-blocks” (Integrated DNA Technologies), pooled, and transformed into the yeast strain EBY100 (by electroporation if >100 designs, by the lithium acetate method otherwise) along with a pETCON3 backbone linearized at NdeI and XhoI (for Aga2p and c-Myc fusion) (4, 5). The transformed pool was inoculated into CTUG medium (yeast nitrogen base 6.7g/L (difco) + complete amino acids -trp - ura + 2% glucose) and incubated 12-16 hours at 30°C with shaking, then diluted 200uL + 2mL into SGCAA (yeast nitrogen base 6.7g/L + complete amino acids 5g/L (Bacto) + 90mM Na2HPO4 + 2% galactose + 0.1% glucose) and incubated 12-16 hours to induce binder expression and display. For flow sorting, around 107 cells were harvested, washed 3x in TBSF (50mM Tris-HCl pH8.0, 150mM NaCl, 1% bovine serum albumin), incubated in TBSF with biotinylated binding target for 30 minutes at room temperature, washed 1x in TBSF, incubated for 30 minutes at room temperature in 0.1mg/mL FITC anti-c-Myc (ICL Lab) and 70mg/mL streptavidin R-phycoerythrin (PE) conjugate (Invitrogen), and washed 3x in TBSF. The binding target and FITC/PE were added in the same incubation when labeling with avidity. Cells were sorted on a Sony SH800 flow sorter and 103 - 106 FITC+/PE+ cells were collected. The cells were either cultured in liquid CTUG for another round of sorting, or plated onto CTUG agar and individual colonies Sanger-sequenced to identify the designs. For trRosetta-hallucinated PD-L1 binders and Mdm2 binders, clonal yeast cultures expressing a single design were analyzed in binding assays to confirm the results of sorting as well as to assess the binding affinity of designs. In this case, yeast culture and binding were performed identically as above except that an Attune NxT (Invitrogen) flow cytometer was used

to analyze the cells. For all other problems, hits identified by yeast display were followed up by E. coli expression and purification.

Surface plasmon resonance (SPR) to assess RSV-F site V binding: SPR measurements were performed on a Biacore 8K (GE Healthcare) in 10 mM HEPES pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.005% v/v Surfactant P20 (GE Healthcare). Ligands were immobilized on a CM5 chip (GE Healthcare) via amine coupling. The preRSVF and RSVF-site V immunogens were immobilized at approximately 300-500 response units (RU). The site V specific RSV90 Fab was injected as analyte in two-fold serial dilutions. The flow rate was 30 µl/min for a contact time of 120 s followed by 400 s dissociation time. After each injection, the surface was regenerated using 0.1 M glycine at pH 3.0. KD values were obtained by fitting the maximum response versus log10 Fab concentration to a sigmoid function using GraphPad PRISM.

Biolayer interferometry (BLI) to assess bivalent TrkA binding: BLI binding experiments were performed on an Octet Red96 (ForteBio), with streptavidin coated tips (Sartorius Item no. 18-5019) and BLI buffer (10 fold dilution of 10x HBS-EP+ buffer [Cytiva Item no. BR100669] supplemented with 0.1% w/v bovine serum albumin). Tips were preincubated in BLI buffer for at least 30 minutes before use. To collect binding data, the tips were incubated in BLI buffer for 100 s, loaded with biotinylated TrkA (30 nM in BLI buffer; a kind gift from Chris Garcia's lab) for 300 s, equilibrated in BLI buffer to obtain a baseline for 150 s , dipped into BLI buffer with the designed proteins for 900 s (association phase) and finally returned to BLI buffer for 900 s (dissociation phase). Reported responses are the change in wavelength between the beginning and end of the association phase.

## 2.13 Supplementary Figures

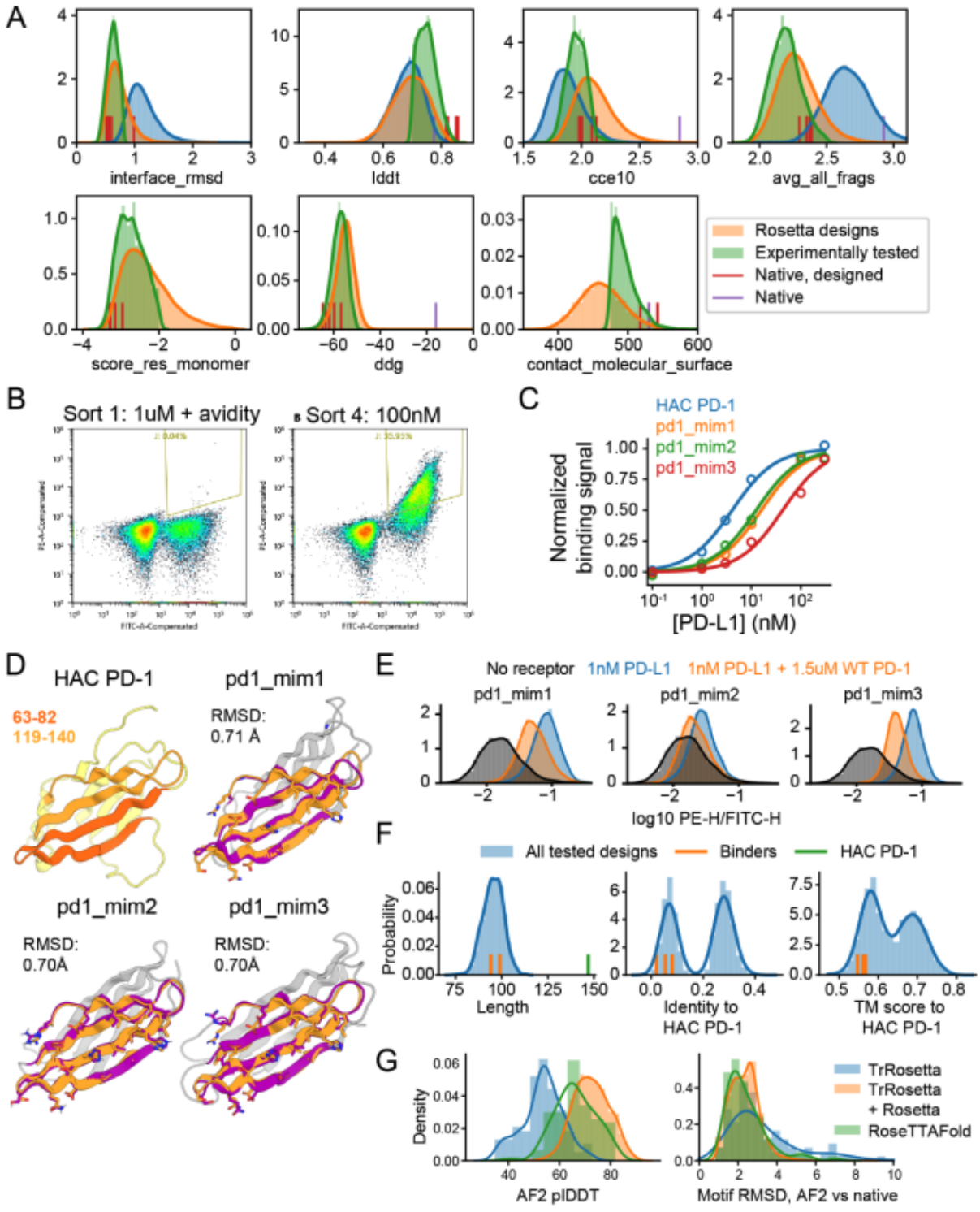


**Figure 2.S1. trRosetta-based hallucination and testing of PD-1 mimetics** (A) Distributions of metrics for PD-1 hallucinations, Rosetta designs, and experimental library (Supplementary Text). (B) PE (binding) vs FITC (surface displayed protein) signal during FACS sorting of PD-1 mimetics. Sort 2 (1 μM PD-L1 with avidity) and 3 (1 μM PD-L1, no avidity) are not shown. (C) Binding signal (Methods) from clonal yeast cultures versus receptor concentration for HAC PD-1 and designs isolated from pooled sorting. Apparent Kd values in nM are: HAC PD-1: 4.10; pd1_mim1: 15.9; pd1_mim2: 12.5; pd1_mim3: 42.9. (D) Crystal structure of HAC PD-1 (discontinuous interface motif in 2 shades of orange) and design models of 3 experimentally isolated binders. "RMSD" denotes the backbone RMSD between design model and template motif at 22 interface residues (Methods). (E) Normalized PE (binding) signal for clonal yeast cultures expressing the 3 binders in the presence of receptor and receptor + unlabeled purified wildtype PD-1. (F) Distribution of sequence length, amino-acid identity to HAC PD-1, and TM-score to HAC PD-1 for the 3,038 experimentally tested designs. The values for HAC PD-1 and the 3 binders shown in (D) are plotted as vertical bars. (G) Comparison of trRosetta and RosettaFold for hallucinating PD-1 mimetics. AlphaFold predicted lDDT and motif backbone RMSD (AF model versus native motif) for hallucinations generated using trRosetta, RosettaFold, or trRosetta followed by Rosetta-based sequence design.

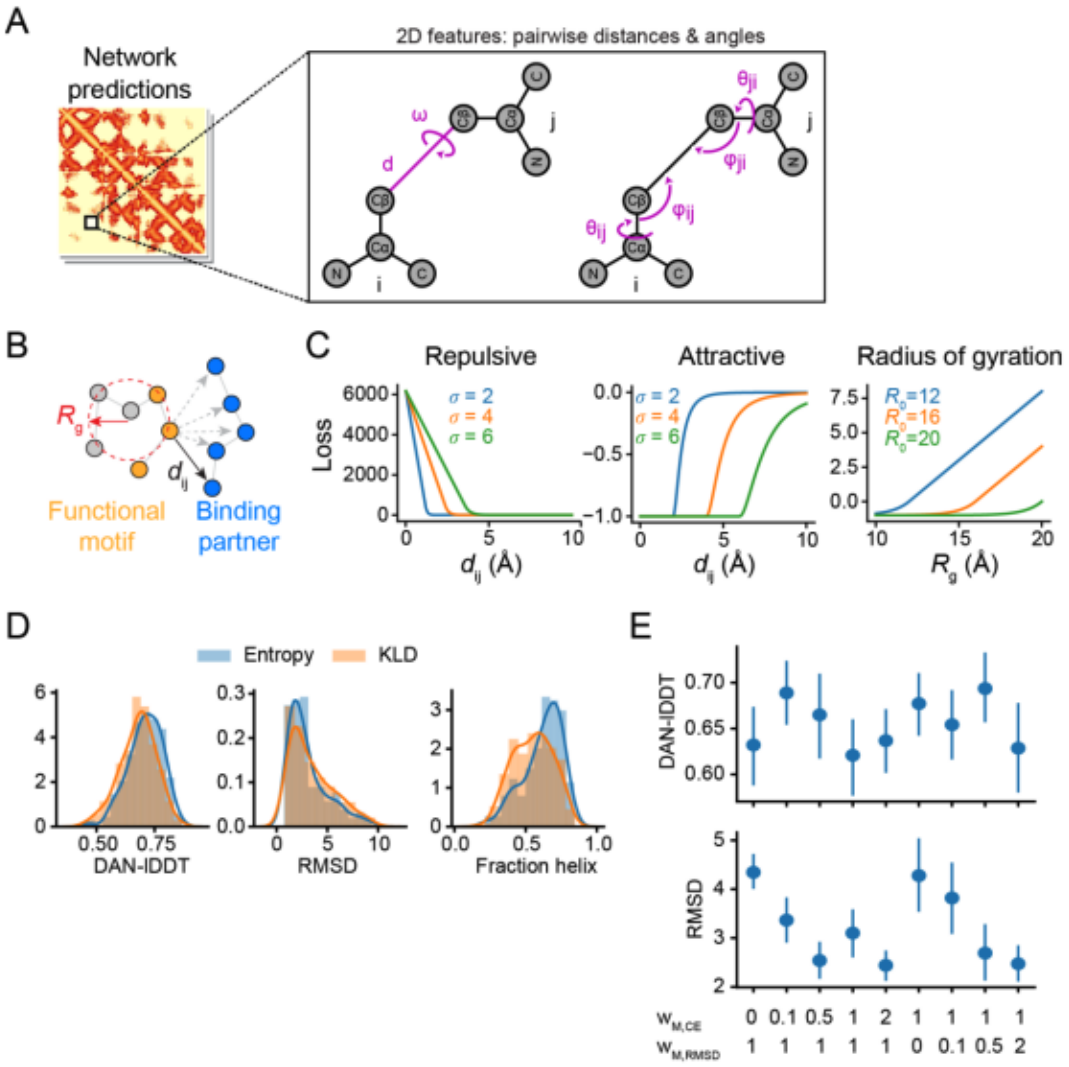



**Figure 2.S2. Auxiliary and alternative loss terms** (A) Schematic of the pairwise distances and orientation angles whose distributions are predicted by trRosetta and RosettaFold and which are used to define the motif and hallucination losses. (B) Schematic of radius of gyration and distances used to calculate repulsive and attractive losses (Supplementary Text). (C) Functional forms of the losses. (D) Distributions of DAN-lDDT, motif RMSD, and fraction of residues that are helix for designs generated using entropy or KL divergence hallucination losses (Supplementary Text), for scaffolding a 2-segment motif from C3d (1GHQ chain A residues 104-126, 170-185). (E) DAN-lDDT and motif RMSD for the same C3d scaffolding problem as in (D), but with varying the loss term weights for the cross-entropy based motif loss (wM,CE) or RMSD-based motif loss (wM,RMSD).

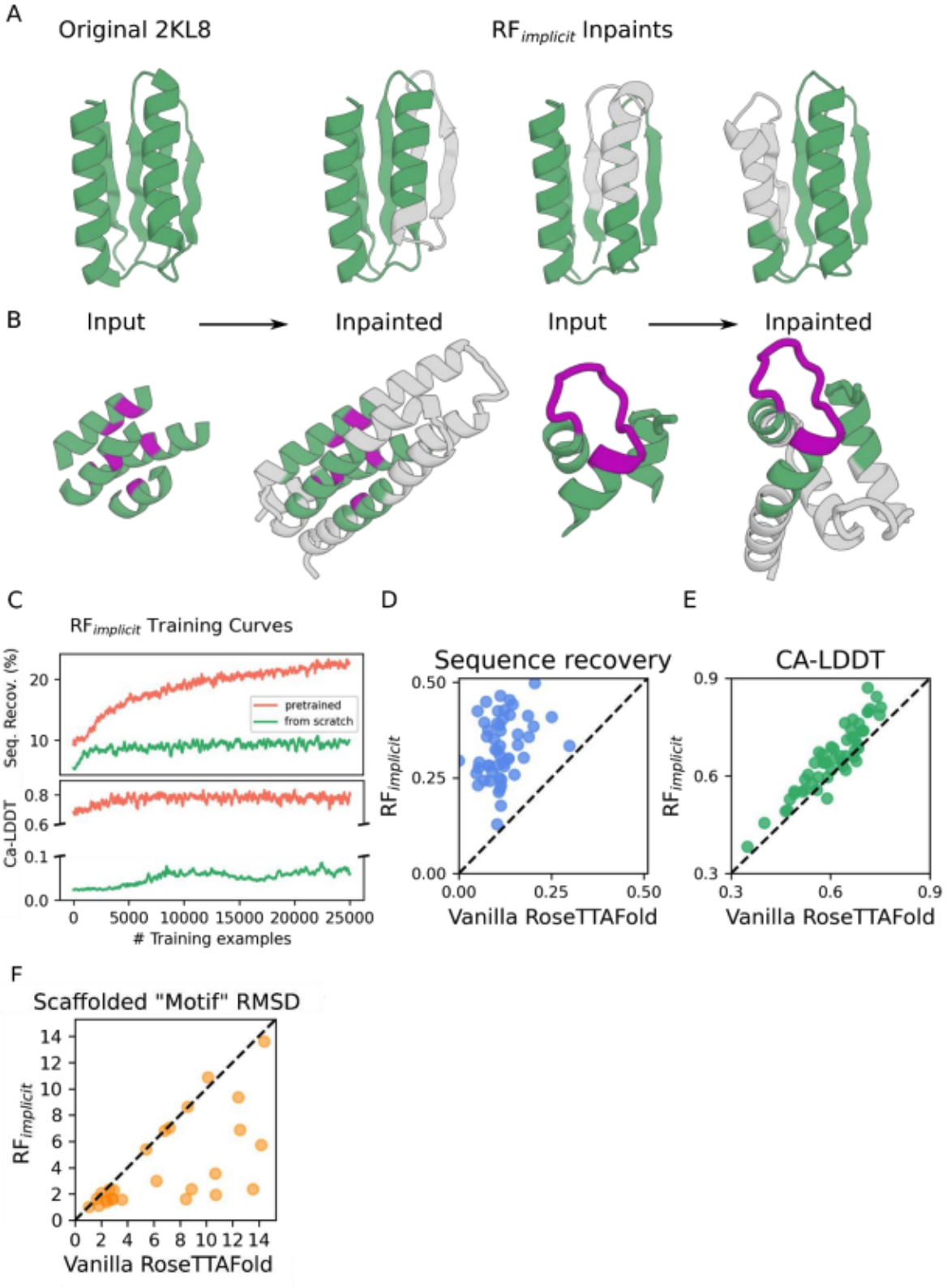



**Figure 2.S3. Training and inpainting with RFimplicit** (A) NMR structure of 2KL8 (left) alongside models of selected inpainting examples of the protein with a masked window size of

20 residues. Green denotes areas of sequence and structure that the network was allowed to see, gray denotes areas that the network inpainted. (B) Functional site scaffolding examples designed with RFimplicit. (Left) AF2 prediction of design EFhand_inp_2 from Fig. S16 scaffolding the EFhand calcium binding site, with RMSD on the motif of 0.7Å between the prediction and the native 1PRW. (Right) AF2 prediction of design dife_impl_1 scaffolding the di-iron binding site from bacterioferritin protein 1BCF, with an RMSD on the motif of 0.5Å between the prediction and the native, and an AF2 pLDDT of 91. (C) Training curves of RFimplicit show that starting the training procedure from a pretrained RosettaFold model (red) results in better sequence design accuracy and structure prediction accuracy than starting from a completely untrained RosettaFold (green). (D) Sequence recovery of RFimplicit vs Vanilla RosettaFold on a set of 52 de novo proteins (Supplementary Text, “Single sequence predictions using AlphaFold”) shows RFimplicit outperforms the baseline model at protein sequence design. (E) CA-LDDT of RFimplicit vs Vanilla RosettaFold shows the model is able to retain its structure prediction capabilities on the same set of 52 de novo proteins even after learning protein sequence design. (F) AF-RMSD of the “motif” (unmasked) region when performing the inpainting benchmark seen in Fig. 1F-G (main text) using RFimplicit vs Vanilla RosettaFold, and a masked window size of 20 residues.

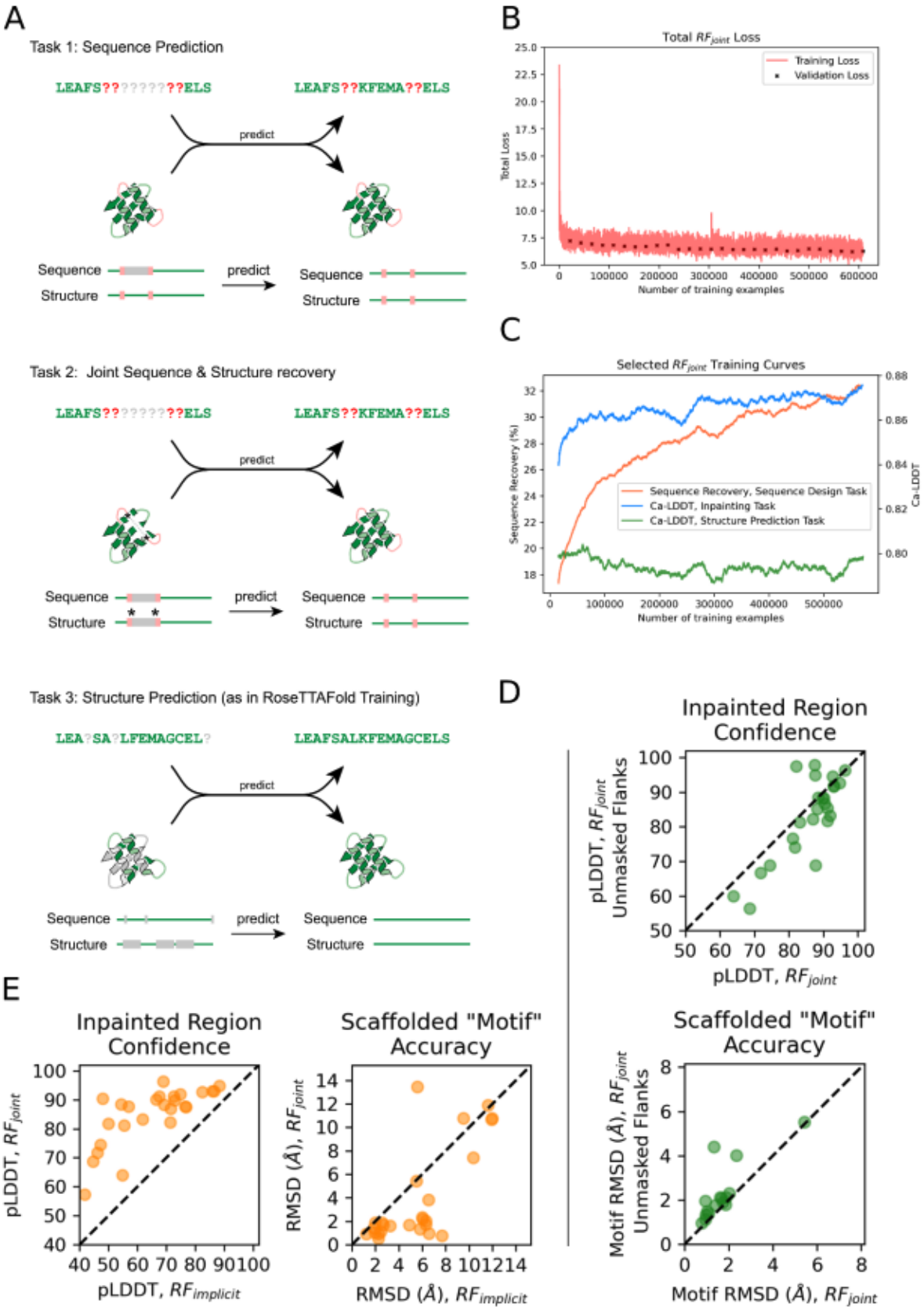

A
Task 1: Sequence Prediction
LEAFS??????ELS
LEAFS??KFEMA??ELS
predict
Sequence
Structure
predict
Sequence
Structure
Task 2: Joint Sequence & Structure recovery
LEAFS??????ELS
LEAFS??KFEMA??ELS
predict
Sequence
Structure
predict
Sequence
Structure
Task 3: Structure Prediction (as in RoseTTAFold Training)
LEA?SA?LFEMAGCEL?
LEAFSALKFEMAGCELS
predict
Sequence
Structure
predict
Sequence
Structure
B
Total RFjoint Loss
Training Loss
Validation Loss
Total Loss
Number of training examples
C
Selected RFjoint Training Curves
Sequence Recovery (%)
Ca-LDDT
Sequence Recovery, Sequence Design Task
Ca-LDDT, Inpainting Task
Ca-LDDT, Structure Prediction Task
Number of training examples
D
Inpainted Region Confidence
pLDDT, RFjoint Unmasked Flanks
pLDDT, RFjoint
Scaffolded "Motif" Accuracy
Motif RMSD (Å), RFjoint Unmasked Flanks
Motif RMSD (Å), RFjoint
E
Inpainted Region Confidence
pLDDT, RFjoint
pLDDT, RFimplicit
Scaffolded "Motif" Accuracy
RMSD (Å), RFjoint
RMSD (Å), RFimplicit

**Figure 2.S4. Training of joint sequence-structure recovery RosettaFold.** (A) Depiction of the three tasks used to train RFjoint, which were trained with equal likelihood (see Algorithm 1). Task 1 comprised a fixed-backbone sequence design task of a continuous segment of a given protein, without the immediate up- and downstream protein visible (see Methods). Task 2 comprised an inpainting task, where the model was tasked with predicting the sequence and structure of a continuous section of protein, also without up- and downstream protein visible. Asterisks indicate "guiding points" provided as inputs during inpainting to Task 3 is the structure prediction task originally used to train RosettaFold. (B) Training curve for RFjoint, showing total training (red) and validation (black crosses) losses decreasing. (C) A selection of different losses associated with each of the three tasks. RFjoint does not severely deteriorate in its ability to predict protein structures (task 3, green line), but its ability to inpaint structure (task 2) improves dramatically (blue line). The model also learns to predict the sequence of a fixed backbone (task 1, orange line). (D) Masking out the structure and sequence of the flanking regions (depicted in (A), Tasks 1 and 2) improves inpainting performance. RFjoint was compared to an identically-trained model, except that flanking regions were not masked during training, on the benchmarking task described in Fig. S5. Both AlphaFold pLDDT in the inpainted region (top), and the "Motif" RMSD of the AlphaFold predictions (bottom) were marginally better for RFjoint. (E) RFjoint outperforms RFimplicit, both in terms of the AlphaFold pLDDT in the inpainted region (left), and in the "Motif" RMSD of the AlphaFold prediction (right). Graphs in D and E correspond to a masked window of 30 residues.

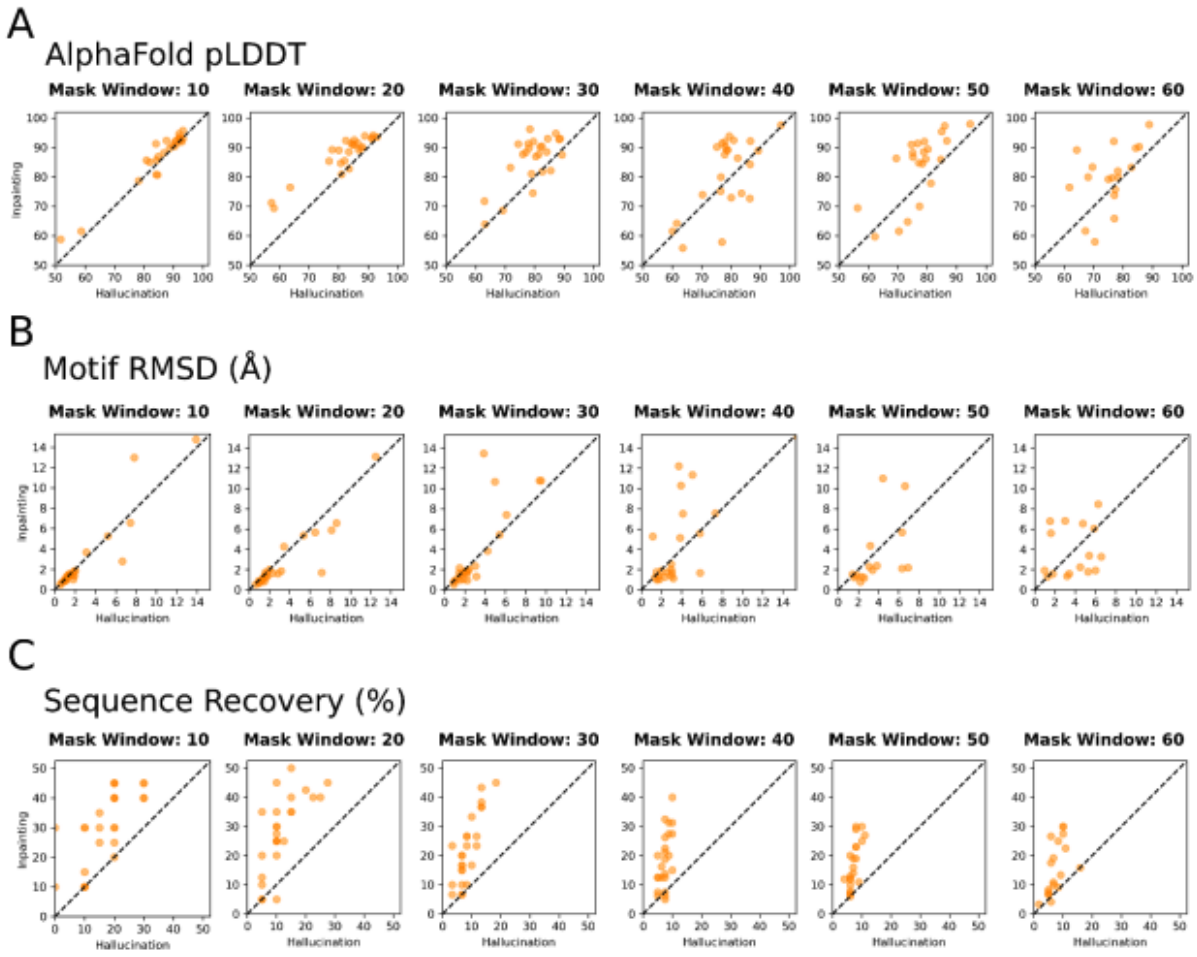



**Figure 2.S5. Comparison of hallucination and inpainting design quality**
(A) Inpainting versus hallucination AlphaFold pLDDT, as a measure of overall design quality, for various window sizes over which sequence and structure were rebuilt by both methods. Each point corresponds to a crystal structure from a benchmarking set of de novo proteins. (B) Inpainting versus hallucination motif AF-RMSD for the same benchmarking set. The "motif" is defined as the region of the protein that was not masked for rebuilding. (C) Percentage sequence recovery in the rebuilt region of protein, in the same benchmarking set.

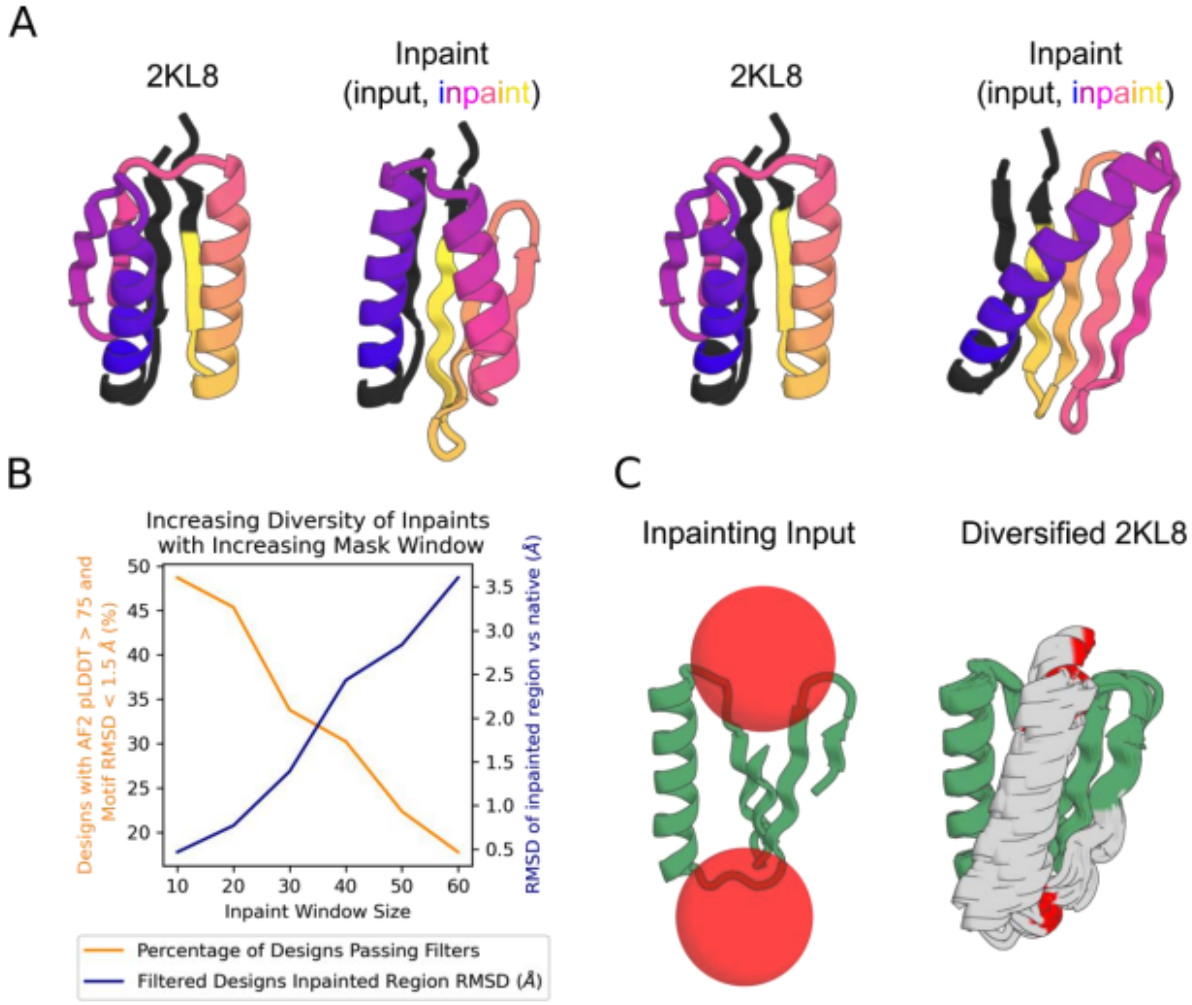



**Figure 2.S6. Generating diversity with inpainting**
(A) With a large region of structure masked, inpainting can sometimes produce confidentlypredicted designs that scaffold the input motif. Two designs are shown, with the dramatically different looping order (left) or topology (right) highlighted with spectrum colors. Both designs scaffold the input "motif" (dark gray). (B) Analysis performed on the inpainting benchmarking data shown in Fig. S5. While the proportion of inpainted designs passing AlphaFold filters (> 75 pLDDT, < 1.5 Å, orange line) decreases with increasing size of the masked window, those designs that do pass filters, and thus successfully scaffold the motif, show more scaffold diversity (as assessed by AF-RMSD to the native masked region) than those designs with a smaller inpainted region (blue line). (C) Further diversity can be explicitly generated by perturbing the input coordinates. During training, RFjoint was trained to Cα-coordinates as approximate positional information (see Methods). Therefore at inference, input Cα-coordinates can be randomly translated (uniformly sampled from within depicted spheres, left), and the model thus outputs diverse inpainted structure (right, gray) capable of supporting the unmasked "Motif" (right, green). All designs shown in (C) have pLDDT (both

total pLDDT and just in the inpainted region) > 80 and “Motif” AF-RMSD < 1.2 Å, and represent examples from each of 30 clusters (clustered at total TM score cutoff of 0.95).

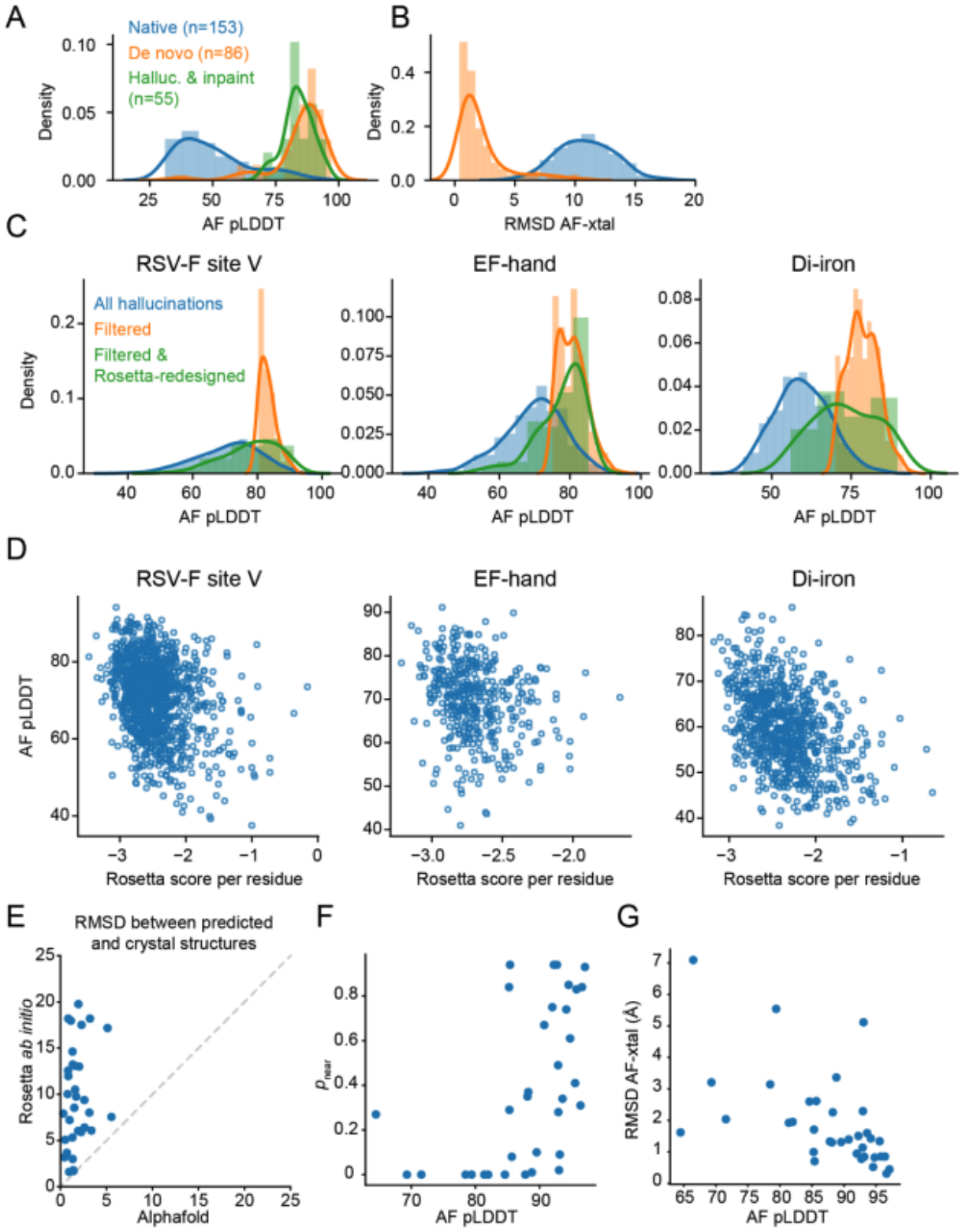



**Figure 2.S7. Using AlphaFold for in silico evaluation of de-novo-designed proteins**

(A) AF pLDDT (mean across all 5 models) using single-sequence input (no multiple-sequence alignment) for a benchmark set of 153 native proteins (79), 86 structurally validated de novo proteins, and experimentally tested or visually displayed designs in this study. (B) RMSDs between AF predictions and crystal structures (averaged across all 5 models) for the same proteins as in (A). (C) AlphaFold pLDDT distributions of hallucinations for 3 representative design problems (blue). The designs are filtered to those with high plDDT and low motif RF-AF RMSD (orange), and then the sequence is redesigned using Rosetta Fastdesign and scored again by AlphaFold (green). (D) Scatterplots of AF pLDDT versus Rosetta score (energy) per residue, showing that AF quality estimates correlate with energy-function-based quality estimates. (E) RMSD between predicted and crystal structure via Rosetta ab initio (“forward folding”) versus AlphaFold for 34 de novo designs not in the AF training set (Supplementary Text). All predictions used only single-sequence input. (F) pnear, a measure of the confidence of an ab initio prediction, versus AF pLDDT, for the de novo designs in (E). (G) RMSD between AF predictions and crystal structures versus AF pLDDT for the de novo designs in (E).

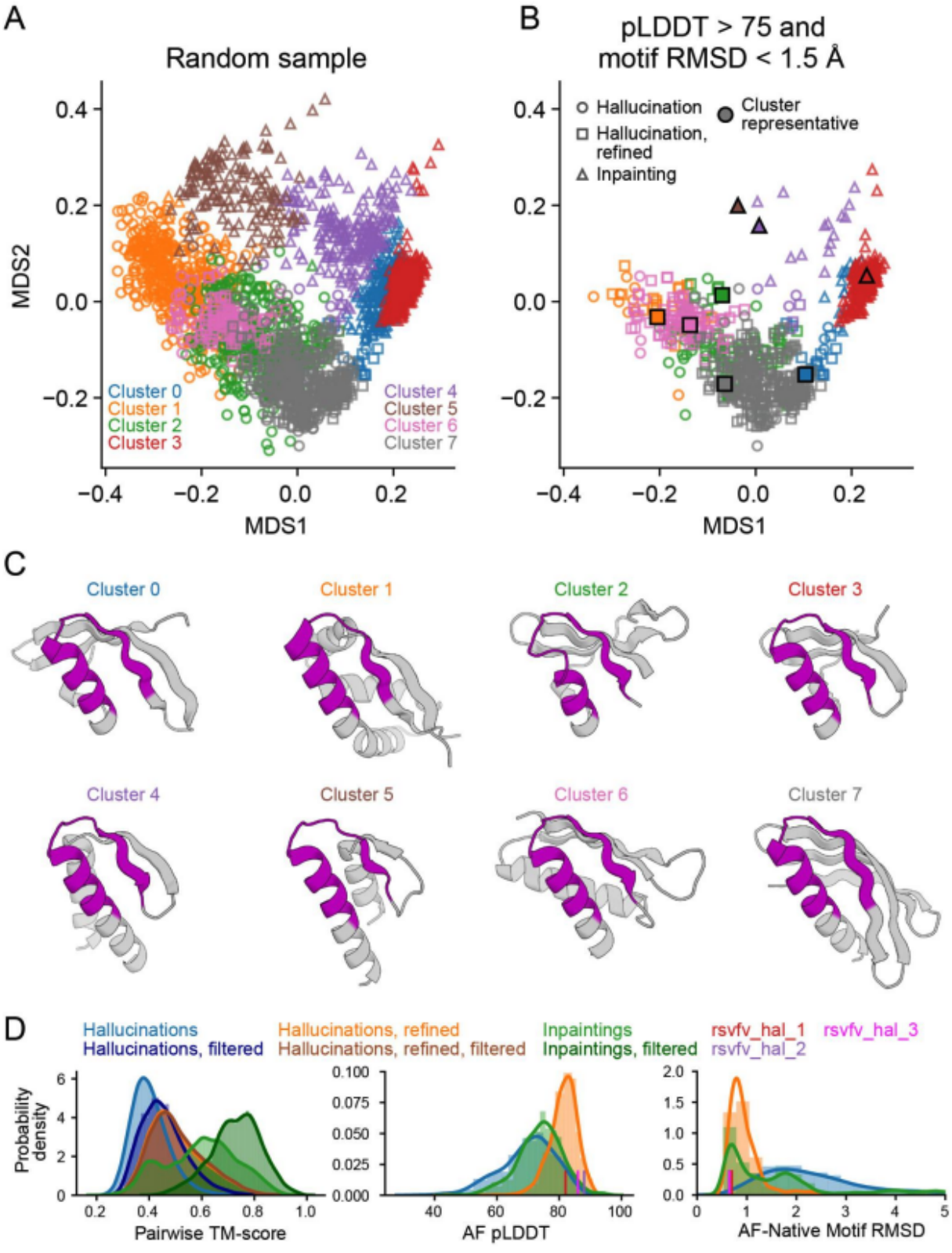

A
Random sample
MDS2
MDS1
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
B
pLDDT > 75 and
motif RMSD < 1.5 Å
Hallucination
Hallucination, refined
Inpainting
Cluster representative
C
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
D
Hallucinations
Hallucinations, filtered
Hallucinations, refined
Hallucinations, refined, filtered
Inpaintings
Inpaintings, filtered
rsvfv_hal_1
rsvfv_hal_2
rsvfv_hal_3
Probability density
Pairwise TM-score
AF pLDDT
AF-Native Motif RMSD

**Figure 2.S8. Structural diversity of hallucinated and inpainted RSV-F site V scaffolds**
(A) Random subsample of 1000 hallucinations, 500 refined hallucinations (Materials and Methods), and 1000 inpaintings for the RSV-F site V epitope scaffolding problem, and (B) subset of designs with AF pLDDT > 75 and motif AF-RMSD < 1.5 Å. All pairwise structural distances (1 - TM-score) were projected into 2 dimensions using classic multidimensional scaling. 8 clusters were identified using k-means, and design models of cluster representatives (black-outlined markers) with highest pLDDT are shown in (C) with motif region in purple. The number of k-means clusters was chosen arbitrarily. Inpaintings (triangles) and hallucinations (circles, squares) occupy different regions of structure space. (D) Distributions of AlphaFold pLDDT, motif AF-RMSD, and pairwise TM-scores within hallucinations, refined hallucinations, and inpaintings, either in full set or only designs with pLDDT > 75 and motif AF-RMSD < 1.5 Å ("filtered").

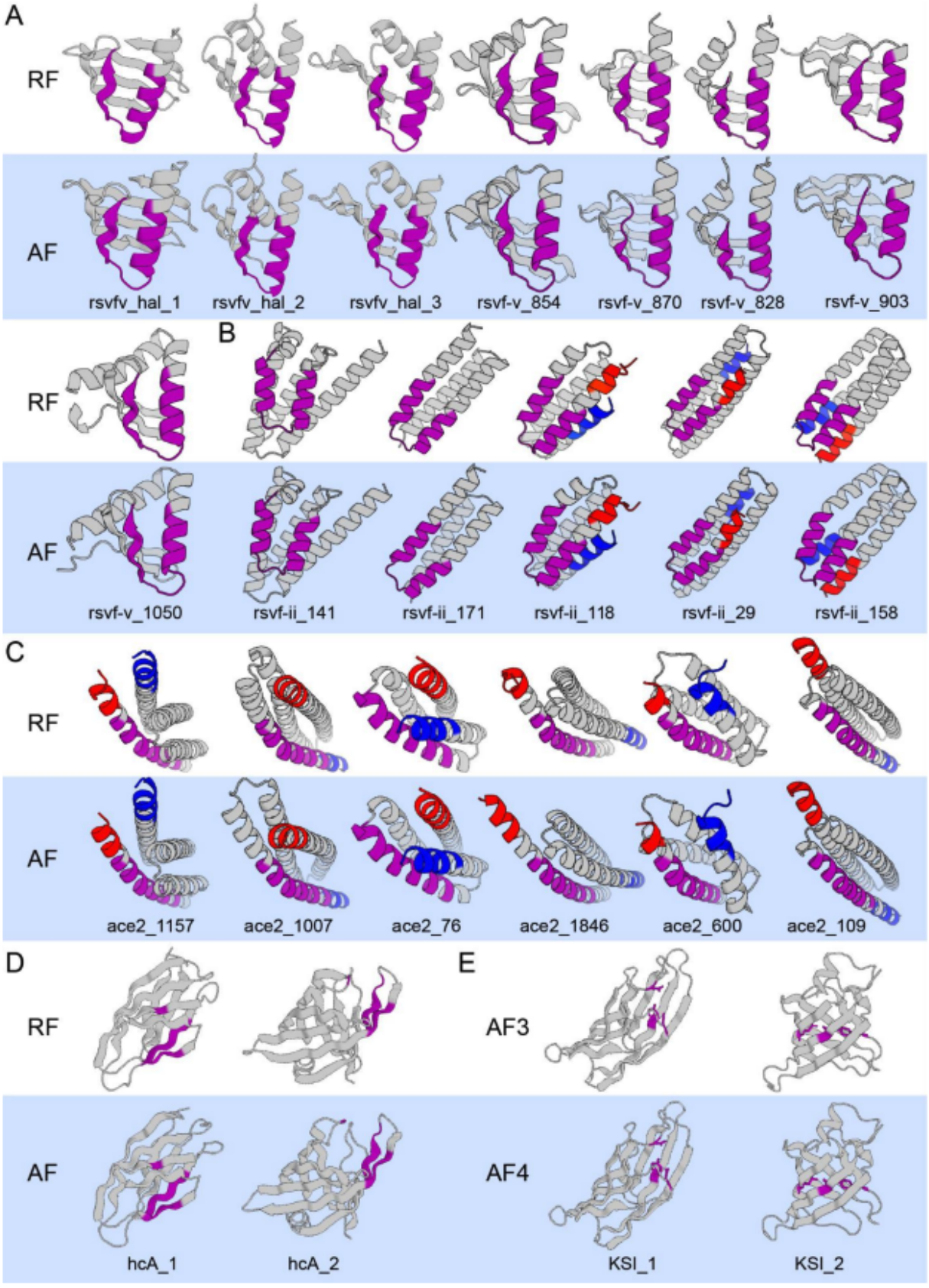

A
RF
AF
rsvfv_hal_1
rsvfv_hal_2
rsvfv_hal_3
rsvf-v_854
rsvf-v_870
rsvf-v_828
rsvf-v_903
B
RF
AF
rsvf-v_1050
rsvf-ii_141
rsvf-ii_171
rsvf-ii_118
rsvf-ii_29
rsvf-ii_158
C
RF
AF
ace2_1157
ace2_1007
ace2_76
ace2_1846
ace2_600
ace2_109
D
RF
AF
hcA_1
hcA_2
E
AF3
AF4
KSI_1
KSI_2

**Figure 2.S9. RosettaFold and AlphaFold models of hallucinations RosettaFold (RF) and AlphaFold (AF) models of hallucinations for**
(A) RSV-F site V and (B) site II epitope scaffolds, (C) ACE2 receptor traps, and (D) carbonic anhydrase and (E) ketosteroid isomerase (KSI) active-site scaffolds. These include the designs shown in the main figures, as well as additional designs. Functional motifs are highlighted in purple. The N- and C-termini in some designs have been colored blue and red (respectively) to highlight that hallucination can find diverse topological solutions, despite having similar overall folds. Because the KSI designs were hallucinated using AlphaFold model 3 (AF3), validation models were predicted with AF model 4 (AF4). Detailed metrics for these designs can be found in Table S2.

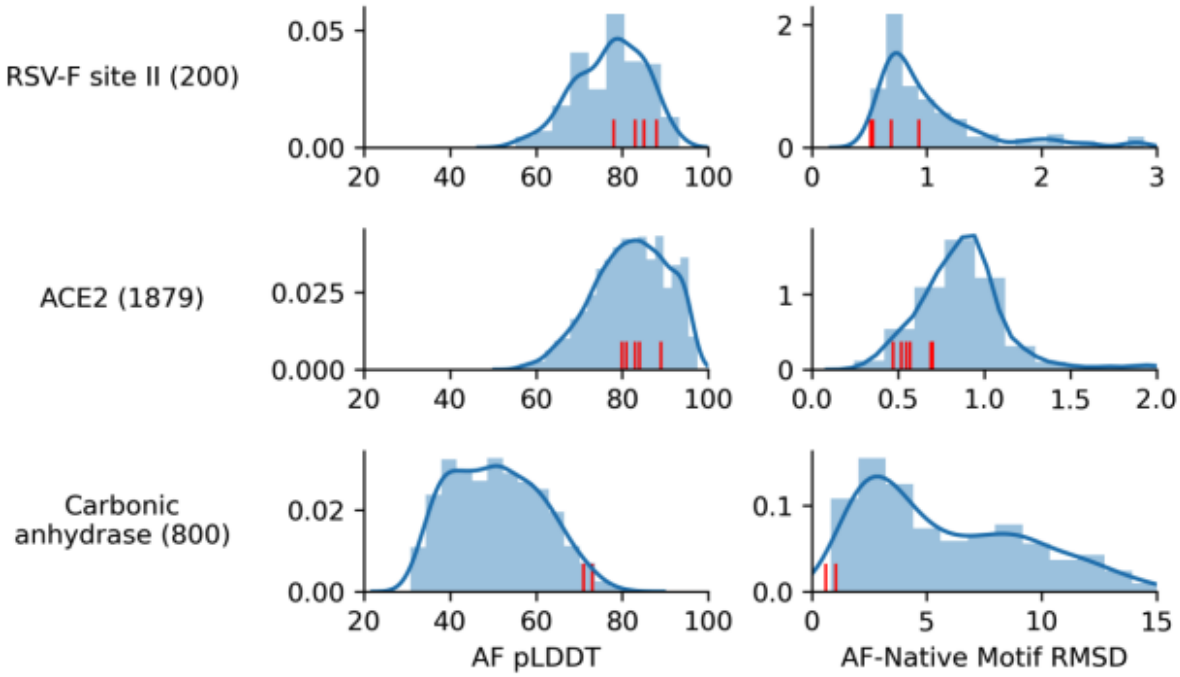



**Figure 2.S10. Distribution of plDDT and motif RMSD of hallucinations before filtering**
Distributions of (A) AlphaFold pLDDT and (B) backbone RMSD between native motif and AF predictions from hallucinated sequences, for design problems presented in Fig. 2-3. Parentheses indicate the number of designs. Red lines indicate designs filtered and chosen for display in main figures.

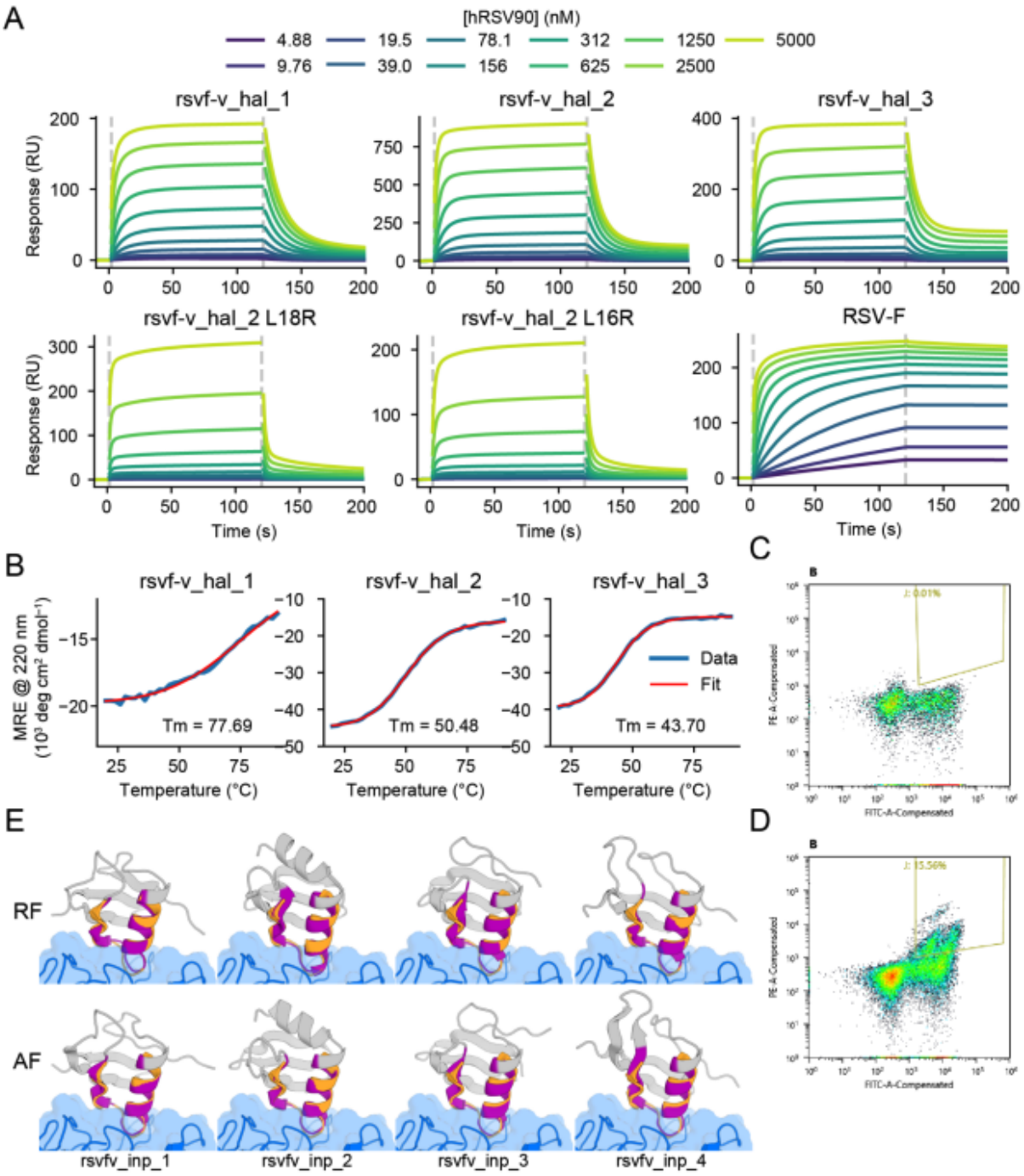



**Figure 2.S11. Experimental characterization of RSV-F site V scaffolds**
(A) Binding response (response units) versus time on SPR for RSV-F site V designs, point mutants, and control RSV F protein. Computed KD values are shown in Fig. 2. (B) Mean residue ellipticity at 220 nm versus temperature from CD. Melting points (Tm) values calculated from a two-state fit are shown in the inset. (C) CD spectra of point mutants with complete loss of activity. (D-E) Compensated PE-A (hRSV90 binding) versus compensated FITC-A (yeast

display) for a pool of 56 RSV-F site V inpaints with (D) no target or (E) 100nM binding target. (F) RosettaFold (RF) and Alphafold2 (AF) models of inpainted designs recovered from the sorted cells in (E).

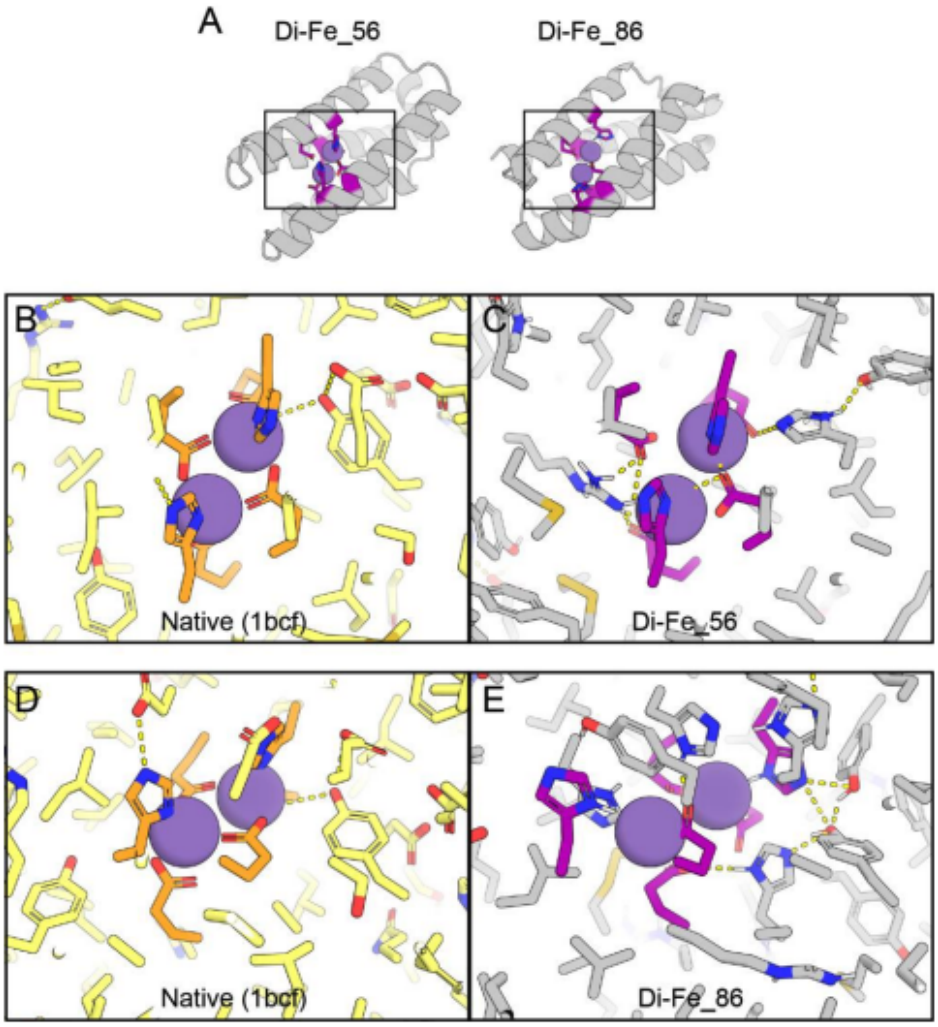



**Figure 2.S12. Di-iron hallucinations containing buried hydrogen-bond networks**
(A) Two di-iron hallucinations and close-ups (C, E) of the residues near the metal binding site. Structures are AF predictions after AMBER relax (80). The native protein used as a hallucination reference is shown in (B, D) after aligning to the hallucinations on the backbone atoms of the functional residues (orange in native, purple in hallucinations). Metals shown in (C, E) are taken from the native structure after superimposition. Note the presence of hallucinated polar residues (gray histidines and tyrosines) to form hydrogen-bonding networks with the functional histidines and glutamates, which were constrained to their native identities during hallucination.

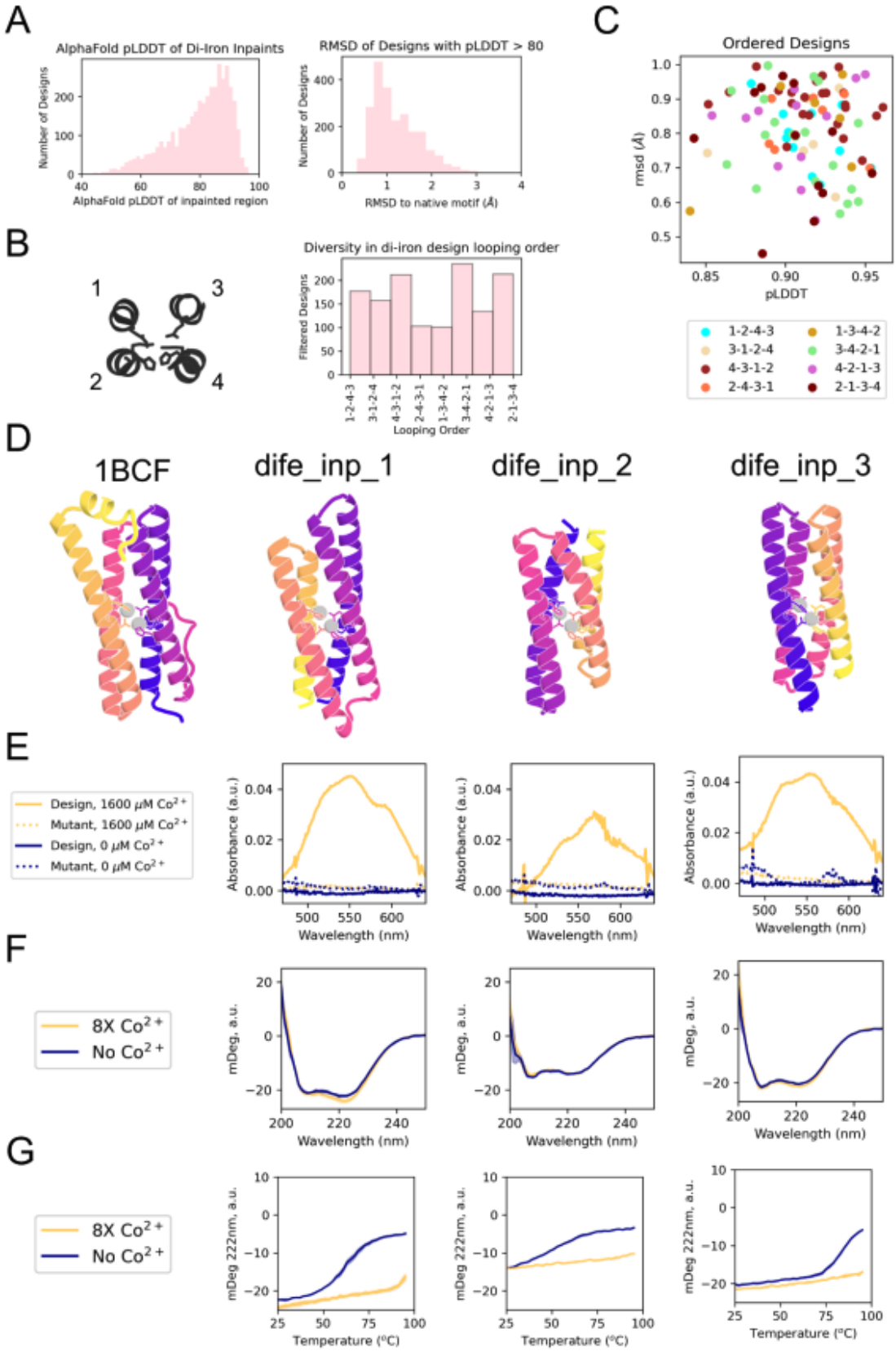

A
AlphaFold pLDDT of Di-Iron Inpaints
Number of Designs
AlphaFold pLDDT of inpainted region
RMSD of Designs with pLDDT > 80
Number of Designs
RMSD to native motif (Å)
B
1
2
3
4
Diversity in di-iron design looping order
Filtered Designs
1-2-4-3
3-1-2-4
4-3-1-2
2-4-3-1
1-3-4-2
3-4-2-1
4-2-1-3
2-1-3-4
Looping Order
C
Ordered Designs
rmsd (Å)
pLDDT
1-2-4-3
3-1-2-4
4-3-1-2
2-4-3-1
1-3-4-2
3-4-2-1
4-2-1-3
2-1-3-4
D
1BCF
dife_inp_1
dife_inp_2
dife_inp_3
E
Design, 1600 μM Co2+
Mutant, 1600 μM Co2+
Design, 0 μM Co2+
Mutant, 0 μM Co2+
Absorbance (a.u.)
Wavelength (nm)
F
8X Co2+
No Co2+
mDeg, a.u.
Wavelength (nm)
G
8X Co2+
No Co2+
mDeg 222nm, a.u.
Temperature (°C)

**Figure 2.S13. A subset of successful di-iron binding proteins designed with RFjoint**
A total of 4000 inpainted designs haboring the bacterioferritin (1BCF) di-iron binding site and encompassing 8 unique looping orders were generated with RFjoint. (A) 57.9% of outputs had AlphaFold pLDDT in the inpainted region > 80 (left), and 43.7% of these designs had a predicted RMSD to the input motif < 1 Å (right). (B) All 8 looping orders produced designs with AlphaFold pLDDT > 80 and motif AF-RMSD < 1 Å. Looping orders are with respect to residueindices in the native bacterioferritin protein (left). (C) After filtering and modest sequence optimization with RFjoint (see supplementary methods), 96 designs were ordered encompassing all 8 looping orders. (D-G) Characterization of three successful designs. (D) AlphaFold predictions of the three designs (right-most three designs), colored to highlight the different looping orders from the native bacterioferritin (left). Iron atoms, aligned to the motif, are depicted in gray for clarity. (pLDDT/Motif AF-RMSD: dife_inp_1: 92/0.65 Å; dife_inp_2: 94/0.64 Å; dife_inp_3: 90/0.76 Å) (E) Designs at 200 μM were incubated with an 8X molar excess of $CoCl_2$. All three designs show absorbance spectra consistent with $Co^{2+}$ binding in a tetra/pentacoordinate state to the designs (solid yellow lines). Such absorbance was not present in the absence of $Co^{2+}$ (solid blue lines), or with mutant designs where the 6 coordinating residues were mutated to alanine (dashed yellow lines). (F) All designs showed circular-dichroism (CD) spectra consistent with helical proteins. (G) Analysis of protein stability by CD-melts. All three designs were stabilized by binding to metal ions (8X molar excess of $Co^{2+}$). Note that dife_inp_1 data (E-G) is the same as in Figure 3, reproduced here for convenience.

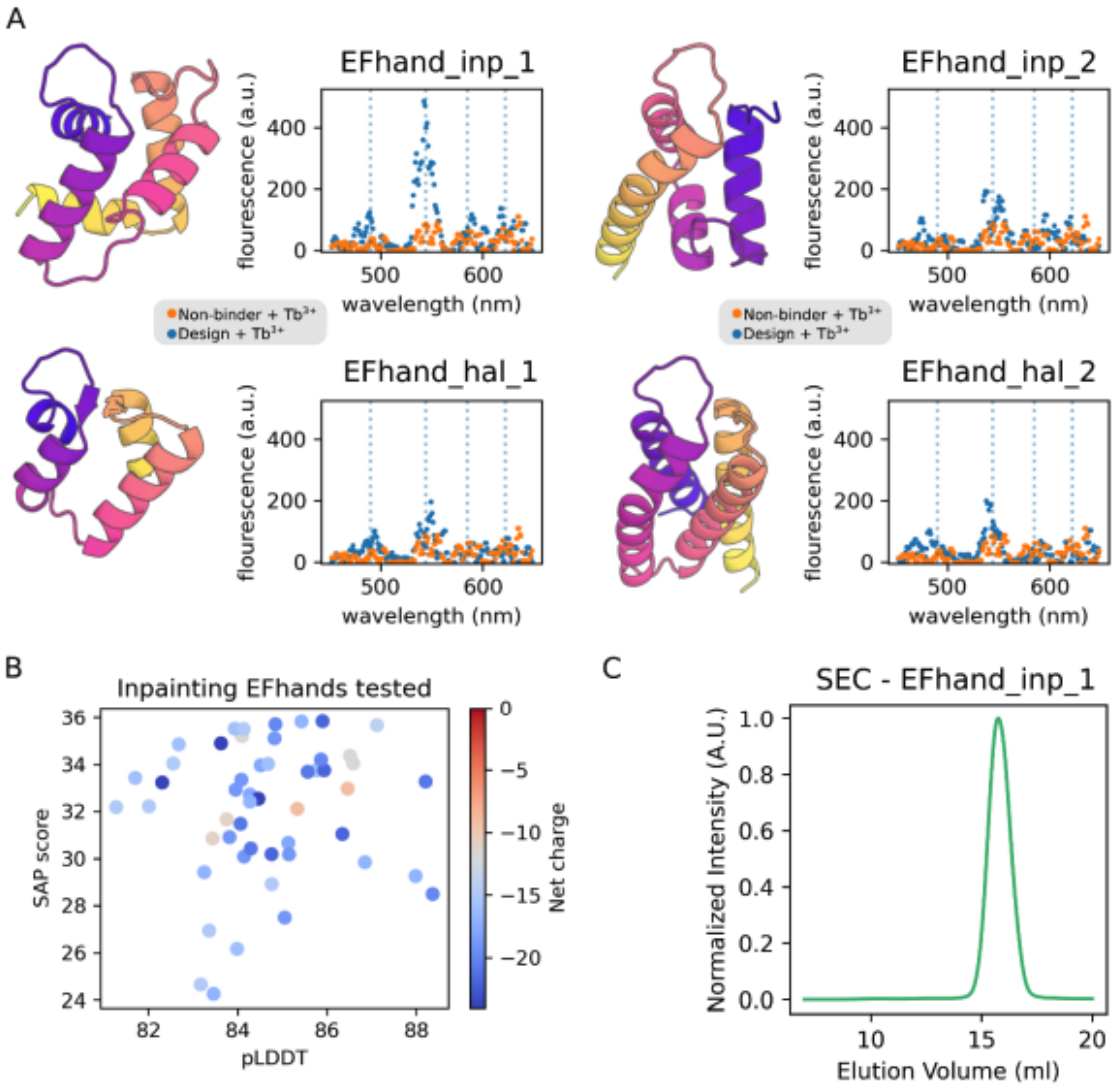



**Figure 2.S14. Characterization of EF-hand designs**
Experimental and computational characterization of EF-hand designs tested experimentally. (A) AF2 prediction of inpainted proteins EFhand_inp_1 and EFhand_inp_2 (top row) and hallucinated proteins EFhand_hal_1 and EFhand_hal_2 (bottom row) next to their terbium fluorescence spectra from a yeast-based initial screen (Materials and Methods). The same negative control spectrum (PDB accession 4DT5, orange) is duplicated across all plots. (B) Computational metrics of inpainted EF-hand designs from RFjoint that were tested by yeast display. In addition to standard filters like motif AF-RMSD and AF2 pLDDT, designs were also filtered by their SAP score and net charge. (C) Size exclusion chromatogram at 280 nm absorbance for EFhand_inp_1 suggests the protein occupies a stable monomeric state.

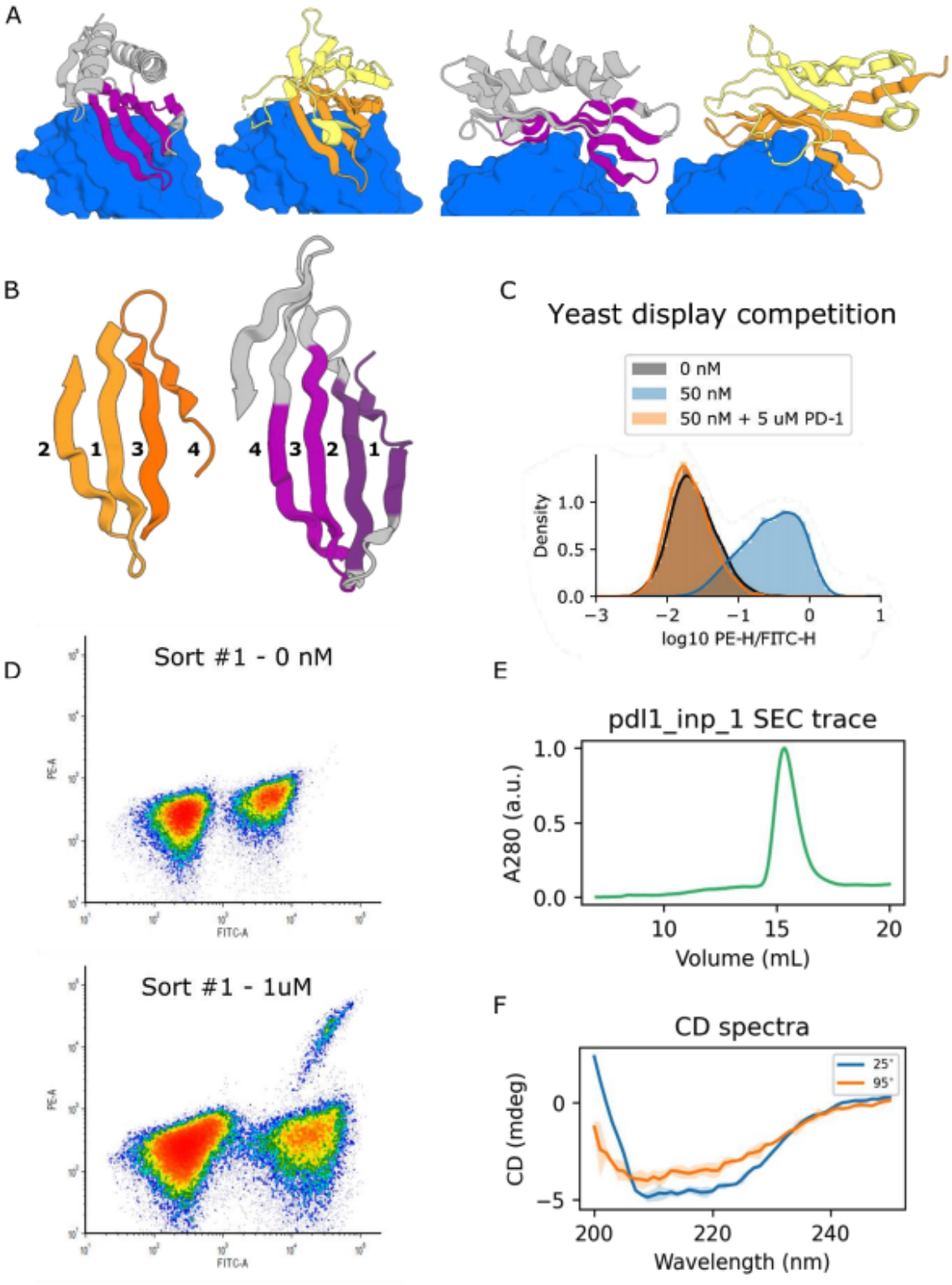



**Figure 2.S15. Experimental characterization of inpainted PD-L1 binder**
(A) Crystal structure of HAC PD-1 (binding interface motif in orange) in complex with PD-L1 (blue) and design model of pdl1_inp_1 (motif in purple). The overall fold of the design is quite

different from HAC PD-1, as the former contains two buttressing helices against the interfacial sheet instead of the original beta-sandwich. The design also includes an additional beta strand which extends the sheet in its C-terminus. (B) The looping order of the interfacial beta strands in the design (purple / dark purple) has changed dramatically from the HAC PD-1, demonstrating the ease of relooping secondary structure elements while maintaining the desired motif with inpainting. Notably, the order in which the two discontiguous strand-loop-strand submotifs appear in primary structure has switched, as well as the order in which strands 3 and 4 from HAC PD-1, which become strands 1 and 2 in the design, respectively. (C) Binding signal (PE-H) normalized to yeast surface expression (FITC-H) of clonal yeast population displaying pdl1_inp_1 labeled with 0 or 50 nM PD-L1, or 50 nM PD-L1 + 5 μM unlabeled PD-1. Loss of binding upon PD-1 competition suggests that pdl1_inp_1 binds PD-L1 at the native PD-1 binding site. (D) Fluorescence activated cell sorting data from yeast display binding experiments. Titles denote the concentration of a disulfide linked homodimeric PD-L1 target present in the binding reaction. Sort #1 denotes the first pooled sort of 31 designs, Sort #2 denotes the second sort performed with the enriched population of yeast displaying binding activity from Sort #1.

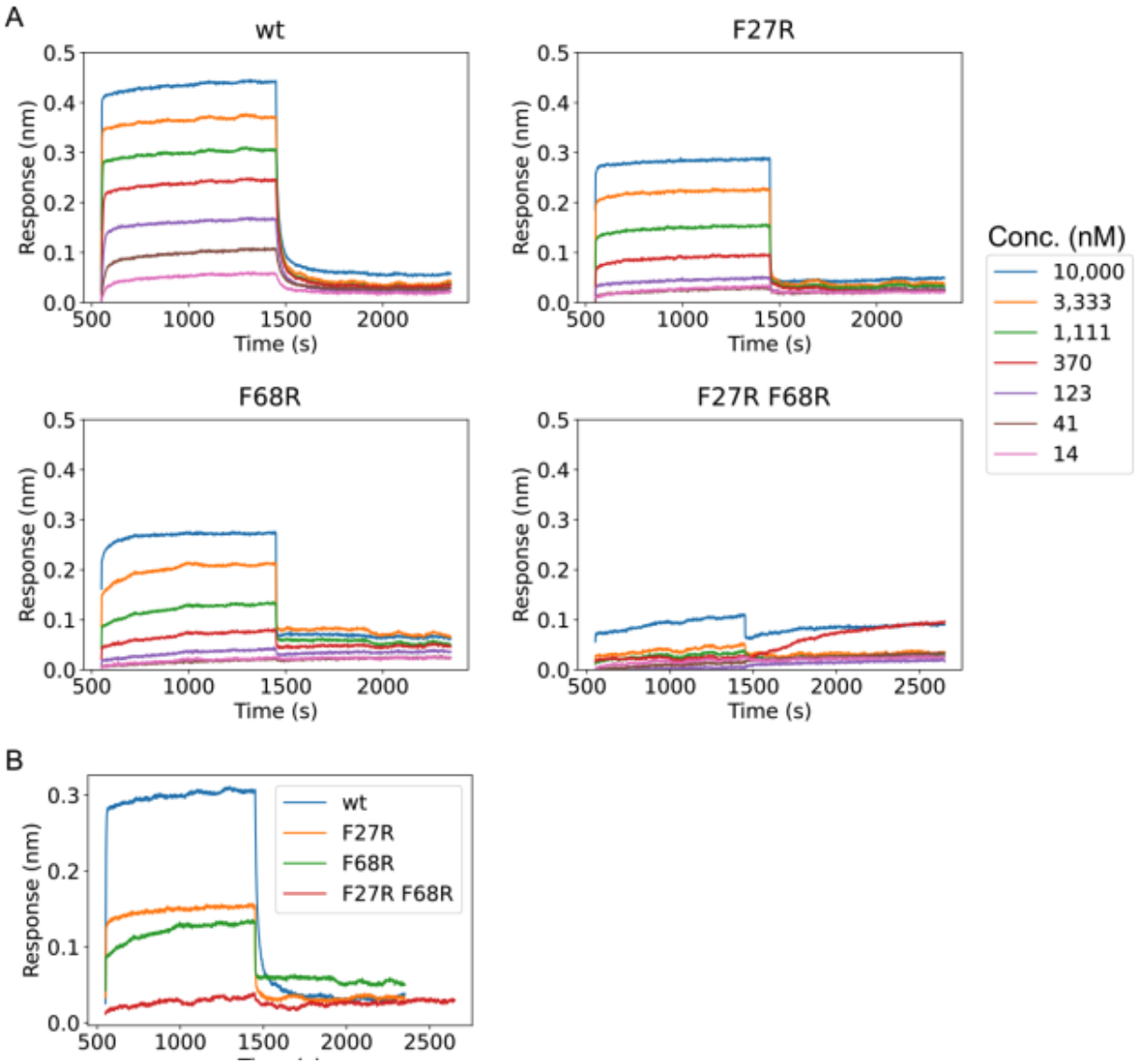



**Figure 2.S16. Experimental characterization of a bivalent TrkA binder**
(A) Association and dissociation kinetics of several TrkA binder variants as measured by biolayer interferometry. WT is the designed binder, F27R and F68R are mutants knocking out either one of the designed binding interfaces, and F27R F68R is a double-mutant knocking out both interfaces. (B) Kinetic traces of all four TrkA binder variants compared at the same concentration (1111 nM) show that wt binds the most TrkA, both single site mutants bind similar amounts, and the double mutant binds negligible amounts of TrkA. These data show that either binding site is sufficient to bind TrkA, indicating that we successfully made a bivalent TrkA binder.

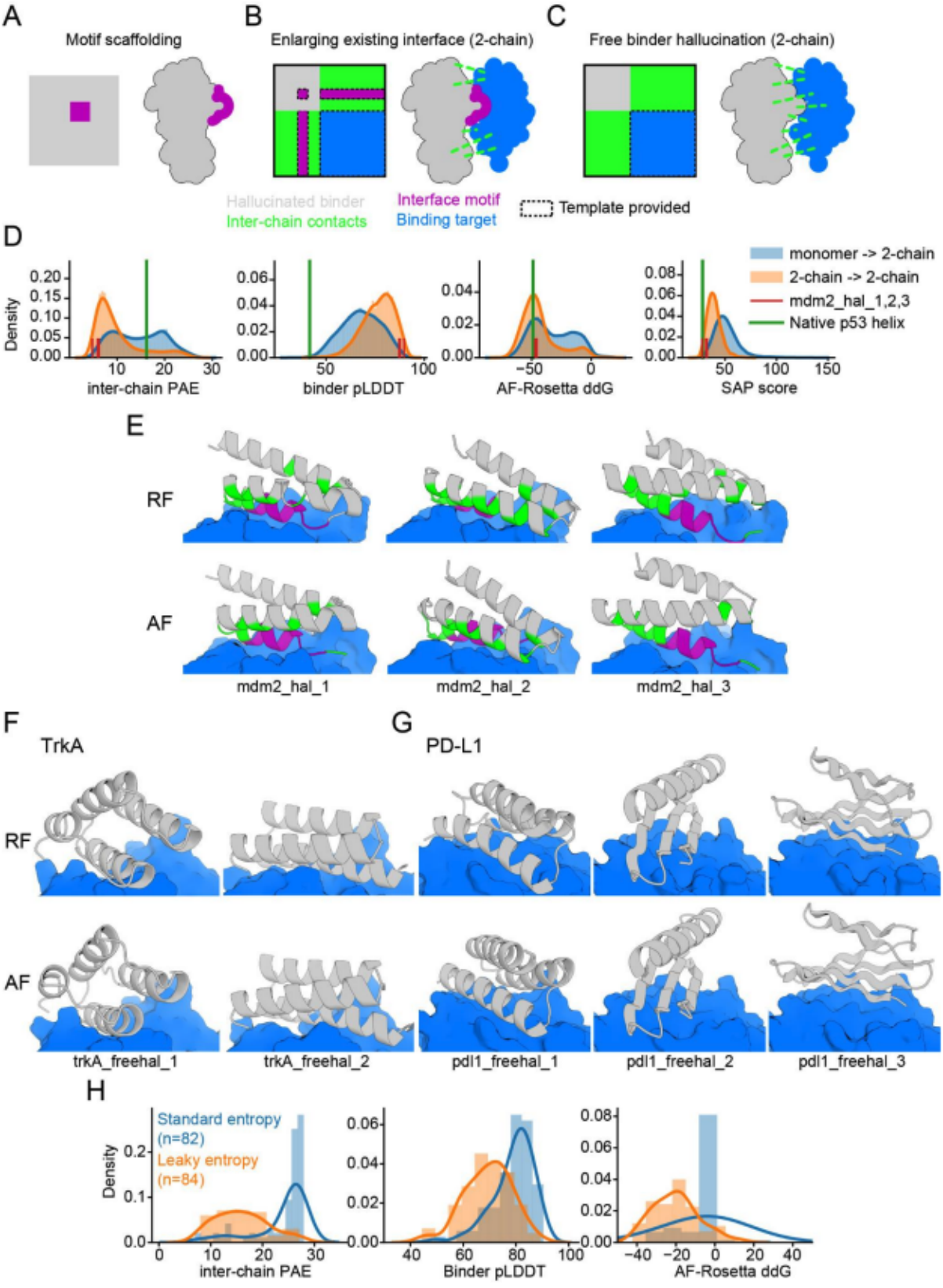

A
Motif scaffolding
B
Enlarging existing interface (2-chain)
C
Free binder hallucination (2-chain)
Hallucinated binder
Inter-chain contacts
Interface motif
Binding target
Template provided
D
Density
inter-chain PAE
binder pLDDT
AF-Rosetta ddG
SAP score
monomer -> 2-chain
2-chain -> 2-chain
mdm2_hal_1,2,3
Native p53 helix
E
RF
AF
mdm2_hal_1
mdm2_hal_2
mdm2_hal_3
F
TrkA
G
PD-L1
RF
AF
trkA_freehal_1
trkA_freehal_2
pdl1_freehal_1
pdl1_freehal_2
pdl1_freehal_3
H
Standard entropy
(n=82)
Leaky entropy
(n=84)
Density
inter-chain PAE
Binder pLDDT
AF-Rosetta ddG

**Figure 2.S17. Multi-chain hallucination for binder design Schematic of the variations on binder hallucination methods. Hallucinated binder gray, binding partners blue, motifs purple.** (A) Motif scaffolding (B) Motif scaffolding while enlarging existing interfaces. (C) Free binder hallucination. (D) Design metrics of 17,450 Mdm2 binder hallucinations. “Monomer -> 2-chain” are designs after one round of two-chain MCMC refinement starting from high-scoring hallucinated monomers (Supplementary Text). “2-chain -> 2-chain” are designs after an additional round of filtering and MCMC refinement. Metrics for the native p53 helix and the 3 highlighted designs are shown in green and red lines, respectively. (E) RF and AF design models of the Mdm2 binder designs shown in Fig. 5G. New binding interactions (hallucinated residues within 5 Å of the target) are in green. (E) Free-hallucinated TrkA and (F) PD-L1 binder designs. (G) Design metrics for free-hallucinated PD-L1 binders using the “leaky” entropy loss (orange), compared to the standard entropy loss (blue) (Supplementary Text).

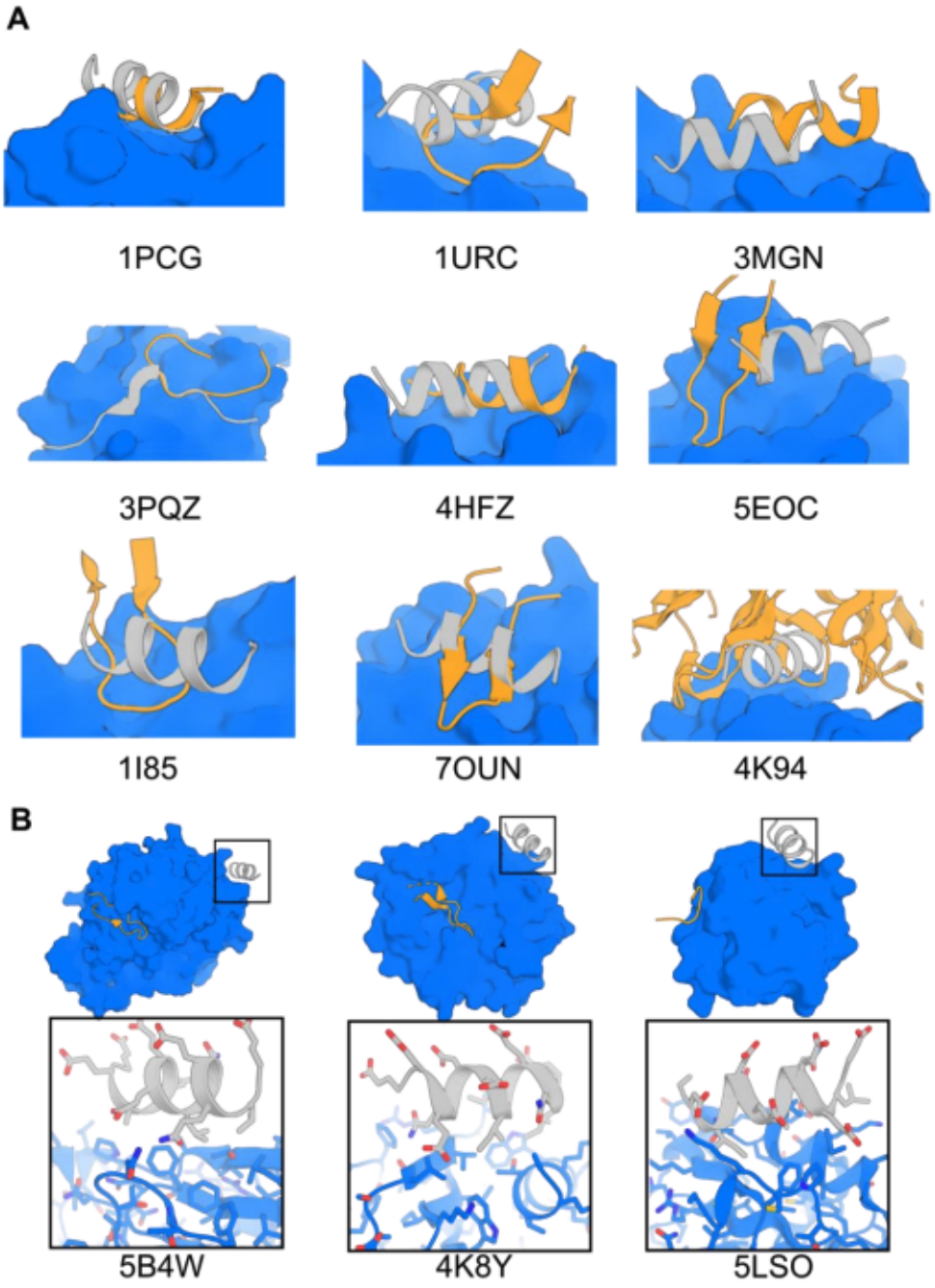



**Figure 2.S18. Free hallucination 12 residue stub placement on native targets** (A) Freely hallucinated 12 residue stubs against native proteins. Gray hallucinated stub; Orange native binder. (B) Hallucinated stubs network docked on alternative hydrophobic grooves to those of native binders. Boxed structures show side chains packing against targets. Structure PDB IDs listed.

## 2.14 Supplementary Tables

**Table 2.S1. Natural proteins used for mimetic design**

"Motif residues" indicate residues that were constrained to native geometry during hallucination. Sometimes only a subset of the motif residues actually comprise a binding interface or catalytic site; these are denoted "functional residues".

| Native protein (Reference) | PDB ID | Chain | Motif residues | Functional residues | Binding partner(s) |
|---|---|---|---|---|---|
| HAC PD-1 (*15*) | 5IUS | A | A63-82, A119-140 | A64, 66, 68, 70, 73-75, 77-78, 81, 85, 89-91, 124, 126, 128, 132, 134, 136, 139 | PD-L1 |
| RSV-F site II (*81*) | 3IXT | P | P254-277 | | Antibody |
| RSV-F site V (*27*) | 5TPN | A | A163-181 | | Antibody |
| ACE2 (*82*) | 6VW1 | A | A24-42 | | SARS-CoV2 receptor binding domain |
| EF-hand (*83*) | 1PRW | A | A21-31,A56-67 | A21-31,A56-67 | $Ca^{2+}$ |
| Di-Fe (*29*) | 1BCF | A | A18-25,A47-54,A94-97,A123-130 | A18, 51, 54, 94, 127, 130 | $Fe^{2+}$ |
| Carbonic anhydrase II (*84*) | 5YUI | A | A62-65,A93-97,A118-120 | A94,A96,A119,A199 | $Zn^{2+}$ |
| $\Delta^5$-3-ketosteroid isomerase (*36*) | 1QJG | A | A14,A38,A99 | A14,A38,A99 | equilenin |
| p53 N-term helix (*85*) | 1YCR | B | B17-27 | A19, 23, 26, 27 | Mdm2 |
| TrkA minibinder (*4*) | 7N3T | A | A5-18 | A5, 6, 9, 10, 12, 13, 14, 16, 17, 18 | TrkA |

**Table 2.S2. RMSDs between native protein, design model, and AlphaFold model**

All RMSDs are in angstroms. Columns in red are the metrics reported in the main text and figures. RMSD values in parentheses (for hcA and KSI) are full-atom RMSDs over the catalytic sidechains. KSI designs are generated using AF, and "Design" refers to models generated using

the ensembling approach over AF models 1,2,3,5 and "AF" refers to AF model 4 (Materials and methods).

| | Overall | | Motif | | |
|---|---|---|---|---|---|
| Design | AF pIDDT | RMSD, Design to AF | RMSD, Design to AF | RMSD, Design to native | RMSD, AF to native |
| rsvfv_hal_1 | 82 | 1.37 | 1.06 | 1.31 | 0.7 |
| rsvfv_hal_2 | 88 | 0.75 | 0.34 | 0.67 | 0.64 |
| rsvfv_hal_3 | 86 | 0.85 | 0.24 | 0.65 | 0.65 |
| rsvf-v_854 | 82 | 2.45 | 0.65 | 0.71 | 0.75 |
| rsv_inp_1 | 83 | 0.91 | 0.5 | 0.51 | 0.59 |
| rsv_inp_2 | 83 | 0.76 | 0.57 | 0.6 | 0.81 |
| rsv_inp_3 | 88 | 1.14 | 0.55 | 0.74 | 0.85 |
| rsv_inp_4 | 81 | 1.69 | 0.64 | 0.5 | 0.87 |
| dife_inp_1 | 92 | 0.3 | 0.24 | 0.61 | 0.65 |
| dife_inp_1_mutant | 87 | n/a | n/a | n/a | 0.71 |
| dife_inp_2 | 94 | 0.91 | 0.39 | 0.54 | 0.64 |
| dife_inp_2_mutant | 95 | n/a | n/a | n/a | 0.79 |
| dife_inp_3 | 90 | 0.54 | 0.31 | 0.72 | 0.76 |
| dife_inp_3_mutant | 92 | n/a | n/a | n/a | 0.89 |
| dife_inp_4 | 88 | 1.04 | 0.77 | 0.32 | 0.85 |
| dife_inp_5 | 90 | 0.82 | 0.67 | 0.39 | 0.71 |
| dife_inp_6 | 93 | 0.77 | 0.39 | 0.99 | 0.92 |
| dife_inp_7 | 95 | 0.4 | 0.27 | 0.64 | 0.68 |
| dife_inp_8 | 90 | 0.72 | 0.62 | 0.31 | 0.8 |
| Di-Fe_86 | 84 | 1.97 | 0.89 | 0.4 | 0.9 |
| Di-Fe_56 | 84 | 2.28 | 0.74 | 0.46 | 0.87 |
| EFhand_inp_1 | 87 | 0.86 | 0.82 | 0.29 | 0.69 |
| EFhand_inp_2 | 87.5 | 1.7 | 0.3 | 0.8 | 0.7 |

| | | | | | |
|---|---|---|---|---|---|
| EFhand_hal_1 | 82.2 | 1.42 | 0.59 | 0.36 | 0.52 |
| EFhand_hal_2 | 82.8 | 0.76 | 0.47 | 0.55 | 0.73 |
| hcA_1 | 73 | 1.44 | 0.73 (2.23) | 0.75 (1.39) | 1.04 (1.97) |
| hcA_2 | 71 | 1.62 | 0.46 (1.74) | 0.46 (1.36) | 0.62 (2.02) |
| ksi_1 (AF) | 84 | 1.04 | 0.30 (0.30) | 0.30 (1.22) | 0.30 (1.20) |
| ksi_2 (AF) | 72 | 1.06 | 0.16 (0.22) | 0.43 (1.63) | 0.53 (1.65) |
| pdl1_inp_1 | 84 | 0.79 | 0.51 | 1 | 1.1 |
| trkA_56 | 89 | 2.53 | 2.06 | 1.15 | 2.34 |
| mdm2_hal_1 | 88.6 | 1.70 | 1.75 | 0.73 | 1.29 |
| mdm2_hal_2 | 84.1 | 1.95 | 0.83 | 0.59 | 0.63 |
| mdm2_hal_3 | 81.7 | 1.14 | 1.00 | 0.77 | 0.68 |

**Table 2.S3. Interface metrics of protein-binder designs**

AlphaFold inter-PAE, binder pLDDT, AF-Rosetta ddG, and target-aligned binder RMSD (Materials and Methods) for protein-binder designs presented in this paper. Note that designs based off of motifs are listed here and in Table S2, but the free hallucinations are only shown here. pdl1_inp_1 and trkA_56 were not designed using 2-chain hallucination, so there were no RF complex design models to use for target-aligned binder RMSD calculations.

| Design | Inter-PAE | Binder pLDDT | AF-Rosetta ddG | Target-aligned binder RMSD |
|---|---|---|---|---|
| **pdl1_inp_1** | 5.695 | 88.5 | -49.9 | N/A |
| **trkA_56** | 8.428 | 88.4 | -51.8 | N/A |
| **mdm2_hal_1** | 5.904 | 87.6 | -47.2 | 2.93 |
| **mdm2_hal_2** | 4.822 | 89.7 | -45.8 | 3.36 |
| **mdm2_hal_3** | 6.208 | 87.1 | -45.9 | 3.48 |
| **trkA_freehal_1** | 6.40 | 87.4 | -32.5 | 3.87 |
| **trkA_freehal_2** | 4.63 | 92.1 | -35.8 | 1.24 |
| **pdl1_freehal_1** | 5.58 | 84.8 | -38.23 | 3.43 |
| **pdl1_freehal_2** | 9.72 | 82.3 | -26.36 | 1.58 |
| **pdl1_freehal_3** | 8.87 | 81.0 | -37.15 | 1.59 |

**Table 2.S4. Frequency of suitable native scaffolds**

| Native protein | PDB ID | Chain | Motif residues | Scaffolds in the PDB with <1Å motif RMSD | |
|---|---|---|---|---|---|
| | | | | Number | Frequency |
| RSV-F site II | 3IXT | P | P254-277 | 0 | 0 |
| RSV-F site V | 5TPN | A | A163-181 | 1 | 3.76e-05 |
| ACE2 | 6VW1 | A | A24-42 | 1874 | 7.05e-02 |
| EF-hand (double) | 1PRW | A | A21-31,A56-67 | 30 | 1.13e-03 |
| EF-hand (single) | 1PRW | A | A56-67 | 77 | 2.90e-03 |
| Di-iron | 1BCF | A | A18-25,A47-54,A94-97,A123-130 | 3 | 1.13e-04 |
| Carbonic anhydrase II | 5YUI | A | A62-65,A93-97,A118-120 | 1 | 3.76e-05 |
| C3d | 1GHQ | A | A104-126,A170-185 | 2 | 7.52e-05 |
| HAC PD-1 | 5IUS | A | A63-82, A119-140 | 56 | 2.11e-03 |

**Table 2.S5. Similarity of designs to native proteins**

Designed proteins were compared to protein in the PDB and the Facebook AF2 models database (64) for structural and sequence similarity with TM-align (71) and blastp (86), respectively (Materials and Methods). TMalign "% ID" refers to the number of identities over the aligned region divided by the number of aligned residues. BLAST "% ID" refers to the number of identities over the best HSP, normalized to the length of the query sequence (design).

| | TMalign to PDB | | | TMalign to FB AF2 | | | BLAST to NR | | |
|---|---|---|---|---|---|---|---|---|---|
| Design | Top hit | TM score | % ID | Top hit | TM score | % ID | Top hit | E-value | % ID |
| dife_inp_1 | 5vju_A | 0.89 | 9.2 | A0A4S8HXL5 | 0.87 | 8.5 | None | NA | NA |
| dife_inp_1_mutant | 5vju_A | 0.88 | 9.2 | A0A4S8HXL5 | 0.90 | 9.2 | None | NA | NA |
| dife_inp_2 | 7jic_B | 0.84 | 3.5 | A0A328DJV2 | 0.87 | 9.7 | WP_000675503.1 | 2.24E-02 | 23 |
| dife_inp_2_mutant | 7jic_B | 0.82 | 3.5 | A0A1D2MQT9 | 0.89 | 8.8 | None | NA | NA |
| dife_inp_3 | 1yo7_A | 0.85 | 12.3 | A0A3S2NBQ8 | 0.87 | 12.5 | None | NA | NA |
| dife_inp_3_mutant | 4phq_B | 0.83 | 13 | A0A2N1PYQ6 | 0.89 | 11.6 | None | NA | NA |
| dife_inp_4 | 6egc_A | 0.84 | 22.4 | A0A1J0A759 | 0.87 | 5.7 | None | NA | NA |
| dife_inp_5 | 6egc_A | 0.85 | 14.8 | A0A1D8FWU8 | 0.84 | 18.6 | None | NA | NA |
| dife_inp_6 | 5vjs_A | 0.80 | 10.2 | A0A131Z7Y1 | 0.87 | 4.7 | None | NA | NA |
| dife_inp_7 | 5vjs_A | 0.84 | 10.5 | A0A1D1UXZ2 | 0.87 | 9.7 | None | NA | NA |
| dife_inp_8 | 5vju_A | 0.85 | 30.7 | J9JP71 | 0.88 | 9.9 | 2LFD_A | 9.30E-03 | 22 |
| rsv_inp_1 | 5a2q_G | 0.61 | 16.1 | A0A2V8WE05 | 0.63 | 15.7 | 1G2C_A | 5.40E-01 | 40 |
| rsv_inp_2 | 6apd_B | 0.64 | 42.1 | A0A2V9VQU5 | 0.65 | 11.1 | XP_021434148.1 | 6.38E+00 | 30 |
| rsv_inp_3 | 5clr_A | 0.60 | 10.2 | I2B993 | 0.69 | 13.3 | WP_120068072.1 | 1.24E+00 | 29 |
| rsv_inp_4 | 5g4y_A | 0.67 | 11.5 | A0A2N5ZAK5 | 0.66 | 8.1 | WP_159887573.1 | 5.70E+00 | 35 |
| trkA_56 | 2d4c_A | 0.80 | 8.1 | UPI00083C0126 | 0.83 | 8.1 | None | NA | NA |
| rsvfv_hal_1 | 6ntr_D | 0.69 | 10.4 | A0A2H6GLY6 | 0.77 | 10.1 | 3KPE_A | 8.90E-01 | 26 |
| rsvfv_hal_2 | 4dmg_A | 0.71 | 11.9 | A0A290HYD2 | 0.78 | 12.1 | RZV56203.1 | 3.12E+00 | 20 |
| rsvfv_hal_3 | 4auk_A | 0.69 | 11.4 | R7HWW9 | 0.76 | 12.9 | WP_154333053.1 | 4.53E+00 | 31 |
| rsvf-v_854 | 5wb0_F | 0.58 | 21.9 | UPI000B354BFA | 0.67 | 8.6 | 1G2C_A | 5.80E-02 | 27 |
| rsvf-v_870 | 5csl_B | 0.67 | 16.7 | A0A413CFN9 | 0.72 | 11.5 | 1G2C_A | 2.63E+00 | 27 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| rsvf-v_828 | 6cp8_A | 0.62 | 13.6 | UPI0009045699 | 0.67 | 10.5 | None | NA | NA |
| rsvf-v_903 | 2x32_B | 0.63 | 5 | UPI0011AE9EE2 | 0.74 | 8.2 | 3KPE_A | 1.56E-01 | 32 |
| rsvf-v_1050 | 5wti_Z | 0.59 | 12.7 | A0A524IGV4 | 0.63 | 1.6 | AIZ95772.1 | 3.16E-02 | 32 |
| rsvf-ii_141 | 6ivm_A | 0.68 | 10 | A0A366EM18 | 0.74 | 9.1 | HHG91166.1 | 6.16E+00 | 17 |
| rsvf-ii_171 | 5j0l_E | 0.86 | 12.5 | A0A1Y6CLD7 | 0.89 | 8.7 | AWV19065.1 | 3.23E-01 | 34 |
| rsvf-ii_118 | 2yfa_A | 0.74 | 15.4 | A0A0M0J6I0 | 0.81 | 9 | CCW60917.1 | 1.54E+00 | 27 |
| rsvf-ii_29 | 4jeh_B | 0.78 | 9.1 | R6XLH6 | 0.82 | 4.5 | RKX18559.1 | 2.83E-01 | 17 |
| rsvf-ii_158 | 2j0o_A | 0.86 | 11.1 | A0A354DBJ4 | 0.88 | 8.2 | WP_068486906.1 | 3.04E-01 | 29 |
| ace2_76 | 7jh6_A | 0.81 | 14.6 | A0A073CH21 | 0.86 | 10.4 | QIN87098.1 | 5.91E+00 | 21 |
| ace2_1157 | 2j0o_A | 0.76 | 11.8 | A0A1Y2MHD8 | 0.83 | 3.5 | WP_100023565.1 | 1.17E+00 | 24 |
| ace2_1007 | 5tqy_A | 0.80 | 10.5 | A0A4R7HW16 | 0.86 | 9.6 | None | NA | NA |
| ace2_1846 | 5iig_A | 0.82 | 9.4 | UPI00041B2217 | 0.81 | 10.3 | None | NA | NA |
| ace2_600 | 4q2g_B | 0.72 | 11.1 | R5GU22 | 0.81 | 5 | EPE07190.1 | 2.41E+00 | 27 |
| ace2_109 | 3zcj_B | 0.80 | 9.2 | A0A3N0EL48 | 0.85 | 6.7 | ROL44962.1 | 1.58E-01 | 22 |
| hcA_1 | 2hb0_A | 0.77 | 17.2 | A0A376L8Y0 | 0.75 | 14.8 | WP_107852251.1 | 3.10E-01 | 30 |
| hcA_2 | 6ohh_B | 0.79 | 15.5 | A0A3D3R120 | 0.81 | 10.5 | WP_021068970.1 | 6.97E+00 | 23 |
| ksi_1 (AF) | 5k59_B | 0.73 | 6.2 | M3UPS5 | 0.78 | 4.2 | WP_147602516.1 | 4.22E-01 | 21 |
| ksi_2 (AF) | 1z8k_A | 0.58 | 5.6 | Q66636 | 0.61 | 7.1 | KAF3849996.1 | 2.21E+00 | 25 |
| Di-Fe_86 | 6h2f_H | 0.76 | 5.9 | X0WN74 | 0.85 | 7.8 | None | NA | NA |
| Di-Fe_56 | 6ezv_X | 0.75 | 3.5 | A0A399XE29 | 0.78 | 3.5 | TGO06933.1 | 1.45E+00 | 19 |
| pdl1_inp_1 | 5ldz_F | 0.61 | 8.3 | A0A2N1TGW6 | 0.67 | 6.7 | WP_071803821.1 | 3.28E-02 | 25 |
| EFhand_inp_1 | 4by5_B | 0.75 | 20.7 | A0A2E7SWA3 | 0.77 | 21.4 | XP_020433196.1 | 1.38E-17 | 52 |
| EFhand_inp_2 | 1juo_A | 0.65 | 15.9 | UPI00052857BB | 0.72 | 23.2 | None | 9.52E-05 | 35 |
| EFhand_hal_1 | 2f8p_A | 0.72 | 26.3 | A0A0A1TVZ3 | 0.82 | 16.9 | XP_019463585.1 | 5.04E-02 | 36 |
| EFhand_hal_2 | 6afs_B | 0.67 | 3.1 | UPI0004131E18 | 0.76 | 12.2 | WP_092746209.1 | 6.09E-02 | 23 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| mdm2_hal_1 | 5h78_A | 0.77 | 17.9 | A0A2T4JJG5 | 0.85 | 15.8 | None | NA | NA |
| mdm2_hal_2 | 1fjg_T | 0.78 | 17.9 | A0A429CN45 | 0.86 | 16.4 | XP_012788760.1 | 2.38E+00 | 27 |
| mdm2_hal_3 | 6w2v_B | 0.86 | 16.7 | A0A1F7QLQ5 | 0.90 | 11.7 | XP_030199201.1 | 7.14E+00 | 27 |
| trkA_freehal_1 | 5wyl_A | 0.75 | 7.9 | UPI0012EDFAF2 | 0.81 | 4.7 | WP_165006269.1 | 3.64E+00 | 25 |
| trkA_freehal_2 | 2oku_A | 0.82 | 6.7 | A0A4R8UL89 | 0.83 | 9.5 | None | NA | NA |
| pdl1_freehal_1 | 3q5d_A | 0.77 | 12.1 | A0A292YNZ8 | 0.80 | 4.9 | WP_132874866.1 | 7.67E+00 | 33 |
| pdl1_freehal_2 | 4jhc_A | 0.78 | 10 | A0A521U212 | 0.86 | 10.3 | MSR05998.1 | 4.94E+00 | 25 |
| pdl1_freehal_3 | 2ygt_A | 0.75 | 3.3 | A0A538M5E7 | 0.79 | 11.9 | PVH99412.1 | 2.20E+00 | 32 |

# Chapter 3. Joint Generation of Protein Sequence and Structure with RoseTTAFold Sequence Space Diffusion

This section contains content previously published as: **Lisanza, S. L.** *et al.* Joint Generation of Protein Sequence and Structure with RoseTTAFold Sequence Space Diffusion. 2023.05.08.539766 Preprint at https://doi.org/10.1101/2023.05.08.539766 (2023).

## 3.1 Abstract

Protein denoising diffusion probabilistic models (DDPMs) show great promise in the *de novo* generation of protein backbones, but are limited in their inability to guide generation of proteins with sequence specific attributes and functional properties. To overcome this limitation, we develop ProteinGenerator, a sequence space diffusion model based on RoseTTAfold that simultaneously generates protein sequences and structures. Beginning from random amino acid sequences, our model generates sequence and structure pairs by iterative denoising, guided by any desired sequence and structural protein attributes. To explore the versatility of this approach, we designed proteins enriched for specific amino acids, with internal sequence repeats, with masked bioactive peptides, with state dependent structures, and with key sequence features of specific protein families. ProteinGenerator readily generates

sequence-structure pairs satisfying the input conditioning (sequence and/or structural) criteria, and experimental validation showed that the designs were monomeric by size exclusion chromatography (SEC), had the desired secondary structure content by circular dichroism (CD), and were thermostable up to 95°C. By enabling the simultaneous optimization of both sequence and structure, ProteinGenerator allows for the design of functional proteins with specific sequence and structural attributes, and paves the way for protein function optimization by active learning on sequence-activity datasets.

## 3.2 Main

Protein function arises from a complex interplay of sequence and structural features, hence designing new protein functions requires reasoning over both sequence and structure space. Many protein design methods sample structures and sequences in separate steps, typically by generating protein backbones first and using inverse folding methods to generate sequences. Traditional methods like Rosetta flexible backbone protein design[5] alternate between structure and sequence design, while recent deep learning based approaches typically generate backbones first and then use sequence design methods such as ProteinMPNN to identify sequences that fold into a given backbone[11–13,73]. Among the latter class of approaches, denoising diffusion probabilistic models[20] (DDPMs), which have shown considerable promise in continuous data domains allow for the generation of protein backbones subject to a wide range of structural constraints[74–76]. DDPMs approximate the probability density function over a data distribution by learning to denoise samples corrupted with Gaussian noise, enabling the generation of high-quality samples from a Gaussian prior; they have been explored less in categorical domains such as text and protein sequences[11,12,73,77]. Simultaneous generation of sequence and structure could have advantages over methods that alternate between optimization in the two domains independently by enabling coordinated guidance with both sequence and structural features. Hallucination approaches that apply activation-maximization to structure prediction networks[16,17,78] can generate sequence-structure pairs without additional training, but these solutions can be adversarial, require a large number of steps to

converge, and robust experimental success requires subsequent sequence design on the hallucinated backbones[73].

We reasoned that diffusion approaches could be powerful for simultaneous generation of sequence and structure while avoiding the adversarial solutions of activation maximization, and set out to develop a diffusion model which jointly generates sequence-structure pairs and can be guided by constraints in both domains. We hypothesized that RoseTTAfold's ability to simultaneously generate protein sequences and structures, as illustrated by RoseTTAFold Joint Inpainting[17], could be adapted for diffusive generation of coherent sequence-structure pairs by finetuning to recover noised native protein sequences while imposing a loss on structure prediction accuracy, and that such a DDPM could be readily guided by constraints in both domains.

## 3.3 DDPM Implementation

We chose to implement diffusion in sequence space by representing amino acid sequences as scaled one hot tensors where true values are set to 1 and all other values set to -1, allowing progressive corruption with Gaussian noise $N(\mu=0,\sigma=1)$[79,80]. This approach is advantageous over other categorical diffusion methods, where diffusion occurs within a learned embedding space of text[81,82], because it simplifies the use of raw sequence based classifiers for guidance. To finetune RoseTTAFold we input the protein sequences progressively noised according to a square root schedule[81], the corresponding time step, and optional structural information. We task the model to generate ground truth sequence-structure pairs by applying a categorical cross entropy loss to the predicted sequence (relative to the ground truth sequence) and FAPE structure loss on the predicted structure. Self-conditioning[79], which allows the model to condition on its previous prediction, was employed to improve training and inference performance. Protein generation begins with an Lx20 dimensional sequence of Gaussian noise, and at each timestep ($\mathbf{x_t}$) the model predicts $\mathbf{x_0}$ from $\mathbf{x_t}$, after which $\mathbf{x_0}$ is noised to $\mathbf{x_{t-1}}$ (Figure 1A, top panel). Conditioning information (guidance) can be combined with $\mathbf{x_0}$ to guide the model towards a constrained sequence space

using activity data, sequence specific potentials, secondary structure features, and more (Figure 1A, bottom)[83].

## 3.4 Unconditional Generation

Starting with a sequence of Gaussian noise, the model generates sequence-structure pairs with amino acid compositions similar to those of native proteins (Figure 1B, left). The generated sequences and structures are internally consistent: AlphaFold2 and ESMFold predictions of the structures adopted by the generated sequences are very close to the generated structures (Figure 1C, S1B) and confident (Figure S1A). Sampling from different noise distributions resulted in different amino acid frequencies and secondary structure compositions in the generated outputs[84] (Figure S1A, S2, S3, S4). Samples of unconditionally generated designs with 100aa, 200aa or 300aa length can be found in Figures S5, S6 and S7. In contrast, unconditionally generated designs on a not fine-tuned RoseTTAFold model are less confident (Figure S8). For the longer lengths the success rate of ProteinGenerator in generating sequences that fold to the designed structures is lower than that of the RoseTTAfold based structure diffusion method RFdiffusion[75] followed by ProteinMPNN[73]; this may reflect intrinsic differences between diffusion in sequence and structure space, or arise from differences in model training.

For experimental characterization, we unconditionally generated 70-80 residue proteins, filtered for high AF2 confidence (pLDDT > 90) and AF2 RMSD to design < 2Å (Table S4). A second subset with high ProteinGenerator confidence (model pLDDT > 90) and AF2 RMSD to design < 2Å, but low AF2 confidence (AF2 pLDDT < 80) was tested as well (Table S4). Synthetic genes encoding the designs were transformed into *E. coli*, and the proteins were expressed and purified using nickel-NTA chromatography. Of the 42 proteins tested, 34 were soluble and monomeric by size exclusion chromatography (SEC) and circular dichroism (CD) experiments showed they had the anticipated secondary structure and were stable up to 95°C (Figure 1D).

## 3.5 Conditioning on Single or Multi-state Structural Information

ProteinGenerator can be conditioned on either explicit 3D coordinates or on secondary structure as described by per-residue DSSP features. *In silico* tests show that when conditioned on 3D structural motif information the model generates proteins accurately recapitulating these motifs (Figure S9). During training coordinates for structural motifs are provided 40% of the time either as continuous spans of 4-9 residues or as 5-10 sparse residues, distant in sequence space. For lower resolution secondary structure guidance, DSSP[85] features were specified on a per-residue level 25% of the time and masked between 0% and 90% (randomly sampled). This allows guidance of a part or all of a structure towards a specific secondary structure type or fold, which the ProteinGenerator does quite well (Figure 1F, 3A).

Designing an amino acid sequence that can adopt distinct structural conformations upon an external trigger is a challenging task, as the energy landscape must contain two discrete minima with free energy differences small enough for a trigger to induce state switching[86]. We reasoned that ProteinGenerator was well equipped for this task because of its understanding of sequence-structure relationships and its ability to apply constraints in both domains.

We experimented with going beyond single-state structural specification by seeking to condition on distinct structural features of two different states. We applied multistate conditioning to design fold switching proteins by inputting the same sequence with two (or more) input sets of structural constraints at each step and averaging the output logits (Figure 1E). This allows the model to search sequence space for high-confidence solutions that satisfy all constraints. We used this approach to generate designs consisting of two fragments separated by a protease cleavage site which adopt different secondary structures following: in the intact parent state beta strand conditioning is used in the region flanking the cleavage site, while alpha helical conditioning is used for the two resulting subsequences. Logits from the parent and children sequences are averaged together at each step to arrive at a single sequence. As designed, AF2 structure predictions for the intact parent sequence have beta sheets in this region, whereas the two fragments (predicted independently) are entirely helical; the 3D structures of both the intact

parent and the two children are very close to the design models (Figure 1F). A similar multistate approach can be applied to design monomers that adopt multiple oligomeric states and other conformationally switching systems.

## 3.6 Sequence Guidance with Amino-Acid Based Potentials

An advantage of diffusion in sequence space is that sequence-based guiding functions can be readily implemented and applied. As a first test of this, we sought to design proteins with high frequencies of specific amino acids conferring structural or functional properties (cysteines can form disulfide bonds to make stable proteins, tryptophans possess spectroscopic properties, and histidines can confer pH sensitivity). Given a specification of the desired fraction of a given amino acid, at each denoising step positions are ranked based on the extent to which the output logits favor the amino acid, and the desired fraction are biased further in this direction (Figure 2A). We found this allowed more fine grained control in generating sequences than imposing a global bias towards a particular amino acid. We used this procedure to generate proteins with high frequencies (20%) of tryptophan, cysteine, valine, histidine, and methionine one at a time (Figure 2B, S10). We obtain compositionally biased protein sequences composed of nearly 20% of the desired amino acids that are strongly predicted to adopt the corresponding structures.

To evaluate the compositionally biased designs experimentally, we generated 70 to 80 residue proteins with different amino acids upweighted, filtered on AF2 pLDDT > 90 and AF2 RMSD to design < 2Å, and experimentally characterized the top 96 designs (Table S4). Of the characterized designs, SEC traces indicated the proteins were monomeric for 4/5 upweighted cysteine proteins, 8/19 upweighted tryptophan proteins, 19/22 upweighted valine proteins, 10/12 upweighted histidine proteins, and 10/10 upweighted methionine proteins. CD spectra were obtained for a subset of the monomeric designs, and in all cases indicated secondary structure was consistent with the designed structure (Figure 2D,E). Guiding for high cysteine content at the sequence level resulted in the formation of 3 to 5 disulfide bonds per

protein without any structural conditioning as indicated by mass spectrometry in the presence and absence of the reducing agent TCEP at 50mM (Figure 2D, Table S2). Proteins designed with upweighted tryptophans exhibited high absorbance at 280 nm, and proteins with upweighted valine exhibited higher beta sheet content by CD (Figure 2D middle, right). These results indicate the model understands general sequence to structure relationships beyond the typical sequence space of native proteins (Figure 2C).

We next explored the generation of proteins with prespecified charge composition, isoelectric points and hydrophobicity which can influence solubility, activity, subcellular location[87], pharmacokinetic clearance, and retention[88]. Biasing away from hydrophobic amino acids can lead to better expression and solubility[89], and designing towards hydrophobic interfaces is advantageous for protein-protein interactions[90]. We implemented sequence based potentials to guide[83] the diffusive process towards these characteristics to enable fine-tuned control over physical properties of the output sequence. This approach enabled the design of proteins with a range of user-defined hydrophobicities (Figure 2F) and isoelectric points (Figure 2G).

## 3.7 Scaffolding bioactive sequences

The design of proteins with activities conditional on an outside input is of considerable general interest, and could enable generation of therapeutics with spatial and temporal control[91]. As a first exploration of the use of ProteinGenerator for such proteins, we sought to scaffold bioactive peptide sequences within an inert protein cage. Unlike our previous LOCKR[92,93] sensor system, in which the bioactive sequence must be in a helical conformation and make specific interactions with the caging scaffold, the generality of ProteinGenerator requires only that the sequence of the bioactive peptide be specified–neither the structure this adopts nor the structure of the overall cage need be decided on in advance. ProteinGenerator is able to generate structures containing peptide sequences corresponding to known lytic peptides and the designed sequences are confidently predicted to adopt the designed structures (Figure 3.3B). We used this approach to scaffold a bioactive peptide in the terminus of a protein

that can be conditionally released upon proteolytic cleavage of a terminal loop (Figure 3.3C). We specified the sequence, not the structure, of the bioactive segment, and used DSSP conditioning to force the cleavage motif to be in a loop. We chose to scaffold the pore forming peptide melittin[94] currently being explored as a cancer therapy[95]. Starting with the melittin sequence and a flanking cleavage site, we generated an additional 125 residues to scaffold the peptide into a globular protein. Melittin-scaffolded proteins generated by the model were in agreement with AlphaFold2 models (AF2 pLDDT > 85, AF2 RMSD < 2Å) (Figure 3.3C). We obtained synthetic genes encoding 12 proteins scaffolding melittin and found that 9/12 were monodisperse by SEC and had the correct secondary structure by CD (Figure 3.3C).

## 3.8 Generation of Sequence Repeat Proteins

Repeat proteins containing tandem copies of a sequence-structure unit are ubiquitous in nature and play central roles in molecular recognition and signaling[96]. Previous work in designing repeat proteins has required extensive pre-specification of structural features[97]. We reasoned ProteinGenerator could be used to readily generate repeat proteins given only the sequence length of the repeat unit and number of repeats desired. At each timestep we symmetrize the noised sequence distribution accordingly (Figure 3.3D). Unconditional generation with this approach yielded largely beta solenoid structures which AF2 corroborated. We added helical caps to a subset of designs to promote stability and reduce aggregation[98,99] based on Zorine et. al 2023 (Unpublished). To encourage further exploration of the repeat protein universe we specified the secondary structure for a small percent (2%-10%) of residues, which yielded a wide range of all alpha, all beta, and mixed alpha-beta designs (Figure 3.3E). We generated 165-185 residue repeat proteins, filtered them (AF2 pLDDT>85 and RMSD to design < 2), and experimentally characterized 74 repeat proteins with helical caps and 86 repeat proteins without helical caps. Of these, 27 repeats with caps and 10 repeats without helical caps were soluble and monomeric by SEC, and 7/8 proteins evaluated using circular dichroism had the expected secondary structure (Figure 3.3E, Table S3[100]). We further obtained a crystal structure of one of the designs showing atomic accuracy of the AF2

design structure and the crystal asymmetric unit at 0.472 Å RMSD, and strong agreement with the overall structure at 1.385 Å RMSD.

## 3.9 Guidance with Sequence only Classifiers

Designing proteins with a desired biological activity is a long standing goal of de novo protein design. An advantage of our approach is that diffusion can be directly guided by function classifiers that operate in sequence space. We first sought to guide the network with the DeepGOPlus Gene Ontology (GO) classifier[101] to generate proteins with specific characteristics and functions. Although GO classification scores increased with guidance for nitrogen compound metabolic process (GO:0006807) and membrane (GO:0016020), we found the classifier had a high false positive rate often assigning high scores to native sequences outside the GO domain (Figure S101). In a separate approach, we trained a simple transformer encoder and single linear layer to discriminate unconditionally generated sequences from nanobody sequences and immunoglobulin (IG) folds aggregated from Integrated Nanobody Database for Immunoinformatics[102] and Structural Classification of Proteins database[103,104]. We generated 125 residue proteins, roughly the length of a nanobody, and found when classifier guidance or strand bias (1%) was used alone the classifier scores increased; when used in combination classifier scores increased (Figure 3.4A) along with the fraction of beta-strand containing proteins (Figure S121). 14% of designs made with the classifier alone were found to be beta sandwiches, which increased to 45% when applying a strand bias to 1% of the residues. Of the designs made with the combination of strand bias and classifier guidance 68.7% matched with tm-align > 0.5 to IG folds. AlphaFold models of sequences with high classifier scores matched the design models well (Figure 3.4B).

## 3.10 Guidance using Protein Family Sequence Information

Protein families often have specific residues important for function that are conserved throughout the family. Position-specific scoring matrices (PSSMs) capture this information and have been previously used to generate active enzymes with corresponding sequence composition using consensus sequence

design[105], but these approaches do not consider sequence-structure coherence, while Rosetta PSSM guided flexible backbone structure based sequence design[106] can require expensive MCMC calculations. We sought to use ProteinGenerator to design new members of protein sequence families, conditioning on both family multiple sequence alignments and key structural features associated with function. We generated a PSSM for the GFP fluorescent protein family[107] (all sequences in uniprot having greater than 30% sequence identity to GFP)[108]. At each step in denoising, we used the PSSM to bias the sequence distribution towards that of the family, and the calculations were conditioned on the coordinates of the residues contacting the chromophore (this cannot be done using purely sequence based methods) (Figure 3.4C). To tune guidance the PSSM was scaled by a factor of 0.25, 0.5, 0.75, and 1.0, and we found that sequences clustered closer together became more similar to GFP family members as scaling increased (Figure 3.4D, E, S13). AlphaFold2 and ESM-Fold are unable to predict native GFP from a single sequence (Figure S14) but predict sequences generated by the model with high accuracy to the design (Figure 3.4F). Active site coordinates provided as conditioning are in close agreement between the AF2 models and the native model, and demonstrate the model's ability to condition on both sequence and structural features (Figure 3.4F).

## 3.11 Towards Guidance with Experimental Data

A longstanding goal in protein design is the optimization of desired fitness attributes in as few experimental iterations as possible. ProteinGenerator presents a framework to employ this workflow: it allows for gradient based optimization on experimentally determined activity data from a pool of designs to bias generation for more active sequences. We simulate an iterative guidance process exploring the fitness landscape of the protein GB1 via partial diffusion on four mutation sites[109]. Sampling is guided by employing classifiers trained on the experimentally determined fitness of 96 designs generated in the preceding round (Figure 3.5A). For broad applicability of the discussed methodology it is essential that the optimization of desired fitness attributes can be reached with standard settings and no extensive

hyperparameter turning, such as the number of iterative rounds, batch size, or guidance strength. Consequently, here we discuss one exemplary approach using standard settings, where both the average fitness of the libraries of generated designs and the proportion of designs with a particular high fitness is increasing in each round of iterative guidance, although with the right hyperparameters both metrics could be extensively optimized (Figure 3.5B). In detail, employing iterative guidance on ProteinGenerator achieves an average fitness of 1.67 with libraries of the same size and 2 successive rounds. The average fitness of the libraries in the final round exceeds the fitness of a library with 96 designs sampled based on the highest confidences from a trained classifier. Furthermore, the average fitness as well as the maximum fitness of the library in the final round is higher than of a library generated using a classifier with artificially increased weights in a single diffusion generation round. Hu et al. (2023)[110] employed Bayesian Optimization for generation of GB1 libraries while optimizing the library size and round iteration number as hyperparameters. The designs generated using a library size of 96 samples in 4 successive rounds reach an average fitness of 0.6, whereby all fitness values above 1 were thresholded.

## 3.12 Discussion

By taking advantage of the ability of RoseTTAFold to jointly model protein sequences and structures, ProteinGenerator is able to directly sample in sequence space while ensuring that the 3D structure is coherent and satisfies any desired constraints. While RFdiffusion[75], Chroma[76], and other protein backbone diffusion models have demonstrated success in generating complex backbones with precise control over structural features, ProteinGenerator designs not only protein backbones but also sequences, enabling the design of proteins with any combination of desired sequence and structural attributes. We anticipate that our sequence space diffusion approach could also be employed with large language models that also generate structures, such as ESMfold[4]. A particularly attractive application is active learning-based protein design/engineering optimization: given a sequence-activity predictor, iterating between ProteinGenerator design of structurally coherent proteins predicted to have high activity,

experimental testing, and updating the activity predictor could be a powerful path to achieving high activity.

## 3.13 Methods

### 3.13.1 Sequence representation

To apply the diffusion framework in sequence space, a continuous representation of the categorical sequence data is needed. To implement this we represented the sequence, $\mathbf{x_0}$, with dimensions Lx20 where L corresponds to the protein length with 20 possibilities for each amino acid type. This takes the form of a one-hot encoded vector that is centered at zero by multiplying the Lx20 tensor by 2 and subtracting 1. Each logit within the tensor is a real number, with higher values corresponding to a higher

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

probability for that specific amino acid at that position. With this representation we noise $\mathbf{x_0}$ to obtain $\mathbf{x_t}$ with the below equation following Ho et al. formulation for a standard forward process sampling from Gaussian noise with mean at 0 and standard deviation of 1.

A critical part of the forward diffusion process is selecting the noising schedule. Determining the correct bin of a categorical distribution is trivial at low time steps by argmaxing the input sequence. Therefore more noise should be present at low timesteps to increase the difficulty of the task during training. The square root noise schedule[81] satisfies this requirement and was employed in this study.

### 3.13.2 Training

To train the model we began by sampling t uniformly from [0,T], where t=0 is an un-noised sequence and t=T is pure Gaussian noise. We then noise $\mathbf{x_0}$ to $\mathbf{x_t}$ with equation (1) and tasked the model to predict the un-noised sequence $\mathbf{x_0}$ and its corresponding structure $\mathbf{y}$. The timestep feature was added to the sequence template passed to the model. We applied a categorical cross entropy loss to $\mathbf{x_0}$ and structure

losses to **y** (FAPE, bond angle, bond length, distogram, lddt). An additional KL loss[81] was applied to the calculated $\mathbf{x_{t-1}}$. Self conditioning[79] was implemented to allow the model to condition on the previous $\mathbf{x_0}$ prediction and the back calculated $\mathbf{x_{t-1}}$ during both training and inference. To self condition in practice the model was used with gradients turned off to first predict $\mathbf{x_0}$ from $\mathbf{x_{t+1}}$, which was then passed in as a sequence template to the model. During training, RoseTTAFold was allowed 1 to 3 uniformly sampled "recycle" steps to refine structure predictions via multiple passes through the model[35]. Pseudo training and inference code is available in the supplementary information (Pseudocode S1, S2). In later training iterations secondary structure conditioning was provided to the model by concatenating a tensor representing DSSP features onto the sequence template. These features were provided 25% of the time and masked uniformly between 0% and 90% when provided.

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

### 3.13.3 INFERENCE

During inference starting from $\mathbf{x_t}$ the model predicts $\mathbf{x_0}$ and simultaneously decodes it to **y**. $\mathbf{x_0}$ is then back calculated to $\mathbf{x_{t-1}}$ with equation (1) and passed through the network with the previously predicted $\mathbf{x_0}$ to apply self conditioning. Benchmarking against conditioning on $\mathbf{x_t}$ as done in Ho et al[20] with the below equation, shows this approach performs better (SF 1C), as seen in other categorical diffusion methods[80,81].

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

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

This is done for T steps, but T can be varied, and does not have to be what was used during training. The model finds solutions to some problems in as little as 10 steps (Figure 1C). Furthermore, clamping the model's output logits from -3,3 gives better agreement with AF2 predictions (Fig S1B) $\mathbf{x_{t-1}}$ is sampled from either a zero-mean Normal distribution or a Non-Bayesian Gaussian Mixture distribution with equal mixing probabilities. For the Non-Bayesian Gaussian Mixture models we defined a mixture with two Normals centered at [-1, 1] (GMM2) and a mixture with three Normals centered at [-1, 0, 1] (GMM3).

### 3.13.4 DSSP Guidance

For constructing the DSSP features we calculated each training example's DSSP based on the structure with helix, strand, loop, and masked labels[85]. During training, the calculated per-residue secondary structure features were appended to RoseTTAFold's t1d features, and were one-hot encoded for 25% or 50% of the time and masked for 30% or 80% of the time. During inference, DSSP features are appended to the t1D features as necessary and masked when not.

### 3.13.5 Classifier Guidance

For classifier guidance we utilized the DeepGOPlus model[101] and trained a vanilla transformer model (2 Multihead Attention heads with each 2 layers, Embedding dimension: 64, Hidden Layer dimension: 64) on the INDI database for nanobodies[102]. Classifier guidance was implemented as described by Dhariwal and Nichol, 2021[83].

### 3.13.6 Multistate Guidance

Parent and child protein pairings were generated in 25 steps using DSSP features "XHHHHHHHHHHHHHLLLHHHHHHHHHHHHHLLEEEEEELLLEEEEEXXXXXXEEEEELLLEEEEEELLLHHHHHHHHHHHHHLLLHHHHHHHHHHHHHX" for parent, "XHHHHHHHHHHHHHLLLHHHHHHHHHHHHHLLLHHHHHHHHHHHHHHHHHX" for child A, and "XHHHHHHHHHHHHHLLLHHHHHHHHHHHHHLLLHHHHHHHHHHHHHX" for child B. A mixing coefficient of 0.25 was used to combine parent and child sequences together at each step. Multistate design pseudocode implementation is available in the supplements (Pseudocode S3).

### 3.13.7 Single Sequence Prediction with AlphaFold 2

All designs used in ESMFold benchmarks were modeled by using the curl command to predict single sequence structures. All designs used in AlphaFold2 (AF2) benchmarks and ordered for experimental characterization were predicted in single-sequence structure prediction mode with model 4. Pairwise backbone RMSDs between the design model and AF2 model were calculated for each design.

### 3.13.8 Sequence Identity calculations

Blast alignment was used to examine sequence alignment and similarity using query coverage >90% and target coverage >50%. Alignment to natives was done against Uniref-90.

### 3.13.9 Unconditional Protein Generation

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

confidence were ordered by filtering on design pLDDT > 90, AF2 pLDDT < 80, and AF2 RMSD to design < 5 Å.

### 3.13.10 Compositionally Biased Protein Generation

Proteins ranging from 70-80 amino acids in length with an amino acid compositional potential were generated in 25 steps. Designs were filtered by AF2 pLDDT > 90, AF2 RMSD to design < 2 Å, and SAP score[111] < 30. The top 10 to 22 designs were ordered for each upweighted amino acid type (tryptophan, cysteine, valine, histidine, and methionine). Pseudocode for the implementation of the amino acid compositional potential is provided in the supplements (Pseudocode S4).

### 3.13.11 Charge Biased Protein Generation

Proteins of 50 amino acids in length with charge potentials applied were generated in 25 steps with charge conditioning information. The ground truth charge for each protein was calculated at pH 7.4 by using the Henderson-Hasselbach equation. Pseudocode for the implementation of the charge potential is provided in the supplements (Pseudocode S5A, S5B).

### 3.13.12 Hydrophobic Biased Protein Generation

Proteins of 50 amino acids in length with hydrophobic potentials applied were generated in 25 steps with hydrophobicity conditioning information. The ground truth hydropathy index for each design was calculated by summing the hydropathy index for each residue and dividing by the sequence length[112]. Pseudocode for the implementation of the hydrophobic potential is provided in the supplements (Pseudocode S6A, S6B).

### 3.13.13 PSSM guidance

We formulate guidance with a PSSM by simply adding a precalculated PSSM to the output $\mathbf{x_0}$ prediction. This can be further scaled with a scaling factor, $\lambda$, to either promote stronger agreement with

the PSSM, or lower, to promote more diversity. PSSMs were calculated from MSAs generated by mmseqs2[113] with a 30-90% sequence identity cutoff to the query GFP sequence. Designs were filtered on AF2 pLDDT > 80 and AF2 RMSD to design < 2 Å.

### 3.13.14 Iterative guidance

Iterative guidance has been employed as described in Figure XA. In short, the experimental or in silico characterization of designs is collected to train a classifier, which is used to generate further designs with guidance. In an iterative process, classifiers trained on designs sampled in a preceding round inform network sampling. We investigated the nearly complete fitness landscape of the V39, D40, G41 and V54 amino acid sites of GB1, which is the binding domain of protein G, an immunoglobulin binding protein found in Streptococcal bacteria, as provided in the FLIP[114] paper.

### 3.13.15 GB1

We performed diffusion on the four mutation sites GB1 protein, while providing guidance in a second and third round design processes. In the first round designs are directly sampled from ProteinGenerator. For each round, a classifier is trained on the designs generated in the preceding round using a fitness equal to 1 as the classifier boundary.

### 3.13.16 Repeat protein generation

Repeat proteins ranging from 125-150 amino acids in length were generated in 50 steps with and without DSSP conditioning information. Designed proteins contained 5 repeat units using one of the one of the following DSSP strings, where X represents mask, E represents strand, and H represents helix:

"XXXXEEEEEXXXXXXXXXXXXXXXHHHHHXXXX",

"XXXXEEEEEXXXXXHHHHHXXXXXEEEEEXXXX",

"XXXXHHHHHXXXXXEEEEEXXXXXHHHHHXXXX",

"XXXXHHHHHXXLXXHHHHHXXLXXHHHHHXXXX",

"XXXXEEEEEXXHXXEEEEEXXHXXEEEEEXXXX",

"XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX",

"XXXXEXXXXXXXXXEXXXXXXXXXEXXXXXXXX".

Designs were filtered on AF2 pLDDT > 80 and AF2 RMSD to design < 2 Å.

### 3.13.17 Scaffolding Bioactive Peptides

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

### 3.13.18 Plasmid Construction

Protein designs were cloned into plasmids as in Watson et al[75]. Briefly, designs were ordered as synthetic genes (eBlocks, Integrated DNA Technologies) with BsaI overhangs compatible with a ccdB-encoding expression vector LM0627[16]. Genes cloned into LM0627 result in the following sequence: MSG-design-GSGSHHWGST**HHHHHH**, (SNAC cleavage tag and **6XHis affinity tag** are indicated). We used the NEBridge® Golden Gate Assembly Kit (New England Biolabs) with a total reaction volume of 5 µL and a ratio of 1:2 by mass of LM0627 plasmid DNA to design. We then incubated the reaction mixture at 37 C for 30 minutes, halted the reaction by incubating the reaction mixture at 60 C for 5 minutes, and transformed 1 µL of the reaction mixture into 6 µL of BL21 competent cells (New England Biolabs). After heat shock and recovery in SOC media, transformed BL21 cells were grown overnight in 1.0 mL of LB from which glycerol stock were created and small-scale expression cultures were inoculated.

### 3.13.19 1 mL-Scale Protein Purification

Initially, proteins were expressed with small-scale expression screens as previously reported[16] with small adaptations. Briefly, designs were inoculated with 100 uL of overnight growths and 900 uL of auto-induction media (sterile-filtered TBII media supplemented with 50 µg/mL kanamycin, 2 mM MgSO4, 1X 5052) in deep-well 96-well plates. 16 hours post-inoculation, cells were harvested and lysed in lysis buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM imidazole supplemented with 1X BugBuster, 1 mM PMSF, 0.1 mg/mL lysozyme, 0.1 mg/mL DNase). Clarified lysates were added to a 50 µL bed of Ni-NTA agarose resin in a 96-well fritted plate equilibrated with wash buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM Imidazole). After sample application and flow through, the resin was washed three times with wash buffer, and samples were eluted in 200 µL of elution buffer (50 mM Tris-HCl (pH 8), 0.3 M NaCl, 0.5 M imidazole, 5 mM EDTA (pH 8)). All eluates were sterile filtered with a 96-well 0.22µm filter plate (Agilent 203940-100) prior to size exclusion chromatography (SEC). Protein designs were then screened via SEC using an AKTA FPLC outfitted with an autosampler capable of running samples from a 96-well source plate. Samples were run on a SuperdexS75 Increase 5/150 GL column (Cytiva 29148722; 3,000 to 70,000 Da separation range) in a running buffer (20 mM Tris pH 8, 150 mM NaCl). To improve peak resolution, the SEC column was connected directly in line from the autosampler to the UV detector. 0.25 mL fractions were collected from each run. Absorption spectra were collected by the AKTA U9-M at 230 nm and 280 nm.

### 3.13.20 50 mL-scale protein purification

Proteins selected for further downstream characterization were expressed in 50 mL of auto-induction media. 16 hours post-inoculation, cells were harvested and lysed in lysis buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM imidazole, 1 mM PMSF, 0.1 mg/mL lysozyme, 0.1 mg/mL DNase) through sonication. Clarified lysates were added to a 2 mL bed of Ni-NTA agarose resin in a 20 mL column (Bio-Rad 7321010) equilibrated with wash buffer (50 mM Tris-HCl (pH 8), 0.5 M NaCl, 30 mM

Imidazole). After sample application and flow through, the resin was washed 3 times with 10 mL wash buffer, and samples were eluted in 2 mL elution buffer (50mM Tris-HCl (pH 8), 0.5M NaCl, 200mM Imidazole). All eluates were sterile filtered with a 3 mL 0.22uM filter plate prior to SEC. Protein designs were then screened via SEC using an AKTA FPLC outfitted with an autosampler capable of running samples from a 96-well source plate. Samples were run on a SuperdexS75 Increase 10/300 GL column (Cytiva 29148721; 3,000 to 70,000 Da separation range) in a running buffer (20 mM Tris pH 8, 150 mM NaCl). 1 mL fractions were collected from each run. Absorption spectra were collected by the AKTA U9-M at 230 nm and 280 nm.

### 3.13.21 Cysteine bias protein expression

Proteins guided towards high cysteine content were transformed into and expressed in Rosetta-gami B(DE3) Competent Cells (Novagen 71137). The 1 mL and 50 mL scale protein purification protocols were otherwise followed.

### 3.13.22 Circular Dichroism

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

### 3.13.23 Mass Spectrometry

To identify the molecular mass of each protein, intact mass spectra were obtained via reverse-phase LC/MS on an Agilent G6230B TOF on an AdvanceBio RP-Desalting column, and subsequently deconvoluted by way of Bioconfirm using a total entropy algorithm. Disulfide formation

was determined by injecting protein at 1.5 mg/mL in the presence and absence of 50 mM TCEP-HCl (Millipore Sigma 646547-10X1ML) and detecting the mass shift.

### 3.13.24 Acknowledgements

We would like to acknowledge Lisa Li, Sanaa Mansoor, Dmitri Zorine, Ian Humphreys, Harley Pyles, Brian Trippe, DéJenaé Ray, Abbas Idris, Xiaochuang Han, Meerit Said, Florence Dou, Linna Ann, Kejia Wu, Derrick Hicks, Hao Nguyen, Elias Kinfu, Adam Chazin-Gray, Quoc Tran, Marlo Zorman, Namrata Anand, and Naveen Jasti for helpful discussions and support. Chris Norn for PSSM scripts. Sergey Ovchinnikov for DSSP scripts. David Chmielewski for help with experimental procedures. Nate Ennist for help with CD. Doug Tischer for developing “contig” class for processing user inputs when running inference. Ivan Anishchenko for scripts to run TM - align, sequence similarity, and multidimensional scaling plots. Jue Wang and Justas Dauparas for benchmarking scripts. Minkyung Baek and Frank DiMaio for training scripts and RoseTTAFold code base. Joe Watson, David Juergens, and Nate Bennett for helpful scripts and conversations. Ian Haydon, Lance Stewart, Luki Goldschimdt, Adam Sadowski, Kandise Van Wormer, and Lauren Carter for general operations.

This work was supported by the Defense Threat Reduction Agency Grant HDTRA1-19-1-0003 (X.L.), by funding from the DARPA program Harnessing Enzymatic Activity for Lifesaving Remedies (HEALR) under award HR0011-21-2-0012 (X.L.), the Juvenile Diabetes Research Foundation International (JDRF) grant # 2-SRA-2018-605-Q-R (X.L.), AMGEN (S.L.), the Helmsley Charitable Trust Type 1 Diabetes (T1D) Program Grant # 2019PG-T1D026 (X.L.), the Bill and Melinda Gates Foundation Grant #OPP1156262 (X.L.), the Audacious Project at the Institute for Protein Design (J.S.), the Howard Hughes Medical Institute (J.S.).

### 3.13.25 Code Availability

The code for this project, with the exception of training scripts, is available at: https://github.com/RosettaCommons/protein_generator. For greater accessibility, thank you to Simon

Dürr and HuggingFace who supplied a GPU grant to run the model interactively in your browser:

https://huggingface.co/spaces/merle/PROTEIN_GENERATOR.

## 3.14 Main Figures

Figure 3.1: **Overview of ProteinGenerator.**

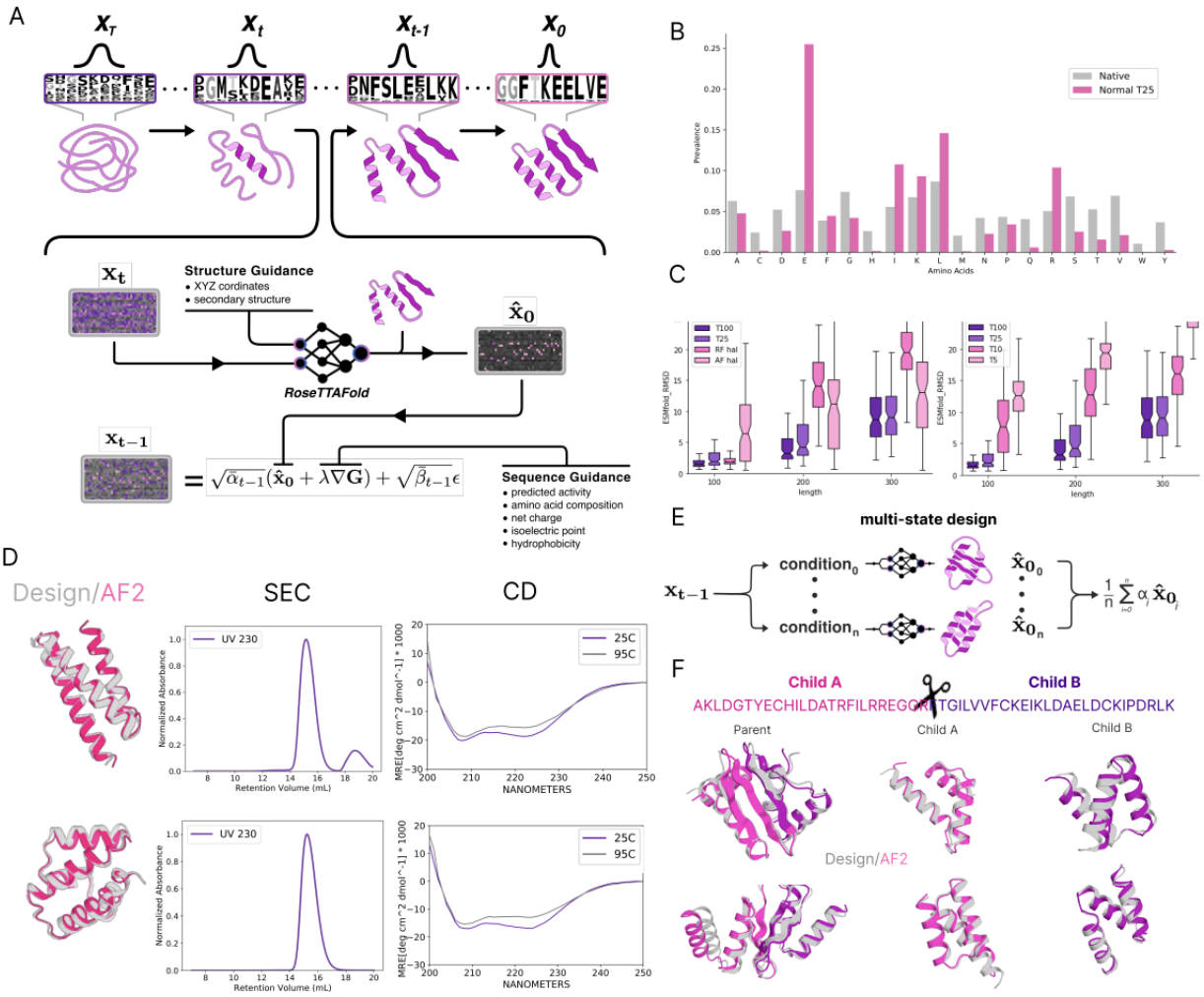



**Figure 3.1**. **Overview of ProteinGenerator.** (A) Inference schematic indicating how a noised sequence $\mathbf{x_t}$ is passed through the model with structural conditioning and updated to $\mathbf{x_{t-1}}$ for the next pass. At each step in the diffusion process $\mathbf{x_0}$ is predicted from $\mathbf{x_t}$ and guidance can be added to the predicted $\mathbf{x_0}$ prior to scaling with noise. This process is repeated for T steps as the sequence-structure pair converges on a high confidence solution. (B) Sampling from a Gaussian distribution yields sequences that approach that of native sequences. (C) Unconditional designs have higher ESMfold pLDDT and lower ESMfold RMSD to design compared other joint models: RosettaFold (RF) and Alphafold2 (AF) hallucination. (D)

Experimental validation of unconditional designs: Design (grey) and AlphaFold2 (pink) models of unconditionally generated proteins. Size exclusion chromatography and circular dichroism experiments show these designs are soluble, monodispersed, and thermostable to 95°C. (E) Multistate guidance allows for the design of a single sequence with a variety of structural conditioning states to converge on a single sequence predicted to adopt multiple states. (F) Use of multistate guidance to generate sequences which upon fragmentation switch from alpha/beta to all alpha secondary structure. Parent design (left) switches secondary structure when split into two child proteins (right). Designed structures of parent and children are shown in grey and AlphaFold2 predictions are overlaid in pink/purple.

Figure 3.2: **Generation of folded proteins with target sequence compositions.**

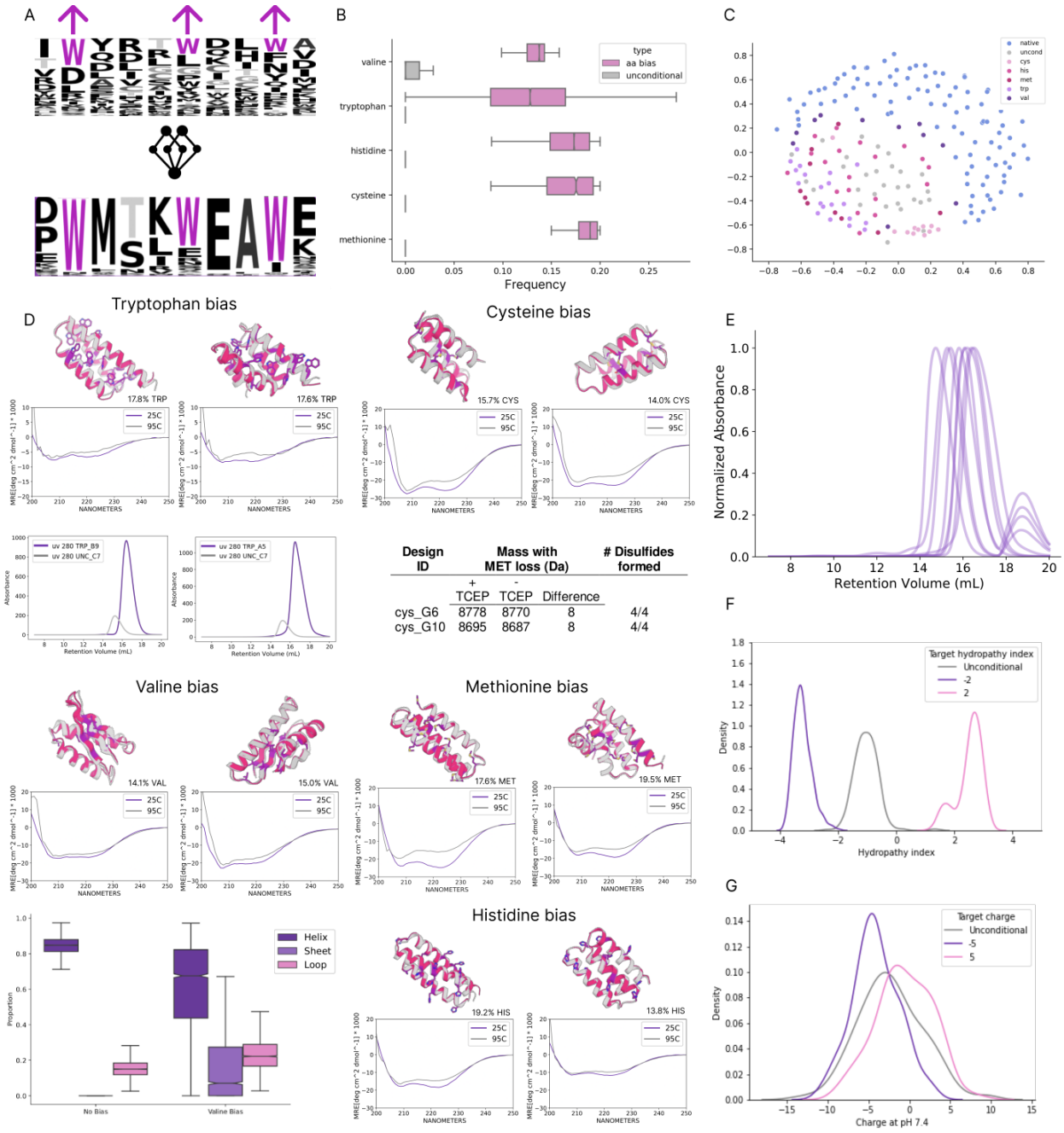



**Figure 3.2**. **Generation of folded proteins with target sequence compositions.** (A) Amino acid compositional bias: Applying an amino acid compositional bias generates proteins with desired sequence composition (tryptophan bias shown as an example). (B) Generation of compositionally biased proteins using ProteinGenerator. Bar plot displaying the average amino acid composition of designed proteins when guiding for 20% amino acid of interest compared to unconditional generation. (C) Sequence diversity representation: Multidimensional scaling of compositionally biased protein sequences occupy distinct spaces. (D) Experimentally validated amino acid compositionally biased proteins: Proteins generated with upweighted tryptophan, cysteine, valine, methionine, and histidine are predicted with high

confidence by AlphaFold2 (pink), match the design model (grey), and are thermostable up to 95°C by circular dichroism. Proteins designed with high tryptophan bias show five-fold higher absorbance at 280nm compared to unconditionally generated proteins. Experimental validation of cysteine-rich proteins under reducing and non-reducing conditions confirms the presence of designed disulfide bonds by mass spectrometry. Proteins designed with high valine bias show increased beta strand propensity compared to unconditionally generated designs. (E) Size exclusion chromatography overlay: Proteins designed with sequence potentials are soluble and monodisperse by size exclusion chromatography. (F) Hydrophobic sequence potential: Biasing the sequence away or toward hydrophobic amino acids results in a shift in the distribution of hydrophobicity scores for the output sequences. (G) Net charge sequence potential: Resulting distribution of net charges when guidance is used to bias sequences toward +/- 5 net charge.

Figure 3.3: **Structural and sequence conditioning can be used to generate proteins with desired secondary structure attributes and scaffold motifs.**

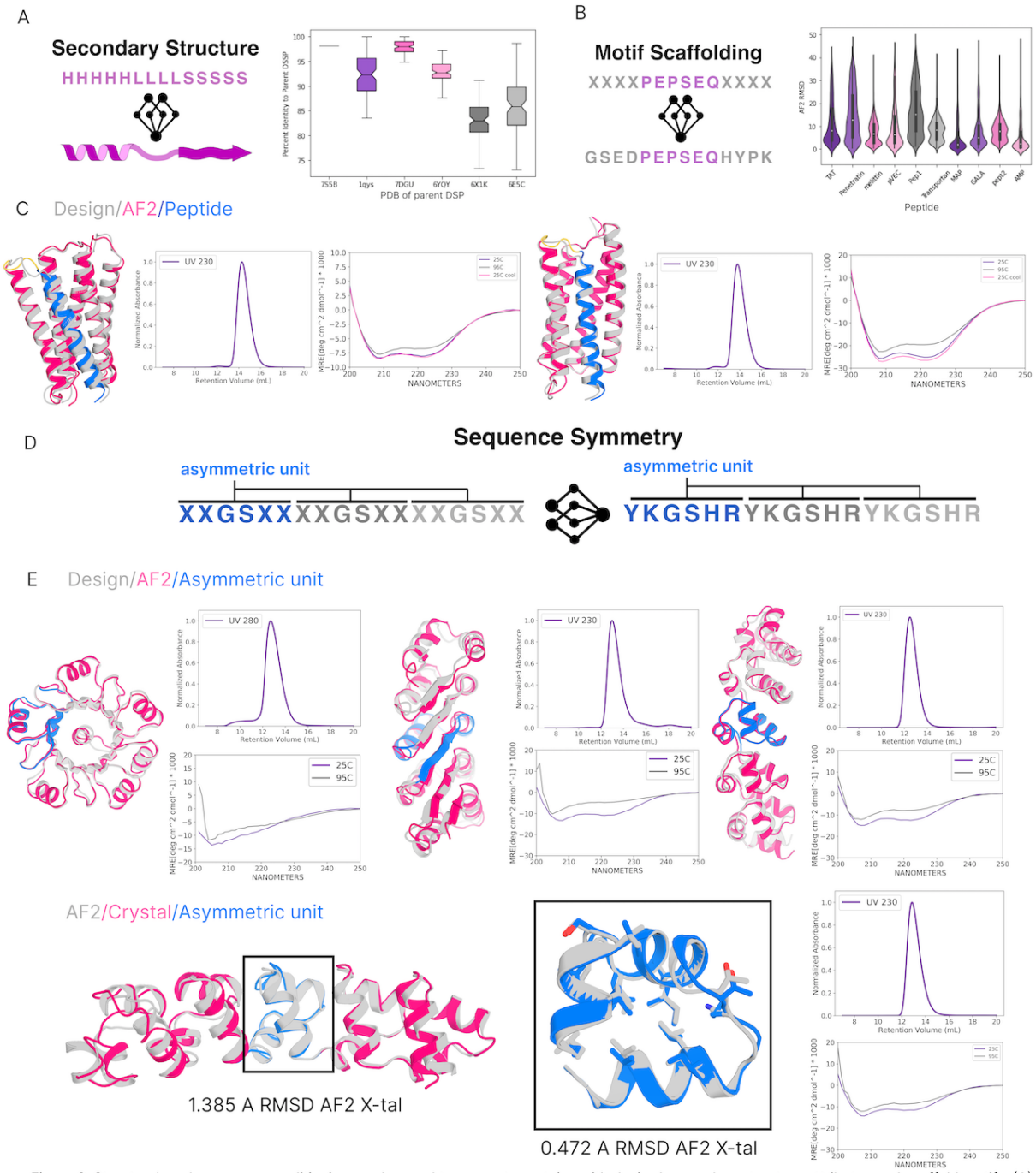



**Figure 3.3**. **Structural and sequence conditioning can be used to generate proteins with desired secondary structure attributes and scaffold motifs.** (A) Secondary structure conditioning: Protein

generation with secondary structure conditioning recapitulates the DSSP of the target protein. (B) Unstructured (sequence only) motif scaffolding: Sequence motif scaffolding of unstructured bioactive peptides yields designs with low AlphaFold2 RMSD to design. (C) Melittin scaffold designs: Scaffolding melittin yields designs (grey) that are corroborated by AlphaFold2 (pink), are soluble and monodispersed by size exclusion chromatography, and thermostable to 95°C by circular dichroism. (D) Sequence symmetry: Sequence repeat symmetry is applied at inference time by symmetrizing update (Xt-1) sequences to generate tandem repeat proteins. (E) Experimentally characterized symmetric repeat proteins: Designed repeat proteins (grey) with secondary structure conditioning are corroborated by AlphaFold2 (pink), are soluble and monodispersed by size exclusion chromatography, and are thermostable up to 95°C by circular dichroism.

Figure 3.4: **Fold and function guided protein generation.**

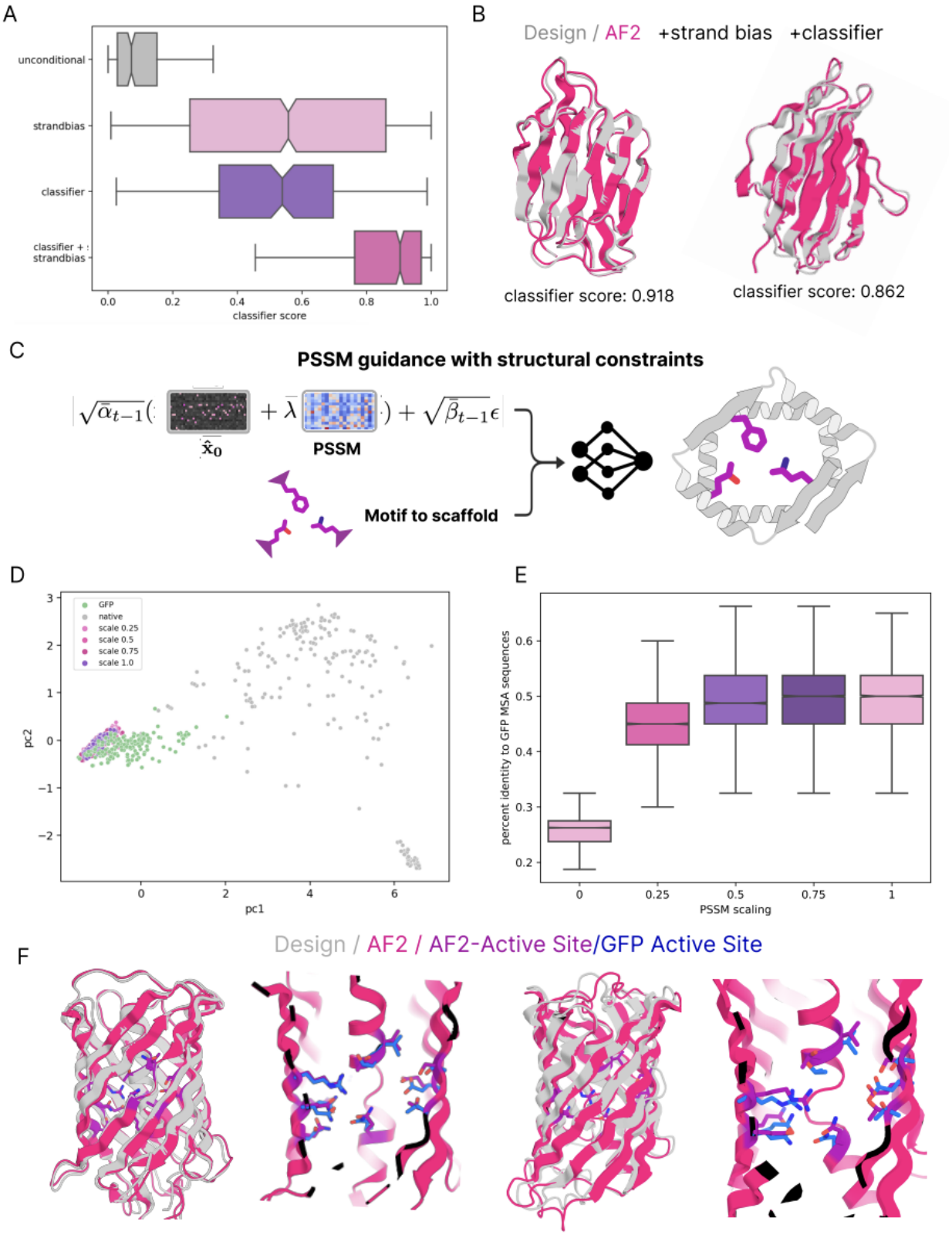
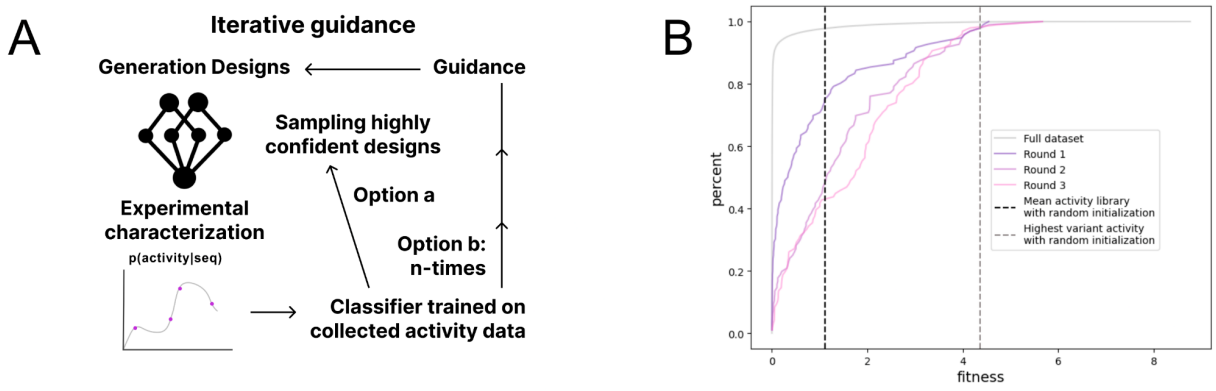


**Figure 3.4**. **Fold and function guided protein generation.** (A) immunoglobulin (IG) sequence based classifier score distributions from outputs generated unconditionally, with strand DSSP conditioning, with gradients from IG classifier, and with combination of IG classifier gradients and secondary structure conditioning. (B) Design model (grey) and AF2 model (pink) of proteins generated with classifier guidance and strand DSSP conditioning. (C) Schematic of PSSM guidance with scaling and motif scaffolding. (D) Guidance by position specific score matrix (PSSM) generates sequence-structure pairs sampled from a similar embedding space as native GFP designs. Designs and natives were embedded with ESM and first and second principle components derived from embeddings are plotted. (E) Sequence identity to native GFPs increase with increasing guidance scaling: Sequence similarity of PSSM designs at varying guide scales to native GFPs. (F) PSSM guided design examples: Design models (grey) and AF2 models (pink) of proteins generated with a PSSM guidance scale of 0 (left) or 8 (right) and 30% DSSP masking. Active site residues on AF2 predicted structures (purple) over-layed with wildtype GFP side chains in blue.

Figure 3.5: **Active learning guided protein generation.**




C

| | Average fitness of library | Highest fitness of library |
|---|---|---|
| Round 1 | 0.11 | 2.77 |
| Round 2 | 1.11 | 4.37 |
| Round 3 | 0.62 | 3.95 |

**Figure 3.5**. **Active learning guided protein generation.** (A) Active learning schematic indicating an iterative process of sampling designs from ProteinGenerator. After sampling designs from the network (1), designs can be subsequently characterized experimentally or using in silico metrics (2). With a classifier trained on the collected activity data (3) the generation of future designs can be guided (4). Thus, designs sampled from the network in (1) are closer to the desired true sequence-function distribution. (B) The GB1 library fitness and portion of variants with a high fitness is increasing for each round (All dataset: 0.03, Round 1: 0.87, Round 2: 1.51, Round 3: 1.67). Guidance through a classifier trained on the designs sampled in the preceding round was applied for sampling designs in round 2 and round 3. (C) Iterative guidance based on randomly chosen designs in round 1 generates libraries with a decreased average and highest fitness in comparison to designs in round 1 sampled unconditionally by

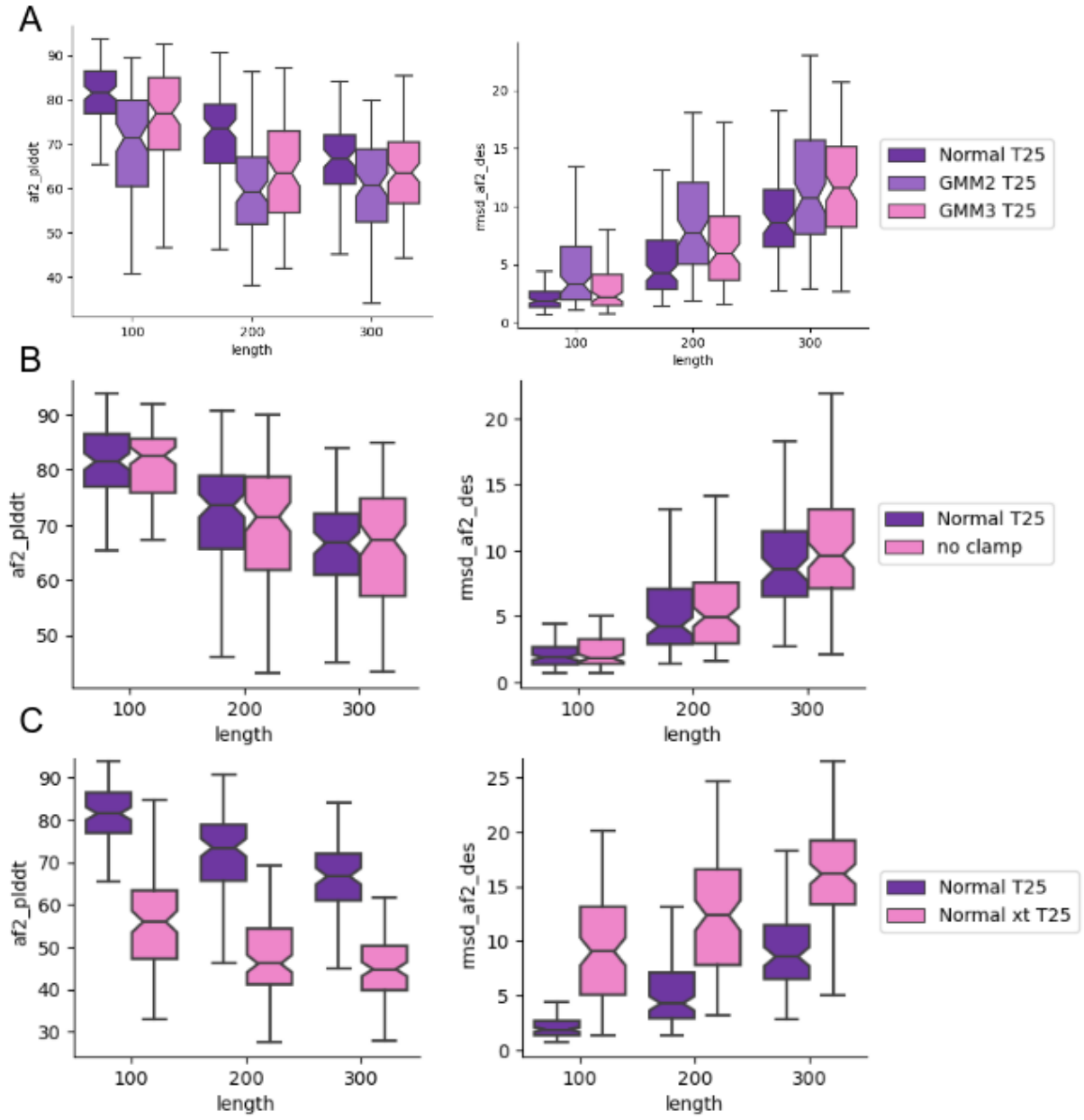
ProteinGenerator. The decrease of the average fitness of the library between round 2 and round 3 indicates that ProteinGenerator guidance converged to a local activity maxima, while not being able to reach regions with higher activity and high confidence by ProteinGenerator.

## 3.15 Supplementary Figures

### Supplementary Figure 1: Inference benchmarks.

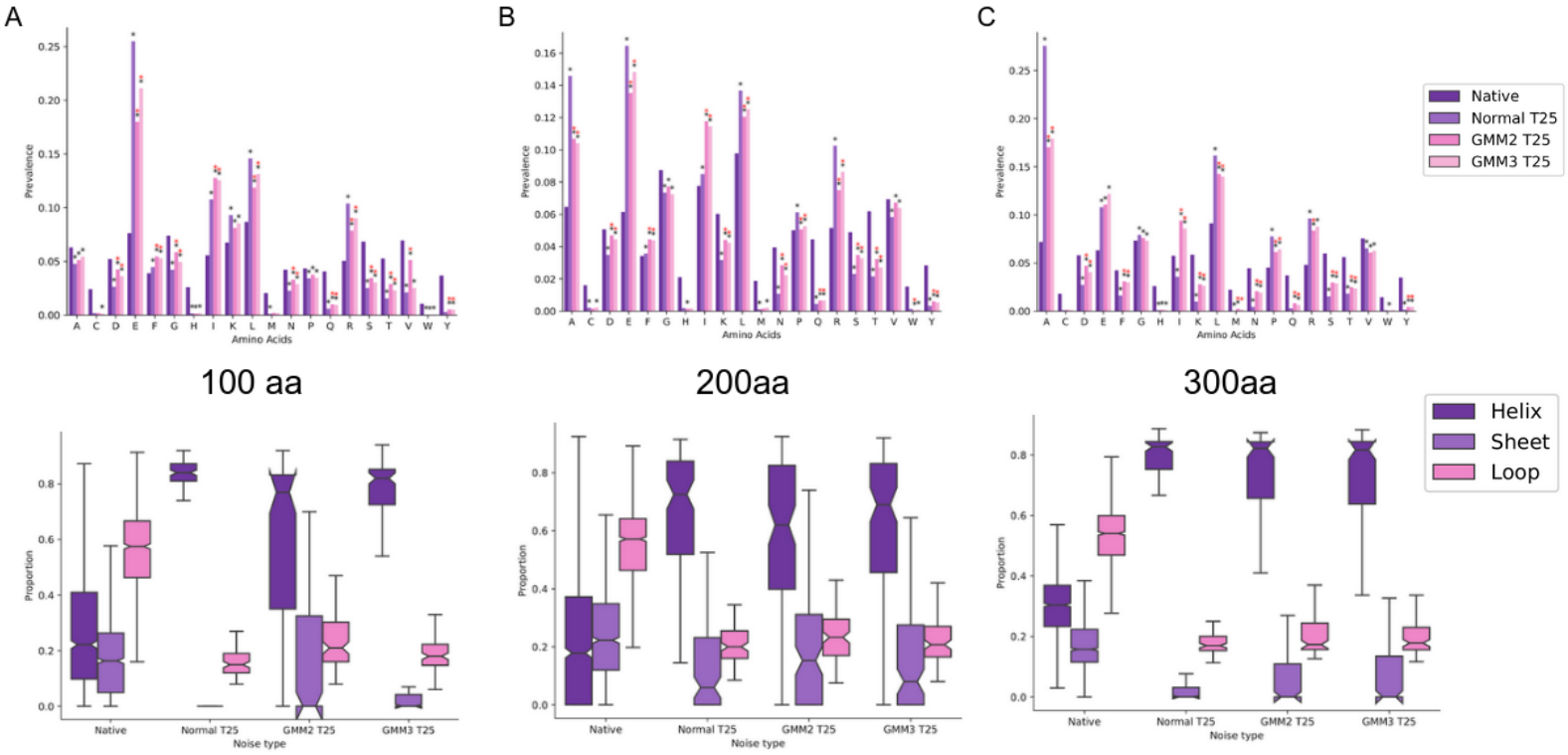
(A) Boxplot of AF2 pLDDT of sequences from model clustered by length on left, right RMSD of AF2 model to design. (B) Boxplot AF2 pLDDT with clamp (-3,3) applied post sampling $\mathbf{x_{t-1}}$ and no clamp on left, AF2 RMSD to design. (C) Comparison of with and without conditioning on $\mathbf{x_t}$ when sampling $\mathbf{x_{t-1}}$,

AF2 pLDDT right, AF2 RMSD to design left.

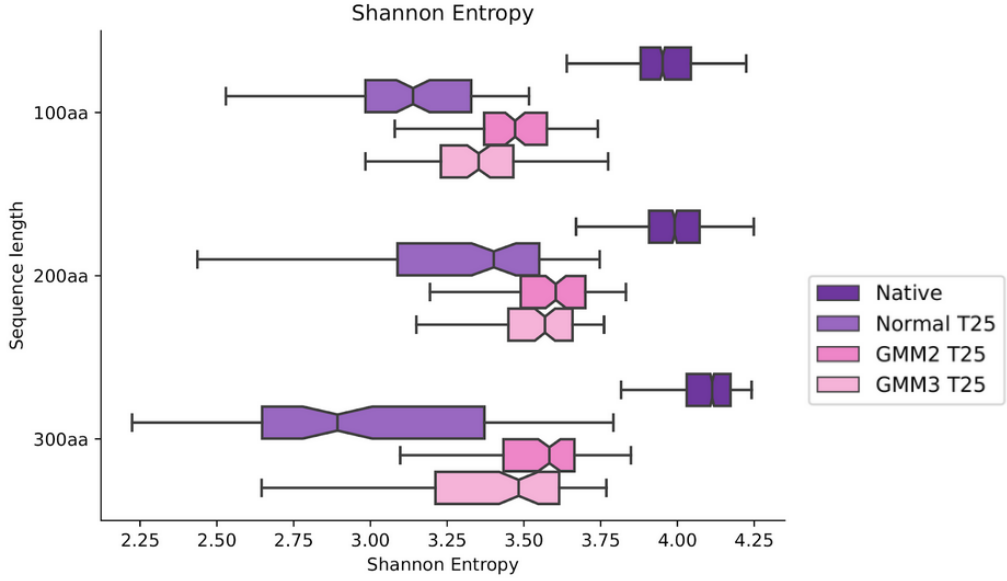


Supplementary Figure 2: Amino acid distributions and secondary structure propensities

(A) 100AA, (B) 200AA, and (C) 300AA length proteins when sampling from normally distributed noise, GMM2, or GMM3. Significant amino prevalence changes between native and unconditional designs as well as between unconditional designs sampled from normal noise and sampled from GMM2 or GMM3 noise are displayed via black asterisk and a red asterisk.

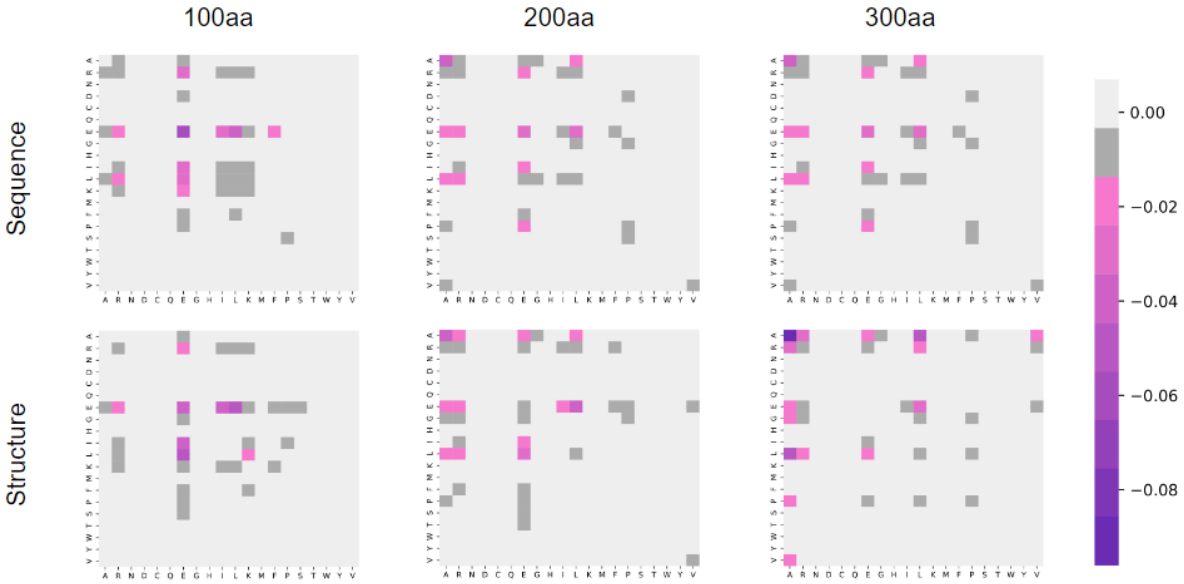



Supplementary Figure 3: Sequence entropy of native proteins, normally sampled sequences, GMM2, and GMM3 for 100AA, 200AA, and 300AA proteins.



## Supplementary Figure 4: Sampling from different noise distributions generates proteins with different sequence and structural neighbors.

Frequencies of amino acid neighbors in generated sequences (top row) and nearest structure neighbors (bottom row). Values are calculated as the difference between native sequences and unconditionally generated sequences for 100AA (left), 200AA (middle) and 300AA (right). Unconditional designs of 200AA and 300AA are characterized by more frequent alanine-leucine, alanine-glutamic acid, and alanine-alanine sequence and structure contacts. In structure space unconditional proteins exhibit lower frequencies of glutamic acid-glutamic acid and glutamic acid-proline neighbors.



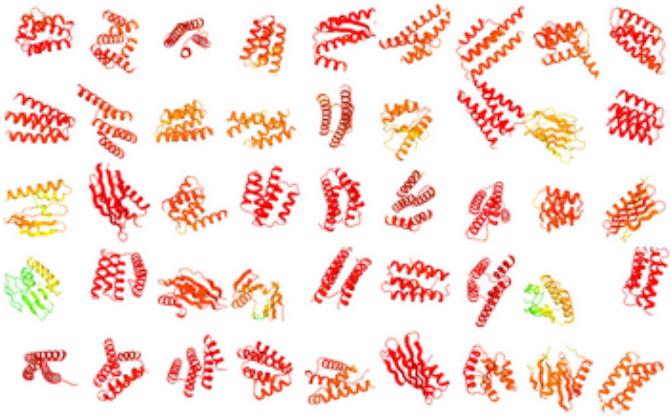

## Supplementary Figure 5: Sampling from different noise distributions generates proteins with more diverse secondary structure.

Representative 100AA unfiltered and unconditionally generated proteins from normal distribution, GMM2, and GMM3. Colored by model pLDDT (red → high confidence).

normal

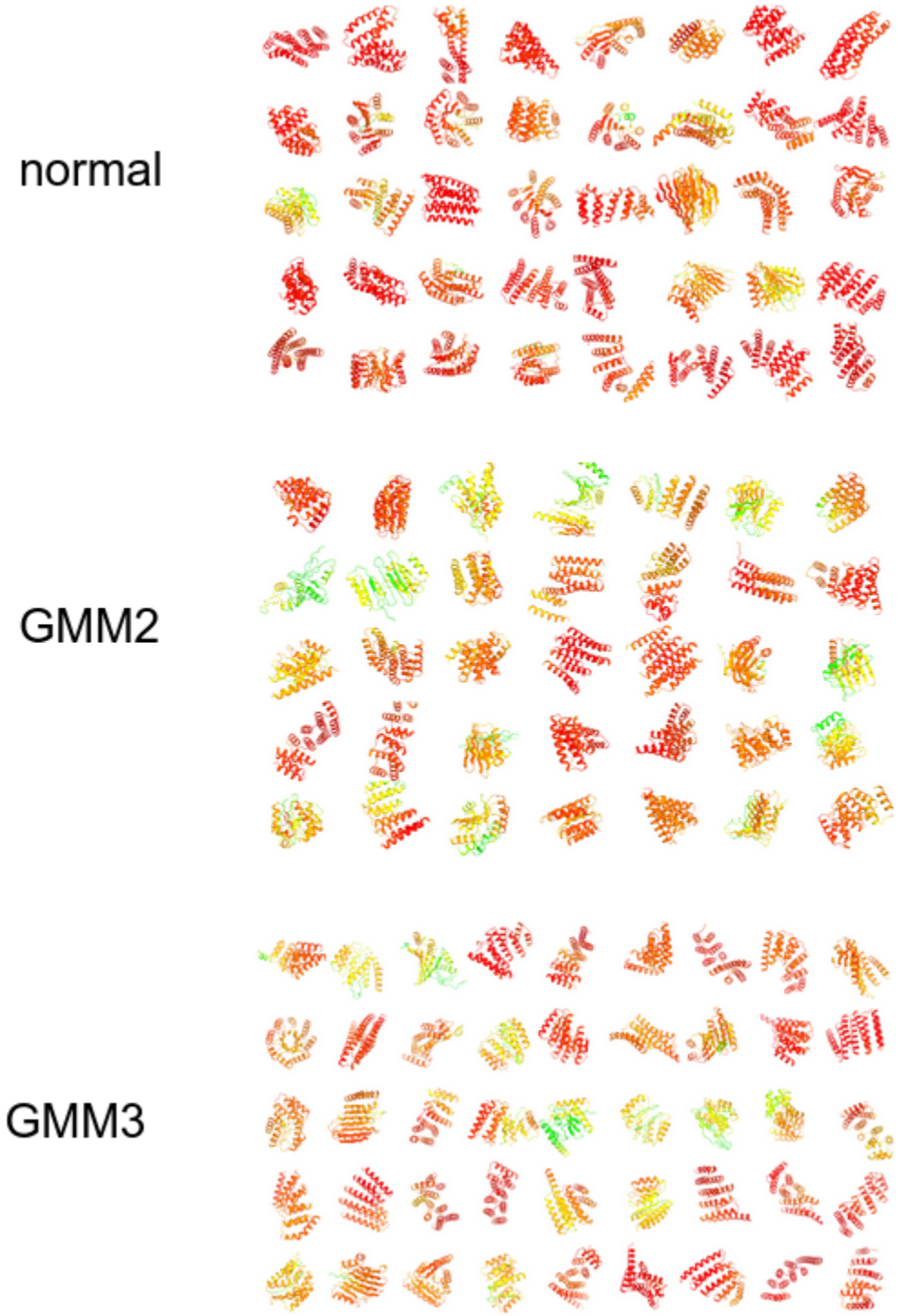
GMM2

GMM3



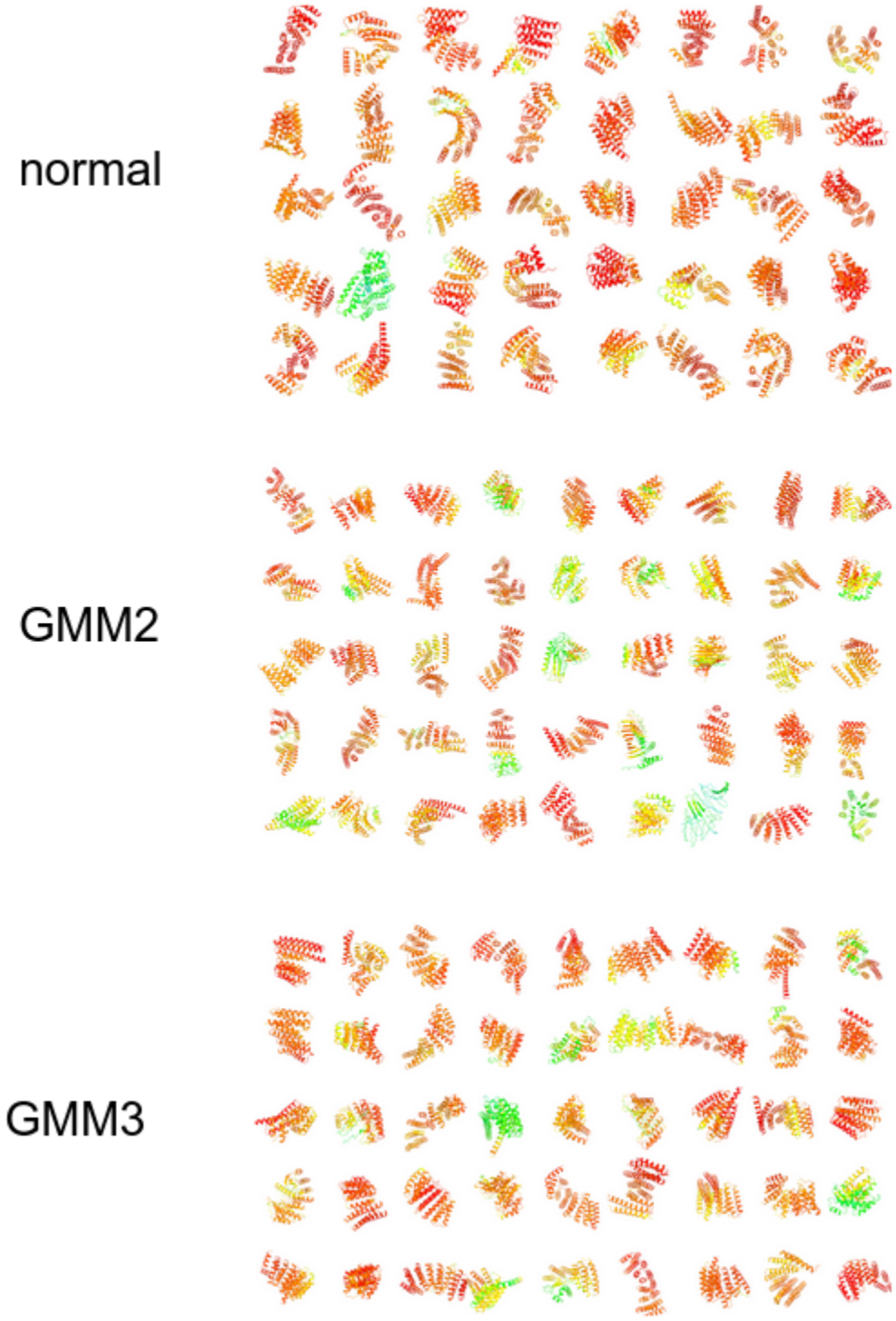
Supplementary Figure 6: Sampling from different noise distributions generates proteins with more diverse secondary structure.

Representative 200AA unfiltered and unconditionally generated proteins from normal distribution, GMM2, and GMM3. Colored by model pLDDT (red → high confidence).



Supplementary Figure 7: Sampling from different noise distributions generates proteins with more diverse secondary structure.

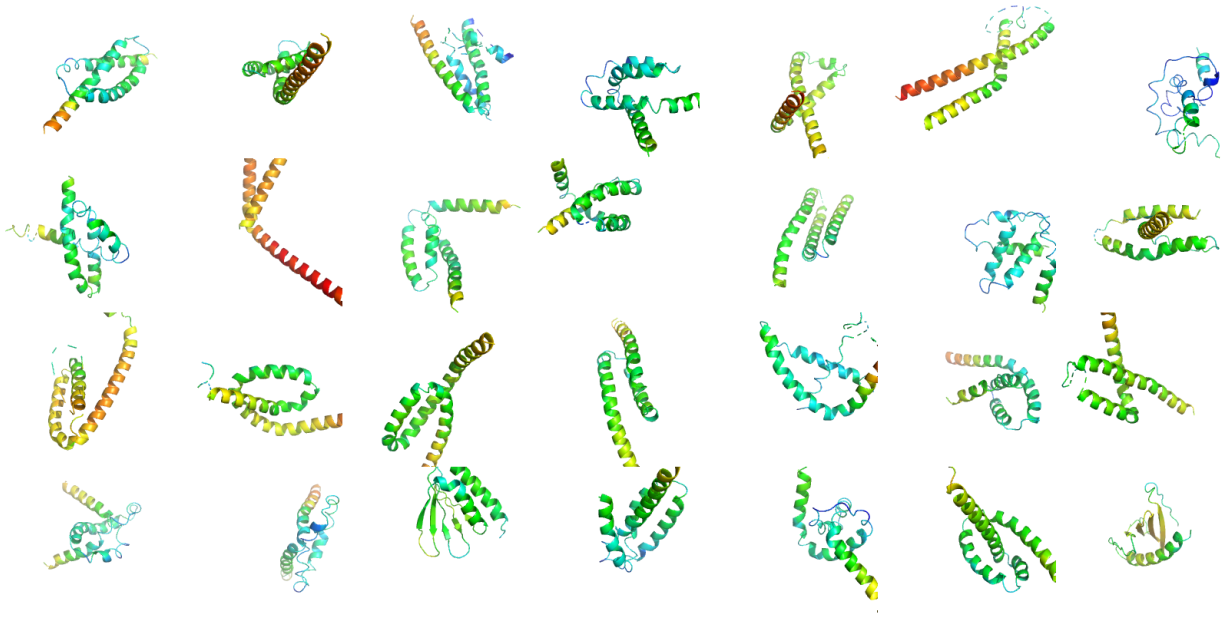
Representative 300AA unfiltered and unconditionally generated proteins from normal distribution, GMM2, and GMM3. Colored by model pLDDT (red → high confidence).



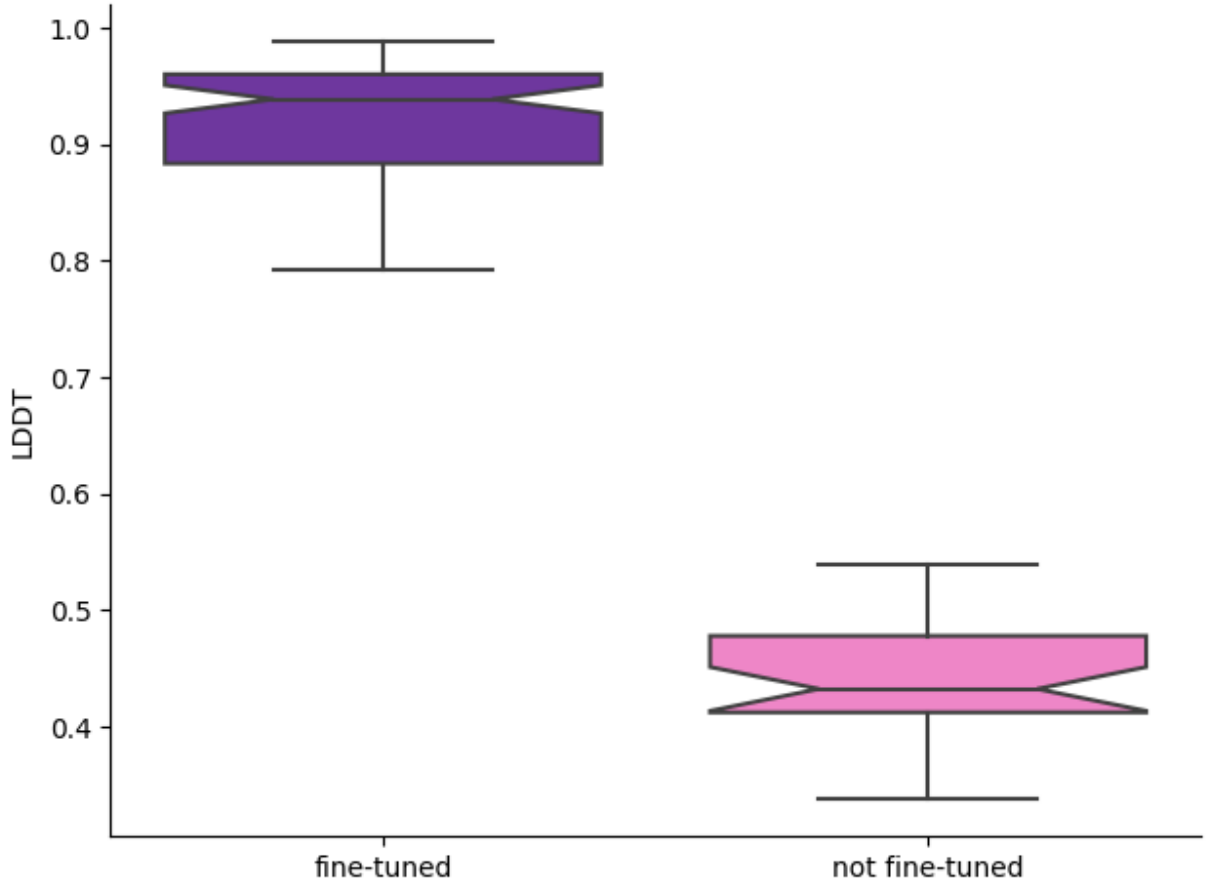
Supplementary Figure 8: Fine-tuning RoseTTAFold is a necessary prerequisite for the generation of high confidence proteins.

(A) Representative 100AA unfiltered and unconditionally generated proteins from a not fine-tuned RoseTTaFold model. Colored by model pLDDT (red → high confidence). (B) Average RoseTTAFold LDDT of proteins unconditionally generated with a fine-tuned and a not fine-tuned RoseTTAFold model.

(A)

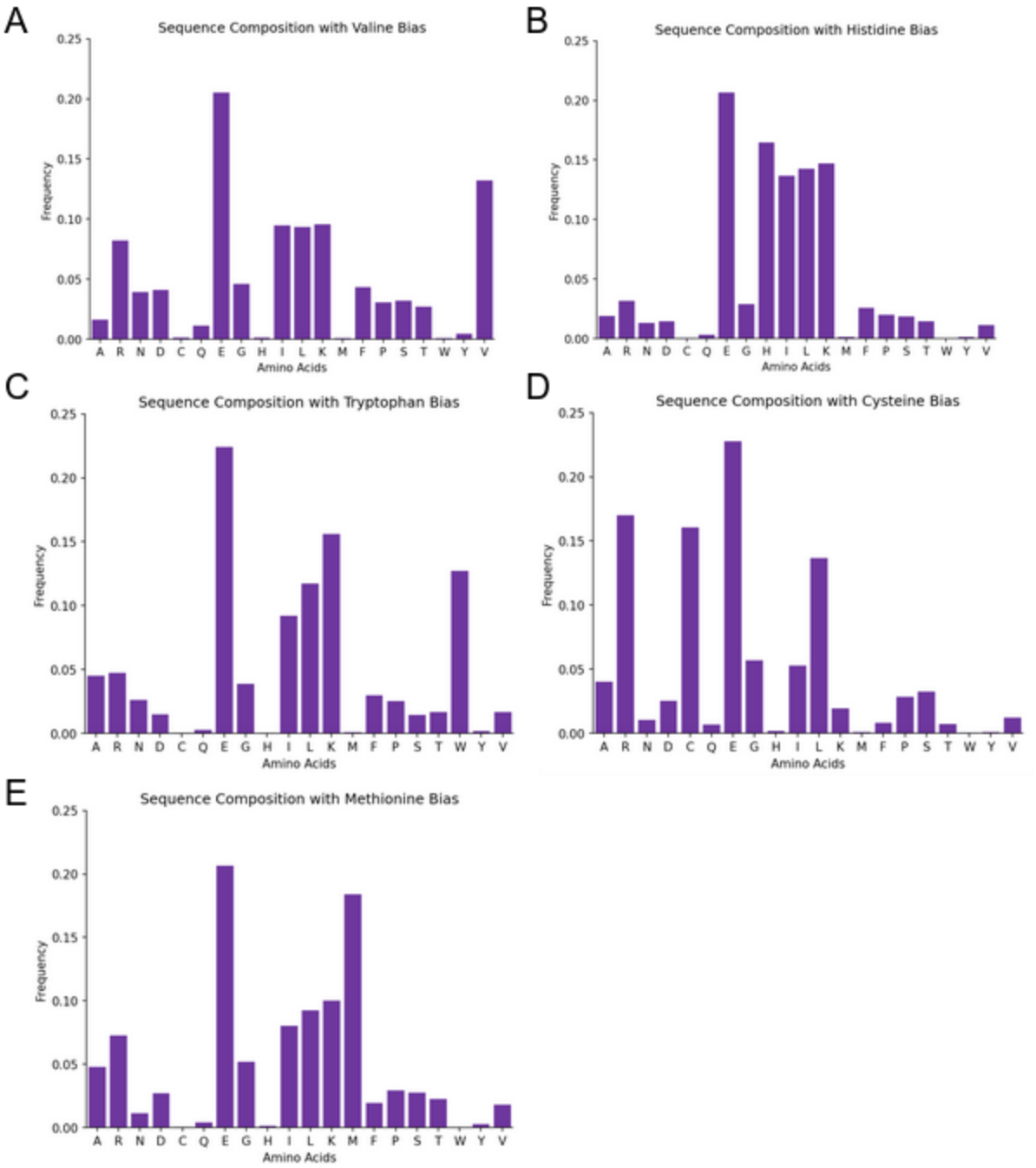


(B)




Supplementary Figure 9: Biasing for specific amino acids results in increased frequency of the specified residue in generated proteins.

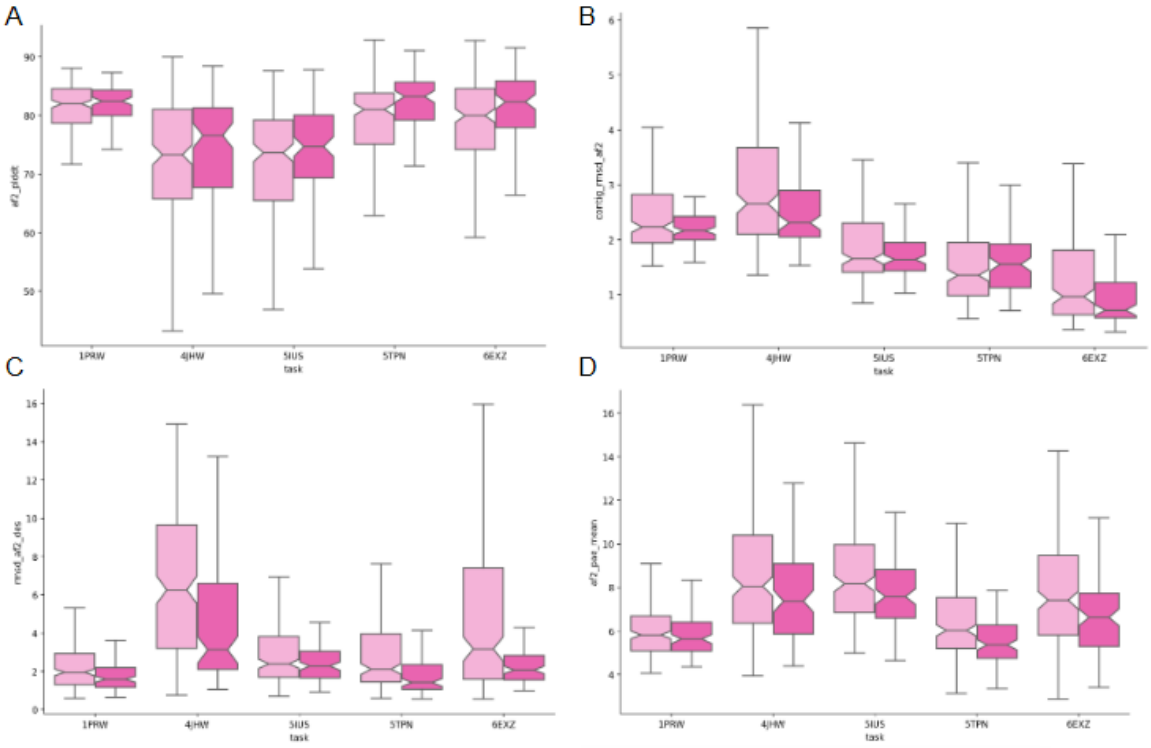
Amino acid distributions of proteins generated with amino acid compositional bias for (A) valine, (B) histidine, (C) tryptophan, (D) cysteine, and (E) methionine.




Supplementary Figure 10: AF2 metrics for scaffolding of structure-sequence motifs in the PDB IDs listed in 25 (light pink) and 100 (dark pink) time steps.

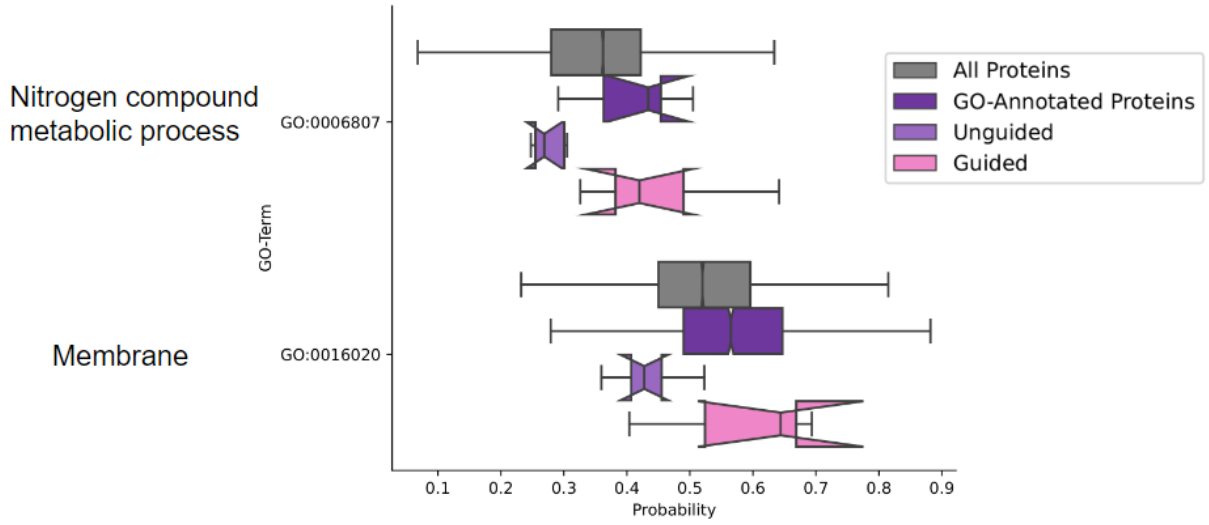
(A) AF2 pLDDT for designs, (B) RMSD of motif predicted by AF2 to design, (C) RMSD of AF2 to design for whole structure, (D) predicted alignmed error (pAE) of designs from AF2. The following contig arguments were used to run motif scaffolding benchmark:
1PRW - contigs 8-20,A21-31,16-25,A56-67,8-20,

6EXZ - contigs 0-95,A28-42,0-95,
5TPN - contigs 10-40,A163-181,10-40,
5IUS - contigs 0-30,A119-140,15-40,A63-82,0-30,
4JHW - contigs 0-25,F196-212,15-30,F63-69,10-25

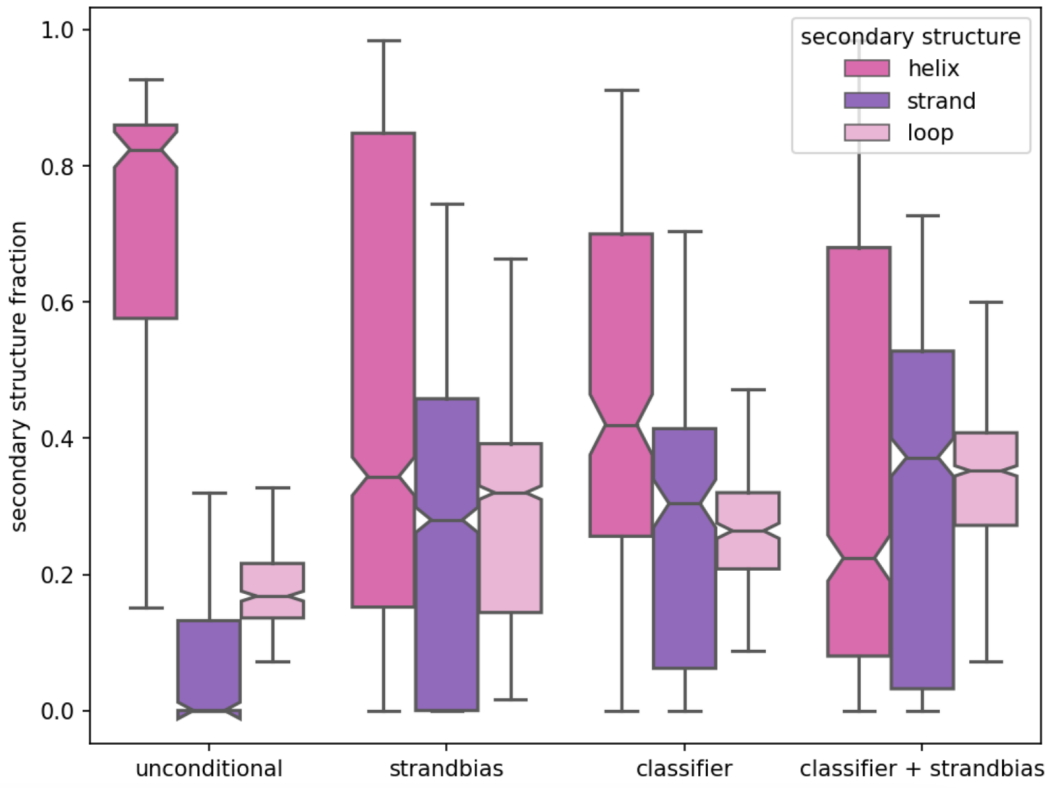



## Supplementary Figure 11: GO-guidance.

The network has been guided with the DeepGOPlus Gene Ontology (GO) classifier to generate proteins with specific characteristics and functions. Exemplary, the classifier GO probability scores for all UniProt proteins, all proteins annotated with the chosen GO term, unconditionally unguided proteins generated with our model and guided proteins generated with our models for the GO terms nitrogen compound metabolic process (GO:0006807) and membrane (GO:0016020) are shown. The classifier has a high false positive rate due to a high mean probability as well as for all UniProt proteins including proteins not annotated with this specific GO term. For both GO terms a shift in the probabilities can be shown for guided proteins in comparison to unguided proteins.

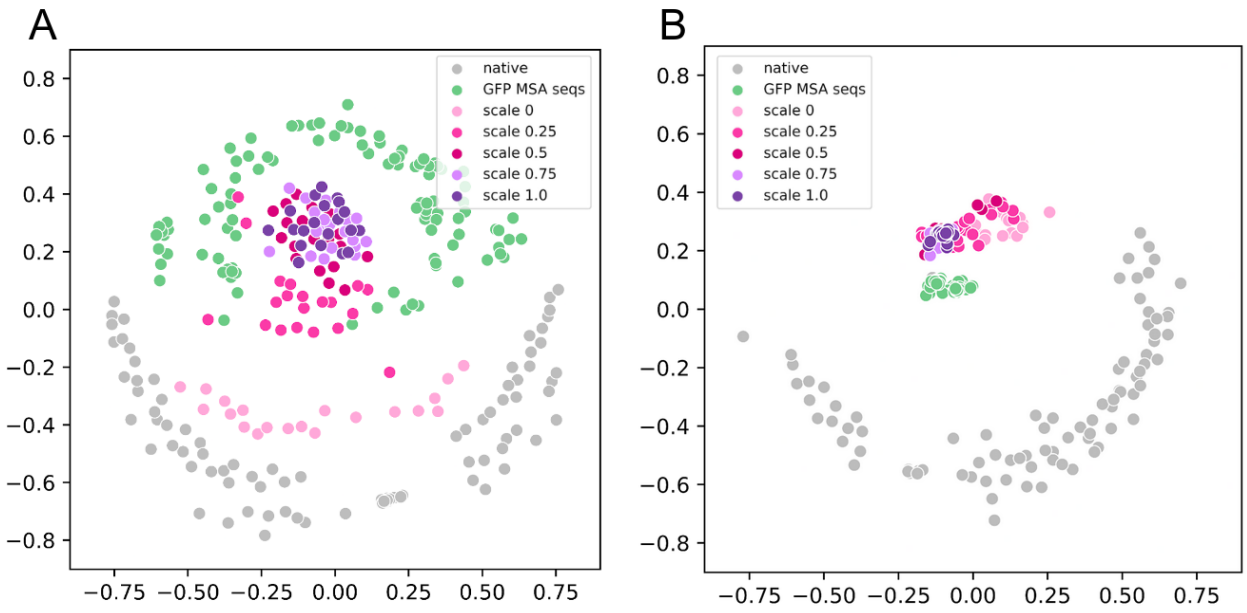



Supplementary Figure 12: Secondary structure composition comparison when generation unconditional designs, designs with strand bias, designs with classifier guidance, and combination of classifier guidance and strand bias.



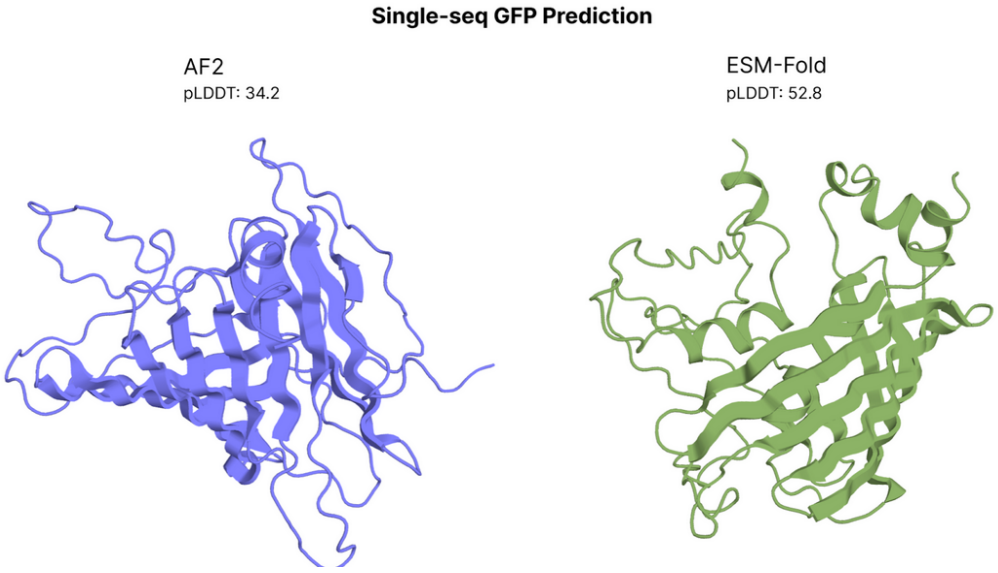

Supplementary Figure 13: Multidimensional scaling plots of proteins generated with increasing GFP PSSM guidance scales.

(A) Higher PSSM scaling increases sequence clustering to native GFPs. Distance metric is percent sequence identity. Green dots are native GFP sequences derived from a GFP MSA with sequence identity cutoffs 30-90% to the query sequence. Grey are randomly sampled native sequences from Uniprot90 (B) Low PSSM scaling results in increased structural diversity and samples of more diverse beta barrels. Higher PSSM scaling reduces structural diversity and clusters closer to native GFPs. Distance metric is TM score. Green dots are structures derived from the same MSA as (A) and grey dots are structures derived from the same set as (A).

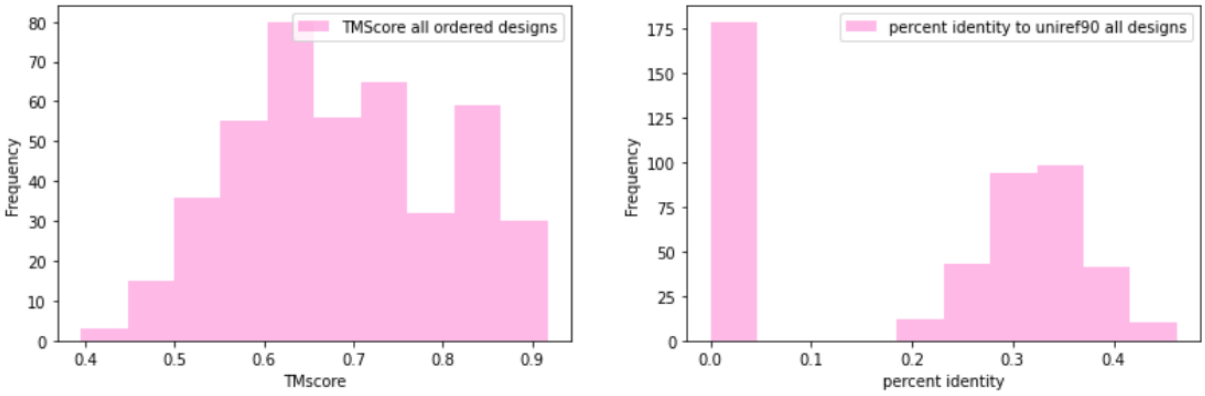



## Supplementary Figure 14: Single-sequence structure predictions of Green Fluorescent Protein (GFP), (PDB 1EMA).

AlphaFold2 (left) and ESM-Fold (right) predictions fail to recover the tertiary structure of GFP when run in single-sequence mode. Both models return structures with low confidence (pLDDT). Models were run with 6 recycles.




Supplementary Figure 15: TMScore (left) and sequence identity (right) distributions of all ordered designs against PDB.



Supplementary Pseudocode 1: Training.

```
def train(seq,xyz,steps,cond,mask_struc,mask_seq):

  #one hot encode
  x=one_hot(seq)*2-1
  t=uniform(0,steps)/steps

  #noise
  eps = normal(mean=0, std=1)
  x_t = sqrt(gamma(t)) * x + sqrt(1 - gamma(t)) * eps

  #masking
  xyz_t=full_like(xyz,Nan)
  xyz_t[~mask_struc]=xyz[~mask_struc]
  x_t[~mask_seq] = x[~mask_seq]

  #self condition
  with stop_gradient:
    x_prev = None
    x_t_1 = sqrt(gamma(t+1)) * x + sqrt(1 - gamma(t+1)) * eps
    x_prev = model(x_t_1,x_prev,t,cond)

  #predict and calc loss
  x_pred,struc_pred = model(x_t,x_prev,t,cond,xyz_t)
  loss  = CCE(x_pred,one_hot(seq))
  loss += KL()
  loss += Structure_Losses(struc_pred,xyz)

  return loss.mean()
```

Supplementary Pseudocode 2: Inference.

```
def generate(L,steps,cond,seq,xyz, mask_struc,mask_seq, classifier):

  #setup
  x_prev=None
  seq_start=full(1,L,20)
  x_pred=one_hot(seq_start)*2-1
  seq=one_hot(seq)

  for step in range(steps):
    #noise
    t=(steps-step)/steps
    eps = normal(mean=0, std=1)
    x_t = sqrt(gamma(t)) * x_pred + sqrt(1 - gamma(t)) * eps

    #masking
    xyz_t=full_like(xyz,Nan)
    xyz_t[~mask_struc]=xyz[~mask_struc]
    x_t[~mask_seq] = seq[~mask_seq]

    #predict x_0
    x_pred=model(x_t,x_prev,t,cond,xyz_t)

    #classfifier guidance
    uf classifier is not None:
      x_pred += grad(classifier(x_pred))

  return argmax(x_pred)
```

Supplementary Pseudocode 3: Multistate design.

```
def generate(L, steps, conditionings, seq, xyz, mask_struc, mask_seq, classifier):

  #setup
  x_prev=None
  seq_start=full(1,L,20)
  x_pred=one_hot(seq_start)*2-1
  seq=one_hot(seq)

  for step in range(steps):
    #list of output seqs
    pred_Xo = list()
    for cond in conditionings:
      #noise
      t=(steps-step)/steps
      eps = normal(mean=0, std=1)
      x_t = sqrt(gamma(t)) * x_pred + sqrt(1 - gamma(t)) * eps

      #masking
      xyz_t=full_like(xyz,Nan)
      xyz_t[~mask_struc]=xyz[~mask_struc]
      x_t[~mask_seq] = seq[~mask_seq]

      #predict x_0
      x_pred=model(x_t,x_prev,t,cond,xyz_t)
      pred_Xo.append(x_pred)

    #updated x_pred with average output seqs
    x_pred = mean(pred_Xo)

  return argmax(x_pred)
```

Supplementary Pseudocode 4: Amino acid composition potential.

```
def aa_bias(seq):

  #softmaxed probability distribution for seq [L,21]
  soft_seq = softmax(seq)

  #stack gradients in list for each AA
  grad_stack = []

  #iterate through each aa and fraction to bias sequence toward:
  for aa, fraction_to_bias in aa_bias_list:

  #set up aa bias by initializing seq of zeros
  potential = zeros_like(seq)

  #set residue type of interest to 1
    potential[:,aa] = 1

    #get mean squared error between soft_seq and potential
    dist = MSE(potential - soft_seq)

    #get gradients of soft_seq w.r.t. dist
    gradients = get_grads(soft_seq, dist)

  #set update gradients
  update_grads = zeros_like(seq)

  #find top-k residues closest to desired aa
  top-k_resi_list = get_topk(soft_seq[:,aa], (L * frac_to_bias))

  #iterate over each residue in the sequence
  for resi in num_residues:

      #neg gradient to bias toward aa of interest
      if resi in top-k_resi_list:
        update_grads[resi,:] = -gradients[resi,:]

      #pos gradient to bias away aa_of_interest
   else:
        update_grads[resi,:] = gradients[resi,:]

  #pos gradient to bias away
  grad_stack.append(update_grads)


  #average over multiple gradients when biasing for more than one aa
  update_grads = mean(grad_stack)

  return update_grads
```

Supplementary Pseudocode 5A: Net charge potential.

```
def charge_at_pH(seq, pH, target_charge):

  #softmax
  soft_seq = softmax(seq)

  #get table of AA partial charges at pH
  #based on Henderson Hasselbach equation
  pos_charges = (1.0 / (10 ** (pH - pos_pKs_matrix)) + 1.0)
  neg_charges = (1.0 / (10 ** (neg_pKs_matrix - pH)) + 1.0)

  #make table based on sequence length of all combinations of
  #positive, negative, and neutral charges that sum to target_charge
  table = make_table(seq.shape[0])

  #classify each position of soft_seq as
  #positive, negative, or neutral
  charge_classification = classify_resis(soft_seq)

  #find closest table entry that sums to target_charge
  #based on the currently classified soft_seq
  target_charge_ratios = get_target_charge_ratios(table, charge_classification)

  #determine gradients at each position based on target_charge_ratio
  #and the current charge_classification
  #return +1 for positions that should have positive gradients, -1 for positions
  #that should have negative gradients, else 0
  Guided_charge_classification = draft_resis(target_charge_ratios,
    charge_classification)

  #sum of soft charges at each position
  soft_charge = sum(soft_seq * (pos_charges - neg_charges), dim = -1)

  #calculate MSE loss with respect to soft_charge
  loss = mean(((guided_charge_classification - soft_charge)**2)**0.5)
  loss.backward()

  #get gradients
  gradients = soft_seq.grad

  return gradients
```

Supplementary Pseudocode 5B: Net charge potential data structures & tables.

```
# pKa lists to account for every residue.
pos_pKs = [[0.0, 12.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 5.98, 0.0, 0.0, 10.0, 0.0,
    0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]]
neg_pKs = [[0.0, 0.0, 0.0, 4.05, 9.0, 0.0, 4.45, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
    0.0, 0.0, 0.0, 0.0, 0.0, 10.0, 0.0, 0.0]]
cterm_pKs = [[0.0, 0.0, 0.0, 4.55, 0.0, 0.0, 4.75, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
    0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]]
nterm_pKs = [[7.59, 0.0, 0.0, 0.0, 0.0, 0.0, 7.7, 0.0, 0.0, 0.0, 0.0, 0.0, 7.0,
    0.0, 8.36, 6.93, 6.82, 0.0, 0.0, 7.44, 0.0]]

# Repeat charged pKs L - 2 times to populate in all non-terminal residue indices
pos_pKs_repeat = self.pos_pKs.repeat(seq.shape[0] - 2, 1)
neg_pKs_repeat = self.neg_pKs.repeat(seq.shape[0] - 2, 1)

# Concatenate all pKs tensors with N-term and C-term pKas to get full L X 21
    charge matrix
self.pos_pKs_matrix = cat((zeros_like(nterm_pKs), pos_pKs_repeat, self.nterm_pKs))
self.neg_pKs_matrix = cat((cterm_pKs, neg_pKs_repeat, zeros_like(self.cterm_pKs)))
```

Supplementary Pseudocode 6A: Hydrophobicity potential.

```
def hydropathy_index(seq, target_score):

  #get table of hydropathy values
  hydropathy_matrix = hydropathy_list.repeat(seq.shape[0])

  #softmax
  soft_seq = torch.softmax(seq)

  #sum of (softmax * hydropathy values)
  hydropathy_score = sum(soft_seq * hydropathy_matrix, dim = -1)

  #calculate MSE loss with respect to soft_seq
  loss = ((hydropathy_score - target_score)**2)**0.5
  loss.backward()

  #get gradients
  gradients = soft_seq.grad

  return gradients
```

Supplementary Pseudocode 6B: Hydrophobicity potential data structures and tables.

```
# AA conversion
conversion_list = list("ARNDCQEGHILKMFPSTWYVX")

# Dictionary to convert amino acids to their hyropathy index
gravy_dict = {'C': 2.5, 'D': -3.5, 'S': -0.8, 'Q': -3.5, 'K': -3.9,
        'I': 4.5, 'P': -1.6, 'T': -0.7, 'F': 2.8, 'N': -3.5,
        'G': -0.4, 'H': -3.2, 'L': 3.8, 'R': -4.5, 'W': -0.9,
        'A': 1.8, 'V':4.2, 'E': -3.5, 'Y': -1.3, 'M': 1.9, 'X': 0, '-': 0}

gravy_list = [self.gravy_dict[a] for a in self.conversion_list]
```

Supplementary Table 1: Observed and predicted mass for designs used in mass spectrometry experiments.

| Design | Observed Mass (Da) | Predicted Mass (Da) | Difference |
|---|---|---|---|
| TrpA5 | 11955 | 12086.18 | -131.18 |
| TrpA9 | 11592 | 11722.61 | -130.61 |
| TrpB9 | 11312 | 11765.34 | -453.34 |
| ValF3 | 10470 | 10600.58 | -130.58 |
| ValF12 | 11203 | 11333.68 | -130.68 |
| ValF11 | 10850 | 10981.23 | -131.23 |
| UncC7 | 11690 | 11820.99 | -130.99 |
| UncC4 | 10851 | 10981.47 | -130.47 |
| UncD5 | 11191 | 11321.68 | -130.67 |
| RCC12 | 23072 | 23202.83 | -130.83 |
| RCE4 | 22880 | 23010.88 | -130.88 |
| RCE8 | 22575 | 22705.43 | -130.43 |
| RCF12 | 22997 | 22996.3 | 0.7 |
| NCG1 | 21162 | 21292.13 | -130.13 |
| NCA1 | 20779 | 20910.78 | -131.78 |
| CysG4 | 8427 | 8561.43 | -134.43 |
| CysH6 | 8675 | 8813.75 | -138.75 |
| CysH11 | 8633 | 8642.57 | -9.57 |
| TEV A1 | 14905 | 15036 | -131 |
| TEV A2 | 14713 | 14844 | -131 |
| TEV A3 | 14005 | 14136 | -131 |
| TEV A4 | 14044 | 14175 | -131 |
| TEV A5 | 14265 | 14396 | -131 |
| TEV A6 | 14178 | 14309 | -131 |
| TEV A7 | 14714 | 14845 | -131 |
| FUR A9 | 22894 | 23025 | -131 |
| FUR A10 | 23556 | 23687 | -131 |
| TEV B1 | 14864 | 14995 | -131 |
| TEV B2 | 14312 | 14443 | -131 |
| TEV B3 | 14561 | 14692 | -131 |
| TEV B6 | 14547 | 14661 | -114 |
| FUR B10 | 19733 | 19916 | -183 |
| FUR C2 | 18934 | 19065 | -131 |
| FUR D2 | 18853 | 18984 | -131 |
| FUR D3 | 18869 | 19000 | -131 |

Supplementary Table 2: Mass spec data of experimentally validated cysteine-rich designs. The mass of each design is reported in the presence and absence of the reducing agent TCEP. The mass difference between reduced and non-reduced designs is used to calculate the number of disulfides formed and compared to the number of designed disulfides.

| Design ID | Mass with MET loss (Da) | | | # Disulfides formed |
|---|---|---|---|---|
| | + TCEP | - TCEP | Difference | |
| cys_G4 | 8431 | 8425 | 6 | 3/3 |
| cys_G6 | 8778 | 8770 | 8 | 4/4 |
| cys_G10 | 8695 | 8687 | 8 | 4/4 |
| cys_G11 | 8368 | 8363 | 5 | 3/3 |
| cys_H6 | 8683 | 8675 | 8 | 4/5 |

Supplementary Table 3: Secondary structure prediction of CD data (200-250 nm) of designs RC_E8 Fig 3E middle top, and RC_F11 Fig 3E middle bottom with BeStSel server indicating high percentage of beta content.

| | CD RC_E8 | CD RC_F11 |
|---|---|---|
| **Helix** | 1.0 | 2.7 |
| regular | 0.2 | 1.2 |
| distorted | 0.8 | 1.5 |
| **Antiparallel** | 40.7 | 26.8 |
| left-twisted | 0.0 | 0.0 |
| relaxed | 24.3 | 12.2 |
| right-twisted | 16.4 | 14.6 |
| **Parallel** | 0.0 | 0.0 |
| **Turn** | 14.0 | 17.9 |
| **Others** | 44.3 | 52.7 |

# Chapter 4. Towards Iterative Optimization with Experimental Feedback with Co-Sequence and Structure Generative Diffusion Models

This work was done in close collaboration with Jacob Gershon, with experimental work done by Kiera Sumida. Figure 4.1 was repurposed from Madison Kennedy.

## 4.1 Abstract

For particularly difficult protein design problems such as the design of highly active enzymes, experimental data feedback is necessary to improve functionality with minimal design iterations. Active learning (AL) and bayesian optimization (BO) approaches provide a principled way to incorporate experimental feedback into the design process, and subsequently minimize the number of iterations cycling between computation and experimental testing to optimize the desired function (enzymatic activity, binding, self-assembly, etc.). Current active learning approaches do not incorporate strong generative priors to bias exploration/exploitations to valid regions of protein space. We hypothesize that coupling a joint sequence and structure diffusion model with bayesian optimization methods will allow for the more efficient search of the sequence activity landscape to find highly active variants. To this end we develop a joint sequence structure generative model, ProteinGenerator2 (PG2), to which we bias generation with zero shot predictors to yield predicted highly active and diverse sequence pools for testing. From which at each experimental round the collected data from preceding rounds will be used to further increase the accuracy of the activity predictor and further bias the design process toward more active variants with the desired function.

## 4.2 Introduction

Directed evolution has been used to successfully engineer proteins with enhanced functionality. This is typically done in an iterative fashion where a random library of mutations is generated for a parent sequence and screened, at which point highly active mutations are fixed (typically the top mutation), and subsequent rounds of mutagenesis and screening are done until a satisfactory stop criterion is met. This approach is now increasingly supplemented with Bayesian optimization (BO) based machine learning methods to better explore the possible mutational space efficiently[115] and enable larger steps through sequence space by learning epistatic interaction not possible with classical directed evolution.

Bayesian optimization techniques build a probabilistic model of the underlying sequence-to-activity landscape via a surrogate model. This surrogate model in addition to predicting mean activity also provides an uncertainty estimation. An acquisition function utilizes this uncertainty, balancing between exploration and exploitation, to select new points to evaluate experimentally. Where the ultimate goal is to guide the sequence space search towards the global optimum while minimizing the number of experimental rounds.

Typical BO supplemented directed evolution involves exhaustively sampling and evaluating the sequence search space in silico only testing highly predicted variants[116]. This strategy is inherently inefficient in that queries to the predictor are exhaustive and there is no way to efficiently bias the search space to plausible sequences. We hypothesize that querying the acquisition function with sequences proposed by a generative model, ProteinGenerator2, that understands the sequence-to-structure mapping will allow for the more efficient search of the sequence to function landscape. In addition, because ProteinGenerator2 is a diffusion model, gradient based guidance can bias the generated outputs to maximize the predicted activity of the

surrogate model(s) further restricting search to a functional subspace. A schematic of the proposed experimental workflow is in Figure 4.1.

## 4.3 Developing Joint Sequence and Structure Generative Model

An ideal generative model for enzyme optimization should be able to consider interactions of the enzyme and its substrates, in addition gradient based guidance should be available at both the sequence and structural level to allow for fine grained optimization of catalytically necessary structural and sequence features. We sought to further develop ProteinGenerator[21] a state of the art joint sequence and structure generative model with these capabilities. ProteinGenerator was initially trained with only noising the amino acid sequence input to the model, and was tasked with not only denoising the noised sequence, but also predicting the corresponding structure. This is limiting in that gradient based guidance could only be applied in sequence space. Further, the model was only able to be conditioned on protein sequence and structural motif, not small molecules or other macromolecules.

To address the model's inability for gradient based structural guidance we incorporated structural 3di tokens previously used in the FoldSeek structural search algorithm[117] in addition to sequence tokens and structural motifs as inputs to the model. We corrupt these structural tokens on the same noising schedule as the amino acid tokens, and task the model with denoising both simultaneously. To address being able to model non-protein substrates we replace RosettaFold2[118] with RosettaFold All-Atom[119] as our denoising network. Self consistency benchmarks with AlphaFold structure prediction of generated sequence-structure pairs show that ProteinGenerator2 improved performance over the original on unconditional generation of large proteins, with comparable results on smaller length designs. (Figure 4.2).

## 4.4 Structure and Sequence Based In Silico Surrogate Models

As an initial application of our proposed diffusion based iterative design approach, we chose to redesign a native Poly(ethylene terephthalate) (PET) hydrolyzing enzyme, PDB ID: 7VVC, for improved catalytic activity and thermostability. PET degradation into reusable subunits is of ecological importance because it composes 70% of synthetic textile fibers and 10% of non fiber plastic packaging[120]. We build on previous work on native redesign with ProteinMPNN coupled with evolutionary features to redesign the petase and screen for activity[121] for an initial design pool to train on. Given this initial dataset of 96 designs with sequence, Alphafold predicted structures, and catalytic rates we sought to see if there were any zero-shot activity predictors which correlate with activity for either sequence or structural inputs, because this number of data-points is inefficient to train a model to map the sequence identity landscape.

Zero-shot predictors of protein function have been found to be useful in restricting the search space in machine learning assisted directed evolution campaigns[116,122]. Inspired by AlphaFold's capability to discriminate decoy structures for a given target sequence[123] we investigated whether AlphaFold could serve as the structural zero-shot predictor given that a common protein design failure mode is that a sequence may not adopt the intended structure. We implemented the AF2rank approach described in Roney et al. 2022[123], and used the template structure pLDDT as a metric for catalytic activity. We found a moderate correlation of 0.26 with activity indicating that this structure based approach could serve as a zero-shot predictor (Figure 4.3.A). It is worth noting we opted for the template approach rather than single sequence structure prediction because AlphaFold fails to predict sequences close to the native sequence

space without an MSA, and folding with an MSA gives overconfident predictions of the structure regardless if the query sequence is valid.

Previous work has shown language models trained on massive corpus of sequence only data information emerges in the learned representations on fundamental properties of proteins such as structure and biological activity[4,124]. These representations have further shown to be able to pick mutations to increase antibody binding affinity and expression in multiple rounds[122]. We sought to determine if ESM2 pseudo-perplexity could serve as a zero shot predictor. We implemented pseudo-perplexity calculations as shown in Verkuil et. al, 2022[19], and calculated it on all of the sequences within our initial training set. We found a strong correlation in activity of -0.5 on the calculated activity rate (Figure 4.3.B). We hypothesize that the reason for this strong correlation is that ESM2 learned the space of plausible mutations for a given sequence, and this space is a much smaller subspace of all possible mutations so sequences within this subspace are more likely to be within the even smaller activity subspace.

## 4.5 Guidance with Zero Shot Predictors

Diffusion models are particularly powerful generative models over other approaches such as GANs, VAEs, or autoregressive language models because of their ability to easily guide generation with off the shelf classifiers[20]. We utilize this capability to train a classifier on the generated sequences of ProteinGenerator2 and their corresponding ESM2 pseudo-perplexity scores to predict pseudo-perplexity. For the generated sequences we provide the model with the wildtype structure of petase with the substrate modeled, and allow redesign of evolutionary nonconserved regions defined by a consensus amino acid not being present in more than 50 percent of the amino acids at a given sequence position in the wild type MSA. We then use classifier guidance as was done in Lisanza et al. 2023[21] with the trained classifier. We chose this

approach rather than backpropagating through ESM2 because the language model is large and backpropagation through would be quite expensive in both time and memory[125]. True ESM2 pseudo-perplexity improves significantly with classifier based guidance (Figure 4.4.A). We then applied this approach to predict AF2 pLDDt values from round 0 designs and then guided generation for round 1 designs. We saw a striking increase in AF2 predicted confidence, with some loss in sequence diversity.

## 4.6 Conclusion and Future Work

We have developed a joint sequence and structure denoising model ProteinGenerator2 that can model non-protein substrates. We have used this model to generate plausible sequences for redesign wild type petase with the ultimate goal of increasing activity and thermostability. To this end we biased generation with gradients from a zero-shot activity predictor (ESM2) that we found to be correlated with experimental activity. ProteinGenerator2 is well positioned to be utilized within active-learning and bayesian optimization pipelines as the one proposed in Figure 4.1 because 1) it serves as generative prior for the more efficient search of the plausible protein sequence landscape, and 2) gradient based guidance can bias the generated outputs to maximize the predicted activity of surrogate model(s); further restricting search to functional sequence subspaces. Future work will couple ProteinGenerator2 with uncertainty collaborated surrogate models trained on experimental data from preceding rounds. Where we envision the predicted activity (mean) of these surrogate models will be used guide, while the predicted uncertainty in conjunction with an acquisition function will choose which sequence to test inorder to navigate the exploration and exploitation tradeoff to generate both increasingly active variants and learn the sequence to activity landscape.

## 4.7 Main Figures

Figure 4.1: **Iterative Design with Experimental Feedback Envisioned Pipeline**

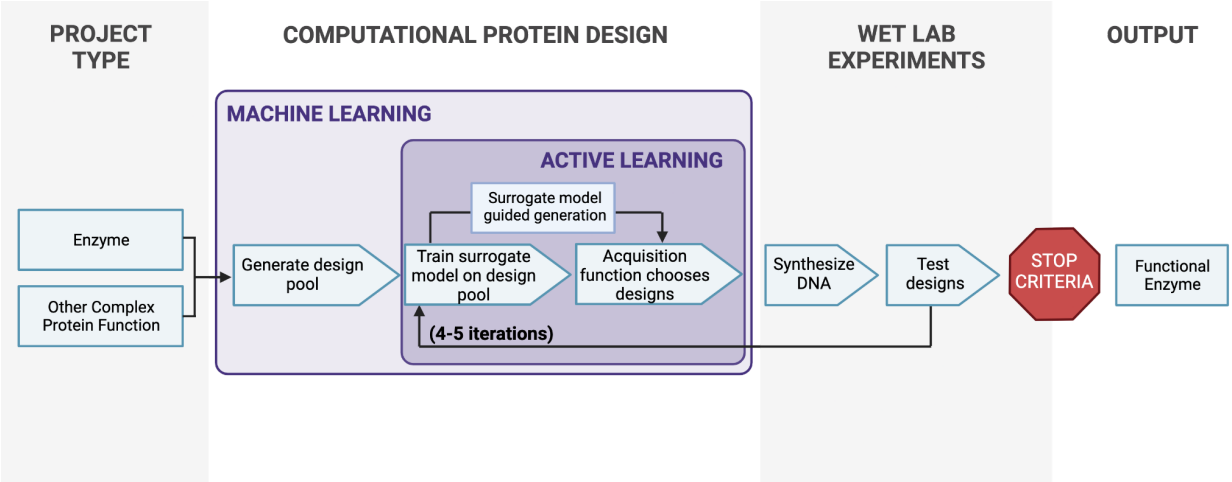



**Figure 4.1: Iterative Design with Experimental Feedback Envisioned Pipeline**

Figure 4.2: **Incorporation of 3di tokens and RF-AA as the denoiser improves AlphaFold2 self consistency performance.**

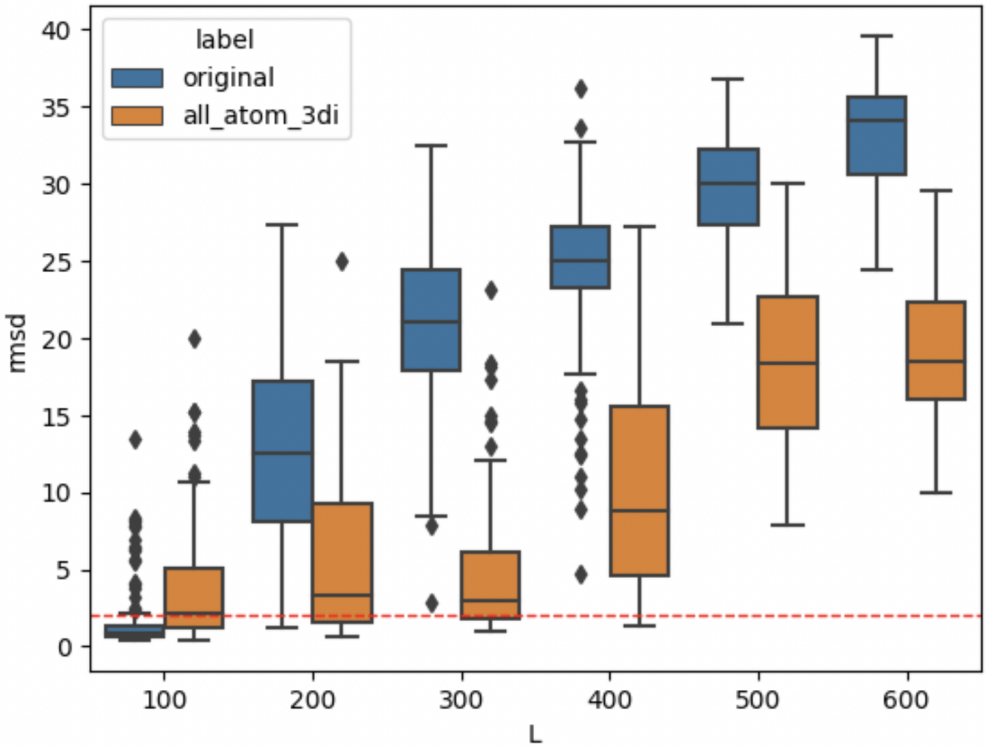


**Figure 4.2: Incorporation of 3di tokens and RF-AA as the denoiser improves AlphaFold2 self consistency performance.** Horizontal dotted red line is 2 angstrom RMSD.

Figure 4.3: **AlphaFold template pLDDT and ESM2 pseudo-perplexity serve as zero-shot predictors of petase activity.**

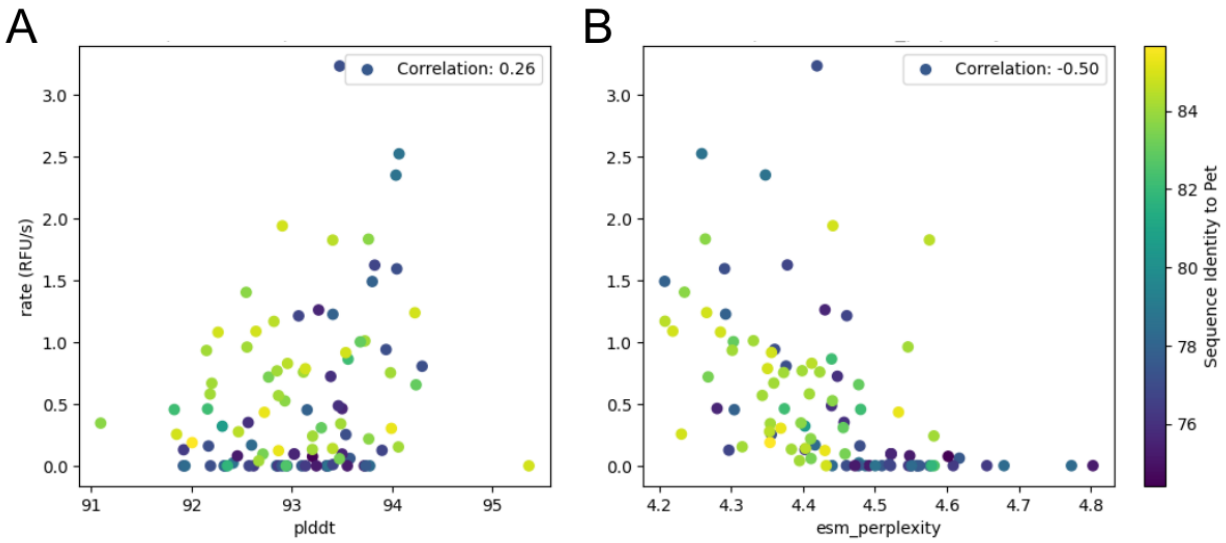
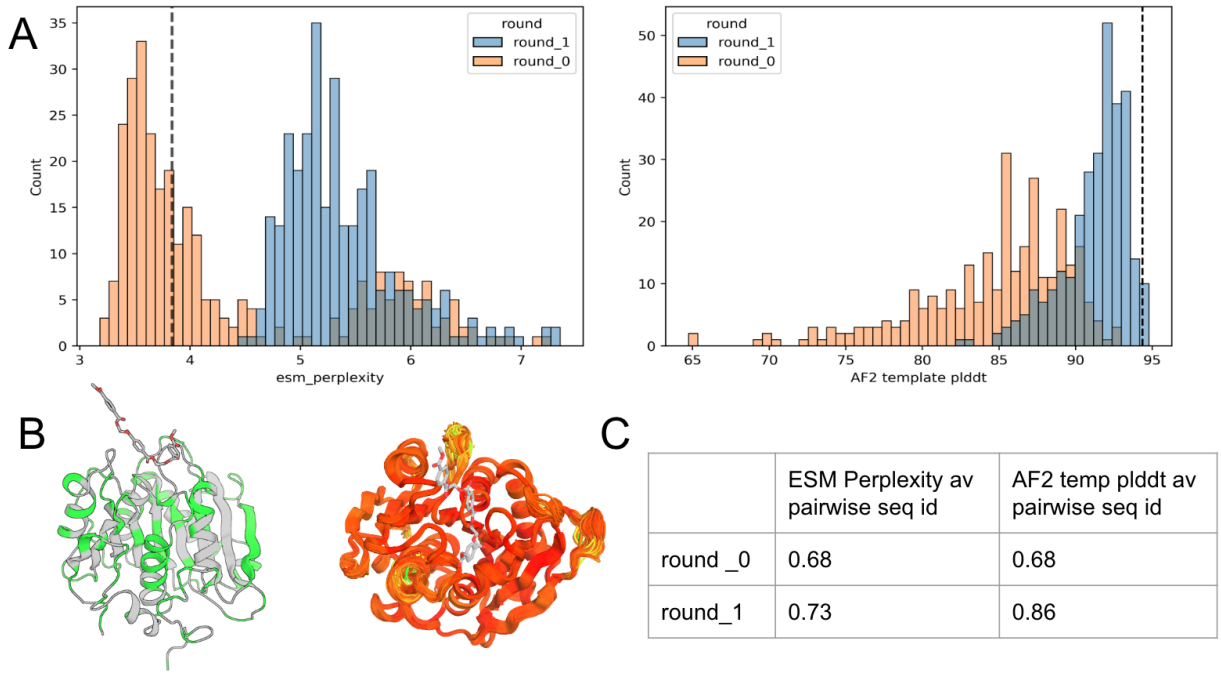



**Figure 4.3: AlphaFold template pLDDT and ESM2 pseudo-perplexity serve as zero-shot predictors of petase activity.** A) Catalytic rate vs AlphaFold template plddt. B) Catalytic rate vs ESM2 pseudo-perplexity.

Figure 4.4: **Guidance with ESM perplexity and AF2 pLDDt classifiers, results in lower true ESM perplexity scores and higher true plddt scores**




| | ESM Perplexity av pairwise seq id | AF2 temp plddt av pairwise seq id |
|---|---|---|
| round _0 | 0.68 | 0.68 |
| round_1 | 0.73 | 0.86 |

**Figure 4.4: Guidance with ESM perplexity and AF2 pLDDt classifiers, results in lower true ESM perplexity scores and higher true plddt scores.** A) Round 0 is without guidance, round 1 is with guidance based on the classifier trained on round 0 designs. The vertical dotted line is wild type pseudo-perplexit or AF2 template plddt. B) green regions allowed to be redesigned, gray regions held fixed. B) Round 0 is without guidance, round 1 is with guidance based on the classifier trained on round 0 designs. The shaded region corresponds to the initial set of 96 designs, and the vertical dotted line is wild type pseudo-perplexit. Sample generated designs with model confidence coloring. Red is high confidence and green is low confidence. C) pairwise sequence identity per round indicates diversity is still maintained even with classifier guidance.

## BIBLIOGRAPHY

1. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

2. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).

3. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).

4. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
5. Huang, P.-S. *et al.* RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLOS ONE* **6**, e24109 (2011).
6. Maguire, J. B. *et al.* Perturbing the energy landscape for improved packing during computational protein design. *Proteins* **89**, 436–449 (2021).
7. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).
8. Yeh, A. H.-W. *et al.* De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).
9. Bale, J. B. *et al.* Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353**, 389–394 (2016).
10. Li, Z. *et al.* Accurate computational design of three-dimensional protein crystals. *Nat. Mater.* (2023) doi:10.1038/s41563-023-01683-1.
11. Anand, N. *et al.* Protein sequence design with a learned potential. *Nat. Commun.* **13**, 746 (2022).
12. Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. Generative Models for Graph-Based Protein Design. (2022).
13. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. 2022.04.10.487779 Preprint at https://doi.org/10.1101/2022.04.10.487779 (2022).
14. Nguyen, A., Yosinski, J. & Clune, J. Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks. Preprint at https://doi.org/10.48550/arXiv.1602.03616 (2016).

15. Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
16. Wicky, B. I. M. *et al.* Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).
17. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
18. Hie, B. *et al.* A high-level programming language for generative protein design. 2022.12.21.521526 Preprint at https://doi.org/10.1101/2022.12.21.521526 (2022).
19. Verkuil, R. *et al.* Language models generalize beyond natural proteins. 2022.12.21.521521 Preprint at https://doi.org/10.1101/2022.12.21.521521 (2022).
20. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. Preprint at http://arxiv.org/abs/2006.11239 (2020).
21. Lisanza, S. L. *et al.* Joint Generation of Protein Sequence and Structure with RoseTTAFold Sequence Space Diffusion. 2023.05.08.539766 Preprint at https://doi.org/10.1101/2023.05.08.539766 (2023).
22. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
23. Siegel, J. B. *et al.* Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **329**, 309–313 (2010).
24. De Novo Computational Design of Retro-Aldol Enzymes | Science. https://www.science.org/doi/10.1126/science.1152692.
25. Chevalier, A. *et al.* Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).

26. Correia, B. E. *et al.* Proof of principle for epitope-focused vaccine design. *Nature* **507**, 201–206 (2014).
27. Procko, E. *et al.* A Computationally Designed Inhibitor of an Epstein-Barr Viral Bcl-2 Protein Induces Apoptosis in Infected Cells. *Cell* **157**, 1644–1656 (2014).
28. Silva, D.-A. *et al.* De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186–191 (2019).
29. De novo protein design enables the precise induction of RSV-neutralizing antibodies | Science. https://www.science.org/doi/10.1126/science.aay5051.
30. Yang, C. *et al.* Bottom-up de novo design of functional proteins with complex structural features. *Nat. Chem. Biol.* **17**, 492–500 (2021).
31. Improved protein structure prediction using predicted interresidue orientations | PNAS. https://www.pnas.org/doi/full/10.1073/pnas.1914677117.
32. Norn, C. *et al.* Protein sequence design by conformational landscape optimization. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2017228118 (2021).
33. Tischer, D. *et al.* Design of proteins presenting discontinuous functional sites using deep learning. 2020.11.29.402743 Preprint at https://doi.org/10.1101/2020.11.29.402743 (2020).
34. Pascolutti, R. *et al.* Structure and Dynamics of PD-L1 and an Ultra-High-Affinity PD-1 Receptor Mutant. *Structure* **24**, 1719–1728 (2016).
35. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
36. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at https://doi.org/10.48550/arXiv.1810.04805 (2019).

37. Yeh, R. A. *et al.* Semantic Image Inpainting with Deep Generative Models. Preprint at https://doi.org/10.48550/arXiv.1607.07539 (2017).

38. Li, Z., Nguyen, S. P., Xu, D. & Shang, Y. Protein Loop Modeling Using Deep Generative Adversarial Network. in *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)* 1085–1091 (2017). doi:10.1109/ICTAI.2017.00166.

39. Anand, N. & Huang, P. Generative modeling for protein structures. in *Advances in Neural Information Processing Systems* vol. 31 (Curran Associates, Inc., 2018).

40. Chowdhury, R. *et al.* Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* **40**, 1617–1623 (2022).

41. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins Struct. Funct. Bioinforma.* **37**, 171–176 (1999).

42. Kim, T.-E. *et al.* Dissecting the stability determinants of a challenging de novo protein fold using massively parallel design and experimentation. *Proc. Natl. Acad. Sci.* **119**, e2122676119 (2022).

43. Pak, M. A. *et al.* Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLOS ONE* **18**, e0282689 (2023).

44. Buel, G. R. & Walters, K. J. Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. Mol. Biol.* **29**, 1–2 (2022).

45. Mousa, J. J., Kose, N., Matta, P., Gilchuk, P. & Crowe, J. E. A novel pre-fusion conformation-specific neutralizing epitope on the respiratory syncytial virus fusion protein. *Nat. Microbiol.* **2**, 1–8 (2017).

46. Linsky, T. W. *et al.* De novo design of potent and resilient hACE2 decoys to neutralize

SARS-CoV-2. *Science* **370**, 1208–1214 (2020).

47. Frolow, F., Kalb (Gilboa), A. J. & Yariv, J. Structure of a unique twofold symmetric haem-binding site. *Nat. Struct. Biol.* **1**, 453–460 (1994).
48. Lombardi, A., Pirro, F., Maglio, O., Chino, M. & DeGrado, W. F. De Novo Design of Four-Helix Bundle Metalloproteins: One Scaffold, Diverse Reactivities. *Acc. Chem. Res.* **52**, 1148–1159 (2019).
49. Calhoun, J. R. *et al.* Artificial diiron proteins: From structure to function. *Pept. Sci.* **80**, 264–278 (2005).
50. Keech, A. M. *et al.* Spectroscopic Studies of Cobalt(II) Binding to Escherichia coli Bacterioferritin*. *J. Biol. Chem.* **272**, 422–429 (1997).
51. Marsh, E. N. G. & DeGrado, W. F. Noncovalent self-assembly of a heterotetrameric diiron protein. *Proc. Natl. Acad. Sci.* **99**, 5150–5154 (2002).
52. Yáñez, M., Gil-Longo, J. & Campos-Toimil, M. Calcium Binding Proteins. in *Calcium Signaling* (ed. Islam, Md. S.) 461–482 (Springer Netherlands, 2012). doi:10.1007/978-94-007-2888-2_19.
53. Caldwell, S. J. *et al.* Tight and specific lanthanide binding in a de novo TIM barrel with a large internal cavity designed by symmetric domain fusion. *Proc. Natl. Acad. Sci.* **117**, 30362–30369 (2020).
54. Cho, H.-S. *et al.* Crystal Structure of Δ5-3-Ketosteroid Isomerase from Pseudomonas testosteroni in Complex with Equilenin Settles the Correct Hydrogen Bonding Scheme for Transition State Stabilization*. *J. Biol. Chem.* **274**, 32863–32868 (1999).
55. Engineering high-affinity PD-1 variants for optimized immunotherapy and immuno-PET imaging | PNAS. https://www.pnas.org/doi/full/10.1073/pnas.1519623112.

56. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

57. Wiesmann, C., Ultsch, M. H., Bass, S. H. & de Vos, A. M. Crystal structure of nerve growth factor in complex with the ligand-binding domain of the TrkA receptor. *Nature* **401**, 184–188 (1999).

58. Humphreys, I. R. *et al.* Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).

59. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

60. Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A. & Kim, P. M. Fast and Flexible Protein Design Using Deep Graph Neural Networks. *Cell Syst.* **11**, 402-411.e4 (2020).

61. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).

62. Repecka, D. *et al.* Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **3**, 324–333 (2021).

63. Shin, J.-E. *et al.* Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* **12**, 2403 (2021).

64. Wu, Z., Johnston, K. E., Arnold, F. H. & Yang, K. K. Protein sequence design with deep generative models. *Curr. Opin. Chem. Biol.* **65**, 18–27 (2021).

65. Madani, A. *et al.* Deep neural language modeling enables functional protein generation across families. 2021.07.18.452833 Preprint at https://doi.org/10.1101/2021.07.18.452833 (2021).

66. Ovchinnikov, S. & Huang, P.-S. Structure-based protein design with deep learning. *Curr. Opin. Chem. Biol.* **65**, 136–144 (2021).

67. Anand, N., Eguchi, R. & Huang, P.-S. *Fully Differentiable Full-Atom Protein Back-Bone Generation*. (2019).

68. Eguchi, R. R., Choe, C. A. & Huang, P.-S. Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. *PLOS Comput. Biol.* **18**, e1010271 (2022).

69. Lin, Z., Sercu, T., LeCun, Y. & Rives, A. Deep generative models create new and diverse protein structures.

70. Jendrusch, M., Korbel, J. O. & Sadiq, S. K. AlphaDesign: A de novo protein design framework based on AlphaFold. 2021.10.11.463937 Preprint at https://doi.org/10.1101/2021.10.11.463937 (2021).

71. Moffat, L., Kandathil, S. M. & Jones, D. T. Design in the DARK: Learning Deep Generative Models for De Novo Protein Design. 2022.01.27.478087 Preprint at https://doi.org/10.1101/2022.01.27.478087 (2022).

72. Moffat, L., Greener, J. G. & Jones, D. T. Using AlphaFold for Rapid and Accurate Fixed Backbone Protein Design. 2021.08.24.457549 Preprint at https://doi.org/10.1101/2021.08.24.457549 (2021).

73. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

74. Anand, N. & Achim, T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. *arXiv.org* https://arxiv.org/abs/2205.15019v1 (2022).

75. Watson, J. L. *et al.* Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. 2022.12.09.519842 Preprint at https://doi.org/10.1101/2022.12.09.519842 (2022).

76. Ingraham, J. *et al.* Illuminating protein space with a programmable generative model. 2022.12.01.518682 Preprint at https://doi.org/10.1101/2022.12.01.518682 (2022).

77. Brown, T. B. *et al.* Language Models are Few-Shot Learners. Preprint at https://doi.org/10.48550/arXiv.2005.14165 (2020).

78. Frank, C. *et al.* Efficient and scalable de novo protein design using a relaxed sequence space. 2023.02.24.529906 Preprint at https://doi.org/10.1101/2023.02.24.529906 (2023).

79. Chen, T., Zhang, R. & Hinton, G. Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning. Preprint at http://arxiv.org/abs/2208.04202 (2022).

80. Han, X., Kumar, S. & Tsvetkov, Y. SSD-LM: Semi-autoregressive Simplex-based Diffusion Language Model for Text Generation and Modular Control. Preprint at http://arxiv.org/abs/2210.17432 (2022).

81. Li, X. L., Thickstun, J., Gulrajani, I., Liang, P. & Hashimoto, T. B. Diffusion-LM Improves Controllable Text Generation. Preprint at https://doi.org/10.48550/arXiv.2205.14217 (2022).

82. Dieleman, S. *et al.* Continuous diffusion for categorical data. Preprint at http://arxiv.org/abs/2211.15089 (2022).

83. Dhariwal, P. & Nichol, A. Diffusion Models Beat GANs on Image Synthesis. *arXiv.org* https://arxiv.org/abs/2105.05233v4 (2021).

84. Nachmani, E., Roman, R. S. & Wolf, L. Non Gaussian Denoising Diffusion Models. Preprint at http://arxiv.org/abs/2106.07582 (2021).

85. Labesse, G., Colloc'h, N., Pothier, J. & Mornon, J. P. P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. *Comput. Appl. Biosci. CABIOS* **13**, 291–295 (1997).

86. Wei, K. Y. *et al.* Computational design of closely related proteins that adopt two well-defined

but structurally divergent folds. *Proc. Natl. Acad. Sci.* **117**, 7208–7215 (2020).

87. Tokmakov, A. A., Kurotani, A. & Sato, K.-I. Protein pI and Intracellular Localization. *Front. Mol. Biosci.* **8**, 775736 (2021).
88. Boswell, C. A. *et al.* Effects of Charge on Antibody Tissue Distribution and Pharmacokinetics. *Bioconjug. Chem.* **21**, 2153–2163 (2010).
89. March, D., Bianco, V. & Franzese, G. Protein Unfolding and Aggregation near a Hydrophobic Interface. *Polymers* **13**, 156 (2021).
90. Rego, N. B., Xi, E. & Patel, A. J. Identifying hydrophobic protein patches to inform protein interaction interfaces. *Proc. Natl. Acad. Sci.* **118**, e2018234118 (2021).
91. Zeng, Z. *et al.* Customized Reversible Stapling for Selective Delivery of Bioactive Peptides. *J. Am. Chem. Soc.* **144**, 23614–23621 (2022).
92. Lajoie, M. J. *et al.* Designed protein logic to target cells with precise combinations of surface antigens. *Science* **369**, 1637–1643 (2020).
93. Quijano-Rubio, A. *et al.* De novo design of modular and tunable protein biosensors. *Nature* **591**, 482–487 (2021).
94. Lee, M.-T., Sun, T.-L., Hung, W.-C. & Huang, H. W. Process of inducing pores in membranes by melittin. *Proc. Natl. Acad. Sci.* **110**, 14243–14248 (2013).
95. Duffy, C. *et al.* Honeybee venom and melittin suppress growth factor receptor activation in HER2-enriched and triple-negative breast cancer. *Npj Precis. Oncol.* **4**, 1–16 (2020).
96. Parmeggiani, F. & Huang, P.-S. Designing repeat proteins: a modular approach to protein design. *Curr. Opin. Struct. Biol.* **45**, 116–123 (2017).
97. Brunette, T. J. *et al.* Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015).

98. Peralta, M. D. R. *et al.* Engineering Amyloid Fibrils from β-Solenoid Proteins for Biomaterials Applications. *ACS Nano* **9**, 449–463 (2015).

99. MacDonald, J. T. *et al.* Synthetic beta-solenoid proteins with the fragment-free computational design of a beta-hairpin extension. *Proc. Natl. Acad. Sci.* **113**, 10346–10351 (2016).

100. Micsonai, A. *et al.* BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy. *Nucleic Acids Res.* **50**, W90–W98 (2022).

101. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2020).

102. Deszyński, P. *et al.* INDI—integrated nanobody database for immunoinformatics. *Nucleic Acids Res.* **50**, D1273–D1281 (2022).

103. Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. & Murzin, A. G. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.* **42**, D310–D314 (2014).

104. Andreeva, A., Kulesha, E., Gough, J. & Murzin, A. G. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* **48**, D376–D382 (2020).

105. Sternke, M., Tripp, K. W. & Barrick, D. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc. Natl. Acad. Sci.* **116**, 11275–11284 (2019).

106. Khersonsky, O. *et al.* Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol. Cell* **72**, 178-186.e5 (2018).

107. Matz, M. V. *et al.* Fluorescent proteins from nonbioluminescent Anthozoa species. *Nat. Biotechnol.* **17**, 969–973 (1999).

108. Ormö, M. *et al.* Crystal Structure of the Aequorea victoria Green Fluorescent Protein. *Science* **273**, 1392–1395 (1996).

109. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5**, e16965 (2016).

110. Hu, R. *et al.* Protein engineering via Bayesian optimization-guided evolutionary algorithm and robotic experiments. *Brief. Bioinform.* **24**, bbac570 (2023).

111. Voynov, V., Chennamsetty, N., Kayser, V., Helk, B. & Trout, B. L. Predictive tools for stabilization of therapeutic proteins. *mAbs* **1**, 580–582 (2009).

112. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

113. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

114. Dallago, C. *et al.* FLIP: Benchmark tasks in fitness landscape inference for proteins. 2021.11.09.467890 Preprint at https://doi.org/10.1101/2021.11.09.467890 (2022).

115. Johnston, K. E. *et al.* Machine Learning for Protein Engineering. Preprint at http://arxiv.org/abs/2305.16634 (2023).

116. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci.* **116**, 8852–8858 (2019).

117. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* 1–4 (2023) doi:10.1038/s41587-023-01773-0.

118. Baek, M. *et al.* Efficient and accurate prediction of protein structure using RoseTTAFold2. 2023.05.24.542179 Preprint at https://doi.org/10.1101/2023.05.24.542179

(2023).

119. Krishna, R. *et al.* Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom. 2023.10.09.561603 Preprint at https://doi.org/10.1101/2023.10.09.561603 (2023).

120. Lu, H. *et al.* Deep learning redesign of PETase for practical PET degrading applications. 2021.10.10.463845 Preprint at https://doi.org/10.1101/2021.10.10.463845 (2021).

121. Sumida, K. H. *et al.* Improving protein expression, stability, and function with ProteinMPNN. 2023.10.03.560713 Preprint at https://doi.org/10.1101/2023.10.03.560713 (2023).

122. Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* 1–9 (2023) doi:10.1038/s41587-023-01763-2.

123. Roney, J. P. & Ovchinnikov, S. State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Phys. Rev. Lett.* **129**, 238101 (2022).

124. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118**, e2016239118 (2021).

125. Jeliazkov, J. R., Alamo, D. del & Karpiak, J. D. ESMFold Hallucinates Native-Like Protein Sequences. 2023.05.23.541774 Preprint at https://doi.org/10.1101/2023.05.23.541774 (2023).

# VITA

Sidney Lisanza was born in Nairobi, Kenya to loving parents Esther Lisanza and Leonard Muaka. After a successful childhood in Nairobi and eager to bring his talents abroad, Sidney moved to the United States of America at the age of 7. There, he began his foray into science and engineering by building marvelously structurally intricate assemblies out of cardboard boxes and plastics found around his house. He slowly graduated to toy coding projects with the Scratch programming language. Excited to continue his scientific journey, Sidney attended the North Carolina School of Science and Mathematics, where he was afforded the opportunity to do research under the guidance of Vikas Bhandawat at Duke University. From there, Sidney attended the University of North Carolina at Chapel Hill where he studied Chemistry, Math, and Computer Science. He then pursued doctoral work at the University of Washington at the Institute for Protein Design under the supervision of David Baker. There, he studied how to develop protein generators, a class of protein design models that generate sequence and structure pairs simultaneously. Upon graduating, Sidney is interested in applying protein design to address some of the unmet needs within the Global South, specifically food sovereignty and medicine.