RESEARCH ARTICLE

PROTEIN DESIGN

Scaffolding protein functional sites using deep learning

Jue Wang[1,2]†, Sidney Lisanza[1,2,3]†, David Juergens[1,2,4]†, Doug Tischer[1,2]†, Joseph L. Watson[1,2]†, Karla M. Castro[5], Robert Ragotte[1,2], Amijai Saragovi[1,2], Lukas F. Milles[1,2], Minkyung Baek[1,2], Ivan Anishchenko[1,2], Wei Yang[1,2], Derrick R. Hicks[1,2], Marc Expòsit[1,2,4], Thomas Schlichthaerle[1,2], Jung-Ho Chun[1,2,3], Justas Dauparas[1,2], Nathaniel Bennett[1,2,4], Basile I. M. Wicky[1,2], Andrew Muenks[1,2], Frank DiMaio[1,2], Bruno Correia[5], Sergey Ovchinnikov[6,7]*, David Baker[1,2,8]*

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.

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 (*1*–*3*), and for protein binders, by fragment docking calculations (*4*, *5*). Alternatively, functional sites can be extracted from a native protein having the desired activity (*6*, *7*). 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 (*6*–*10*), but, with the exception of helical bundles (*8*), 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 (*11*) 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 (*12*). 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 (*12*). 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 (*13*). Given this ability to design both sequence and structure, we reasoned that trRosetta could be adapted to tackle the functional site scaffolding problem.

[1]Department of Biochemistry, University of Washington, Seattle, WA 98105, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA 98105, USA. [3]Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA 98105, USA. [4]Molecular Engineering Graduate Program, University of Washington, Seattle, WA 98105, USA. [5]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland. [6]FAS Division of Science, Harvard University, Cambridge, MA 02138, USA. [7]John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA 02138, USA. [8]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA.
*Corresponding author. Email: dabaker@uw.edu (D.B.); so@fas.harvard.edu (S.O.)
†These authors contributed equally to this work.

Partially constrained hallucination using a multiobjective loss function

To extend existing trRosetta-based design methods to scaffold functional sites (Fig. 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 (*13*)] (Fig. 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. S1A), and hence we used Rosetta design calculations to generate more-optimal sequences (*14*). 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) (*15*), followed by Rosetta design, were found to have binding affinities in the mid-nanomolar range (fig. 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) (*16*), we found that it performed better than trRosetta in guiding protein design by functional site–constrained hallucination (fig. S1G), likely reflecting the better overall modeling of protein sequence-structure relationships (*16*). 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. S2; materials and methods).

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

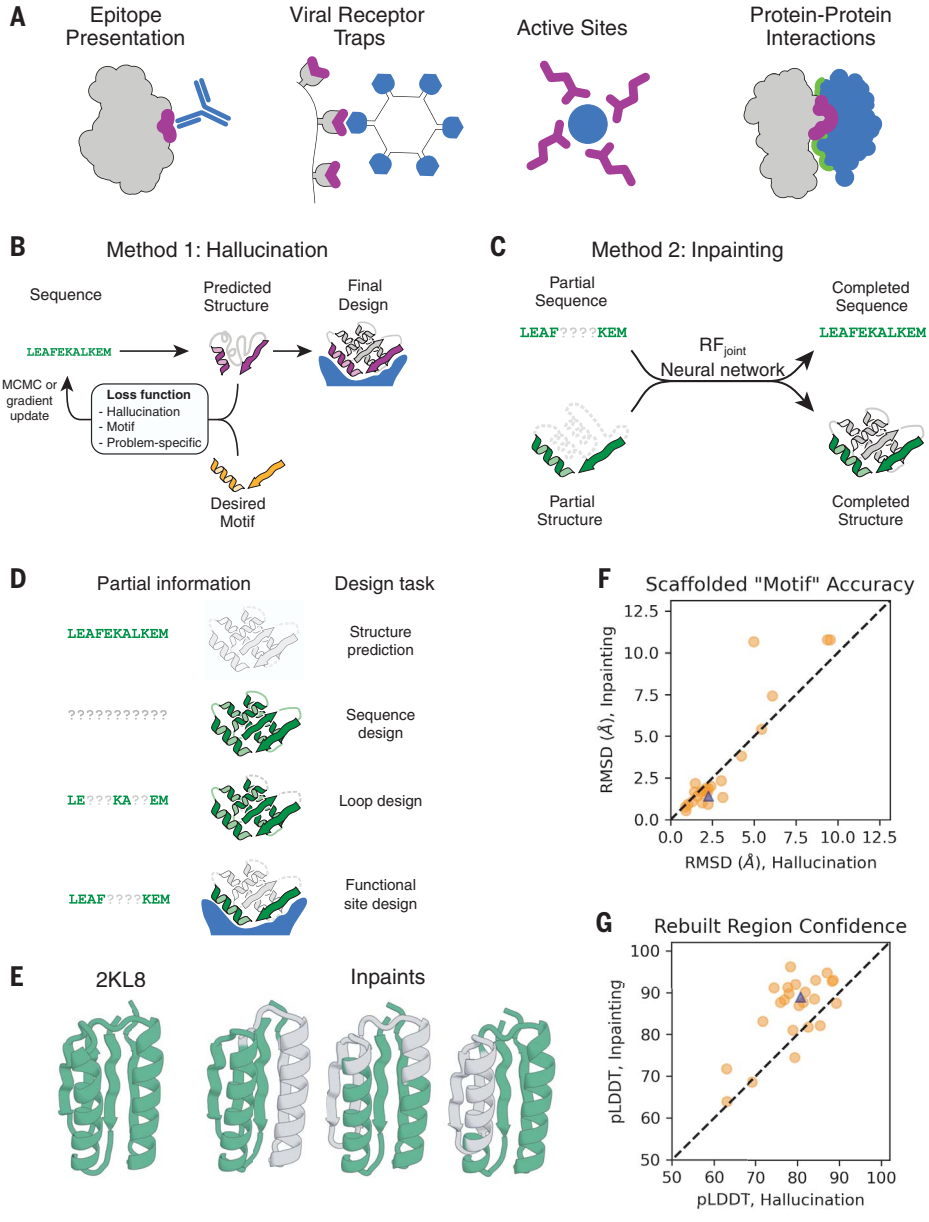

Fig. 1. Methods for protein function design. (**A**) Applications of functional-site scaffolding. (**B** and **C**) Design methods. (B) Constrained hallucination. At each iteration, a sequence is passed to the trRosetta or RoseTTAFold neural network, which predicts 3D coordinates and inter-residue distances 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 RF_{joint}), 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**) RF_{joint} can simultaneously recover structure and sequence of a masked protein region. 2KL8 was fed into RF_{joint} 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 RF_{joint} 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 RF_{joint} and hallucination were tasked with filling in the missing sequence and structure to "scaffold" the unmasked "motif." For this mask length, RF_{joint} 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.

structure in a masked region of protein (Fig. 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 (*17*) or the completion of corrupted images using inpainting (*18*). A wide variety of protein structure prediction and design challenges can be similarly formulated as missing information recovery problems (Fig. 1D). Although protein inpainting has been explored before (*19*, *20*), 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 (*16*) 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. S3; materials and methods). This model, denoted $RF_{implicit}$, was able to recover small, contiguous regions missing both sequence and structure (fig. 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. S4A; materials and methods and algorithm S1). The resulting model was able to inpaint missing regions with high fidelity (Fig. 1E and fig. S4) and performed well at sequence design (32% native sequence recovery during training) and structure prediction (fig. S4C). We call this network RF_{joint} 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 (*21*), which has high accuracy on de novo designed proteins (*22*) (fig. 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 (*13*, *23*). We used AF to compare the ability of hallucination and

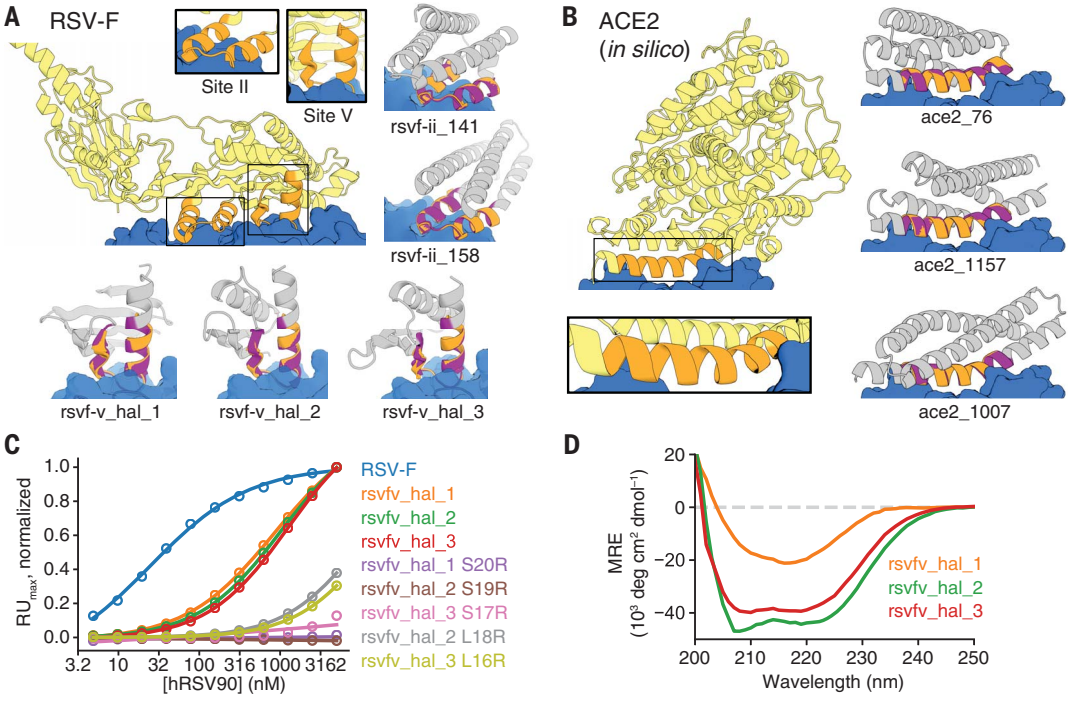

Fig. 2. Design of epitope scaffolds and receptor traps. (**A**) Design of proteins scaffolding immunogenic epitopes on RSV protein F (site II: PDB ID 3IXT chain P residues 254 to 277; site V: PDB ID 5TPN chain A residues 163 to 181). Comparisons of the RF hallucinated models to AF2 structure predictions 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.

inpainting to rebuild missing protein regions (Fig. 1, F and G, and fig. S5). Inpainting yielded solutions with more accurately predicted fixed regions ("AF-RMSD"; Fig. 1G and fig. S5B) and structures overall more confidently predicted from their amino acid sequences ("AF pLDDT"; Fig. 1F and fig. 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. S5) and generated greater structural diversity (fig. 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 to 5 and table 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 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 (*24*) or mutational effects (*25*, *26*)]. More critically, we assessed the activities of the designs experimentally (with the exception of those labeled "in silico" in Figs. 2 to 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. 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 (*7*, *9*, *10*). We scaffolded RSV-F site II, a 24-residue helix-loop-helix motif that had previously been grafted successfully onto a three-helix bundle (*7*), as well as RSV-F site V, a 19-residue helix-loop-strand motif that has not yet been scaffolded successfully (*27*). We were able to hallucinate designs recapitulating both epitopes to sub-angstrom backbone RMSD in a variety of folds [Fig. 2A and fig. 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. S9, S11, and 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. 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 (*27*) with a dissociation constant (K_d) of 0.9 to 1.3 μM (Fig. 2C and fig. 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 (*27*). 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. 2C and fig. S11A), and circular dichroism (CD) spectra were consistent with the designed scaffold structures for both the original hallucinations (Fig. 2D) and the epitope mutants (fig. S11C). Four of the inpainted designs bound hRSV90 by yeast display but were poorly expressed in *E. coli*

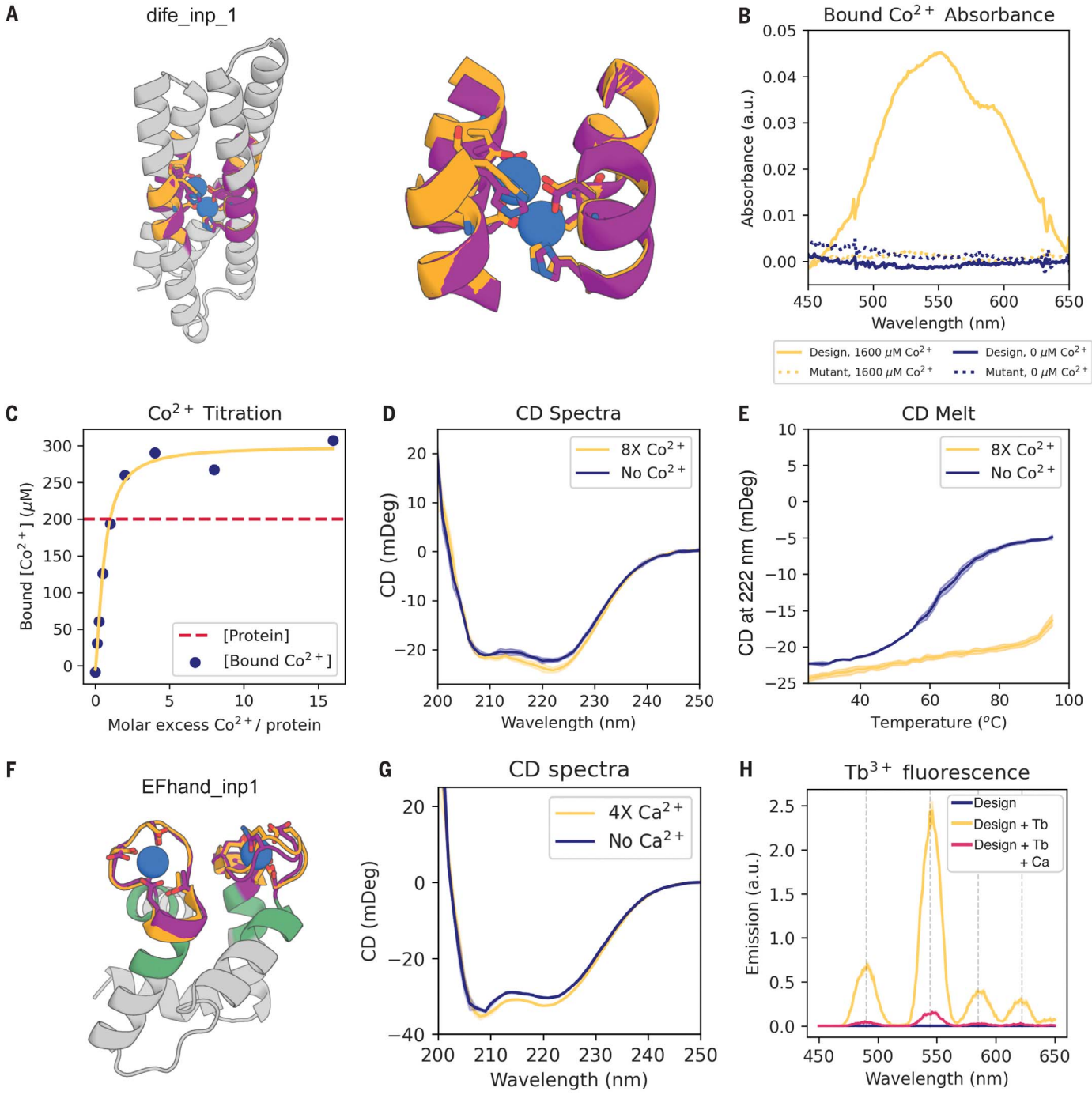

Fig. 3. Design of metal binding. (**A**) Scaffolding of di-iron binding site from *E. coli* cytochrome b1 (PDB ID 1BCF chain A residues 18 to 25, 27 to 54, 94 to 97, and 123 to 130) using inpainting. Colors: native protein scaffold, light yellow; native functional motif, orange; hallucinated scaffold, 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 Co^{2+}. Peaks at 520, 555, and 600 nm, consistent with Co^{2+} 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 Co^{2+} titration (protein concentration: 200 μM). Quantification of the absorbance at 550 nm, using a predicted extinction coefficient of 155 for Co^{2+} 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 Co^{2+} are both consistent with the predicted helical structure. (**E**) Temperature dependence of dife_inp_1 CD signal in the presence and absence of Co^{2+}. Coordination of Co^{2+} in the core stabilizes the protein. Protein concentration: 6.7 μM; Co^{2+} 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 $CaCl_2$ 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 $CaCl_2$ (red). Design metrics (AF pLDDT, motif AF-RMSD): dife_inp_1 (92, 0.65 Å) and EFhand_inp1 (84, 0.7 Å).

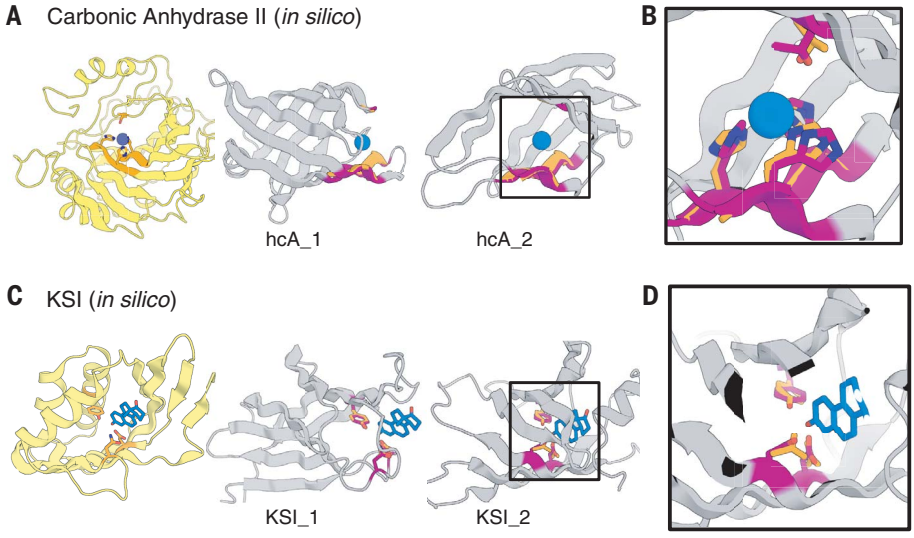

Fig. 4. In silico design of enzyme active sites. (**A** and **B**) Hallucinations using backbone description of site using RF. (**C** and **D**) Hallucination using side-chain description of site using AF2 augmented with trRosetta (materials and methods). (A) Carbonic anhydrase II active site (PDB ID 5YUI chain A residues 62 to 65, 93 to 97, and 118 to 120). (B) Δ^5-3-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).

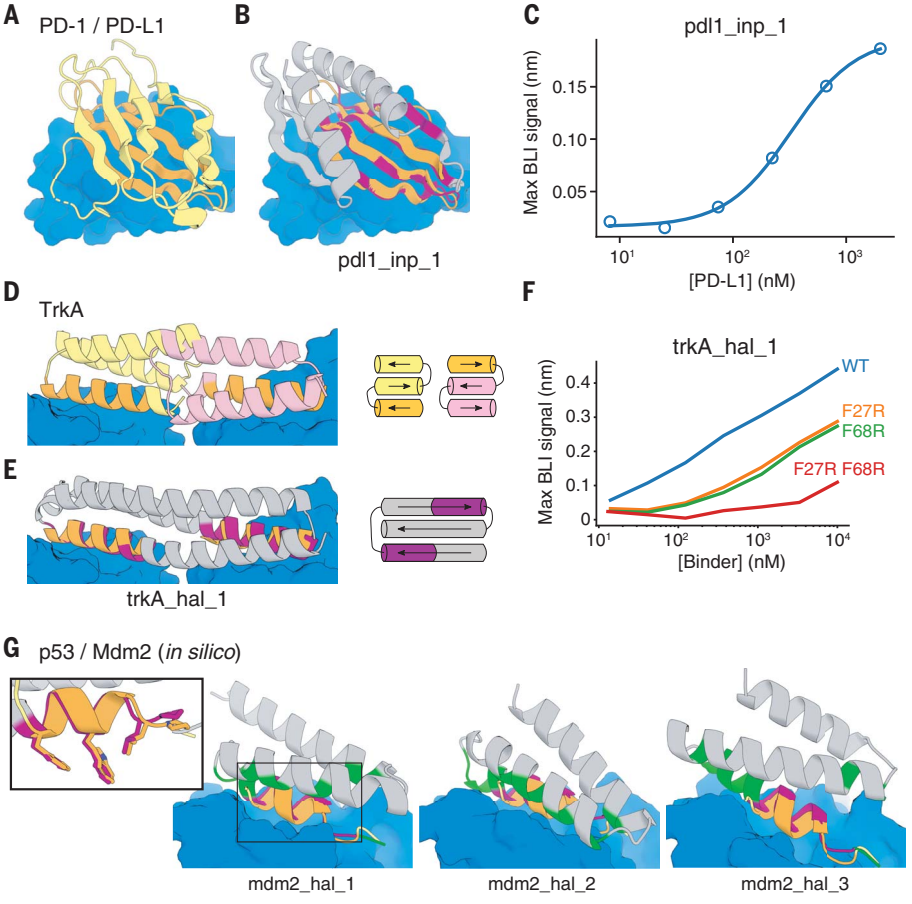

Fig. 5. Design of protein-binding proteins. Designs containing target-binding interfaces built around native-complex–derived binding motifs. Targets are in blue, native scaffolds in yellow or pink, native motifs in orange, designed scaffolds in gray, and designed motifs in purple. (**A**) Crystal structure 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.

(fig. 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 (*28*). The hallucinated hACE2 mimetics have a diverse set of helical topologies, and AF structure predictions recapitulate the binding interface with sub-angstrom accuracy (Fig. 2B and fig. S9C).

Designing metal-coordinating proteins

Di-iron sites are important in biological systems for iron storage (*29*) and can mediate catalysis (*30*, *31*). 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. 3, A to E, and fig. S13) and hallucination (fig. 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. S13B and 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 Co^{2+} binding (a proxy for iron binding) (*32*, *33*), and three (dife_inp_1, dife_inp_2, and dife_inp_3; Fig. 3B and fig. S13E) had CD spectra consistent with the designed fold (Fig. 3D and fig. S13F) and were stabilized by metal binding (Fig. 3E and fig. S13G). Mutation of the metal binding residues abolished binding (Fig. 3B and fig. S13E), and titration analysis of dife_inp_1 suggested that both metal binding sites were successfully scaffolded (Fig. 3C).

We next scaffolded the calcium-binding EF-hand motif (*34*), 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. 3F; fig. 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 (*35*). Six hallucinations and four inpaintings had fluorescence consistent with ion binding [fig. S14A; materials and methods; one of these proteins (EFhand_inp_2) was designed using $RF_{implicit}$ (supplementary text)]. The top hit from yeast, the inpainted EFhand_inp_1, purified from *E. coli* as a monomer (fig. S14C), had the expected CD spectrum (Fig. 3G) and a clear terbium binding signal (Fig. 3H) that was eliminated by $CaCl_2$ competition (Fig. 3H).

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 (*31*–*33*). The active site consists of three Zn^{2+}-coordinating histidines on two strands and a threonine on a loop, which orients the CO_2 (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 Zn^{2+} coordination (Fig. 4A and fig. 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 Δ^5-3-ketosteroid isomerase (KSI) (table S1) involved in steroid hormone biosynthesis (*36*). 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. 4, C and D, and fig. 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 challenges without network retraining.

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. 5A; materials and methods) (*15*). 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. 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) (*37*) but better than wild-type PD-1 (K_d = 3.9 μM) (*37*). The pdl1_inp_1 design expressed as a monomer (fig. S15E), was thermostable, and had a CD spectrum consistent with that of a mixed α-β fold (fig. 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. S15, A and B). The closest Protein Data Bank (PDB) (*38*) 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 (*39*), and starting from the TrkA-NGF crystal structure, we positioned helical segments derived from two copies of a previously designed TrkA binding protein (*4*) and used hallucination followed by inpainting (materials and methods) to scaffold them on a single chain (Fig. 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. 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. 5F and fig. S16), suggesting that the protein binds two molecules of TrkA, as designed.

RoseTTAFold is able to predict the structures of protein complexes (*40*), 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. 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. S17, D and E); three examples are shown in Fig. 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. S17H). Most of the hallucinated peptides bound at native protein interaction sites (fig. S18A); the remainder bound in hydrophobic grooves resembling protein binding sites (fig. 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. 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. S20 and

table S3; supplementary text (*41*), 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. 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.

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 (*42*–*49*), the design of protein structure is relatively underexplored, likely because of the difficulty of efficiently representing and learning structure (*50*). Generative adversarial networks and variational autoencoders have been used to generate protein backbones for specific fold families (*51*–*53*), 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 (*54*–*56*) 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.

REFERENCES AND NOTES

1. D. Röthlisberger *et al*., *Nature* **453**, 190–195 (2008).
2. J. B. Siegel *et al*., *Science* **319**, 1387–1391 (2008).
3. J. B. Siegel *et al*., *Science* **329**, 309–313 (2010).
4. L. Cao *et al*., *Nature* **605**, 551–560 (2022).
5. A. Chevalier *et al*., *Nature* **550**, 74–79 (2017).
6. E. Procko *et al*., *Cell* **157**, 1644–1656 (2014).
7. B. E. Correia *et al*., *Nature* **507**, 201–206 (2014).
8. D.-A. Silva *et al*., *Nature* **565**, 186–191 (2019).
9. F. Sesterhenn *et al*., *Science* **368**, eaay5051 (2020).
10. C. Yang *et al*., *Nat. Chem. Biol.* **17**, 492–500 (2021).
11. J. Yang *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **117**, 1496–1503 (2020).
12. I. Anishchenko *et al*., *Nature* **600**, 547–552 (2021).
13. C. Norn *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2017228118 (2021).
14. D. Tischer *et al*., bioRxiv 2020.11.29.402743 [Preprint] (2020); https://doi.org/10.1101/2020.11.29.402743.
15. R. Pascolutti *et al*., *Structure* **24**, 1719–1728 (2016).
16. M. Baek *et al*., *Science* **373**, 871–876 (2021).
17. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, arXiv:1810.04805 [cs.CL] (2019).
18. R. A. Yeh *et al*., arXiv:1607.07539 [cs.CV] (2017).
19. Z. Li, S. P. Nguyen, D. Xu, Y. Shang, *Proc. Int. Conf. Tools. Artif. Intell.* **29**, 1085–1091 (2017).
20. N. Anand, P. Huang, in *Advances in Neural Information Processing Systems 31*, S. Bengio *et al*., Eds. (Curran Associates, Inc., 2018), pp. 7494–7505.
21. J. Jumper *et al*., *Nature* **596**, 583–589 (2021).
22. R. Chowdhury *et al*., bioRxiv 2021.08.02.454840 [Preprint] (2021); https://doi.org/10.1101/2021.08.02.454840.
23. K. T. Simons, R. Bonneau, I. Ruczinski, D. Baker, *Proteins* **37** (suppl. 3), 171–176 (1999).
24. T.-E. Kim *et al*., bioRxiv 2021.12.17.472837 [Preprint] (2021); https://doi.org/10.1101/2021.12.17.472837.
25. M. A. Pak *et al*., bioRxiv 2021.09.19.460937 [Preprint] (2021); https://doi.org/10.1101/2021.09.19.460937.
26. G. R. Buel, K. J. Walters, *Nat. Struct. Mol. Biol.* **29**, 1–2 (2022).
27. J. J. Mousa, N. Kose, P. Matta, P. Gilchuk, J. E. Crowe Jr., *Nat. Microbiol.* **2**, 16271 (2017).
28. T. W. Linsky *et al*., *Science* **370**, 1208–1214 (2020).
29. F. Frolow, A. J. Kalb, J. Yariv, *Nat. Struct. Biol.* **1**, 453–460 (1994).
30. A. Lombardi, F. Pirro, O. Maglio, M. Chino, W. F. DeGrado, *Acc. Chem. Res.* **52**, 1148–1159 (2019).
31. J. R. Calhoun *et al*., *Biopolymers* **80**, 264–278 (2005).
32. A. M. Keech *et al*., *J. Biol. Chem.* **272**, 422–429 (1997).
33. E. N. G. Marsh, W. F. DeGrado, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 5150–5154 (2002).
34. M. Yáñez, J. Gil-Longo, M. Campos-Toimil, in *Calcium Signaling*, Md. S. Islam, Ed., vol. 740 of *Advances in Experimental Medicine and Biology* (Springer Netherlands, 2012), pp. 461–482.
35. S. J. Caldwell *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30362–30369 (2020).
36. H.-S. Cho *et al*., *J. Biol. Chem.* **274**, 32863–32868 (1999).
37. R. L. Maute *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **112**, E6506–E6514 (2015).
38. H. M. Berman *et al*., *Nucleic Acids Res.* **28**, 235–242 (2000).
39. C. Wiesmann, M. H. Ultsch, S. H. Bass, A. M. de Vos, *Nature* **401**, 184–188 (1999).
40. I. R. Humphreys *et al*., *Science* **374**, eabm4805 (2021).
41. K. Tunyasuvunakool *et al*., *Nature* **596**, 590–596 (2021).
42. J. Ingraham, V. K. Garg, R. Barzilay, T. Jaakkola, "Generative models for graph-based protein design," 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, 8 to 14 December 2019.
43. A. Strokach, D. Becerra, C. Corbi-Verge, A. Perez-Riba, P. M. Kim, *Cell Syst.* **11**, 402–411.e4 (2020).
44. S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, G. M. Church, *Nat. Methods* **18**, 389–396 (2021).
45. D. Repecka *et al*., *Nat. Mach. Intell.* **3**, 324–333 (2021).
46. J.-E. Shin *et al*., *Nat. Commun.* **12**, 2403 (2021).
47. Z. Wu, K. E. Johnston, F. H. Arnold, K. K. Yang, *Curr. Opin. Chem. Biol.* **65**, 18–27 (2021).
48. N. Anand *et al*., *Nat. Commun.* **13**, 746 (2022).
49. A. Madani *et al*., bioRxiv 2021.07.18.452833 [Preprint] (2021); https://doi.org/10.1101/2021.07.18.452833.
50. S. Ovchinnikov, P.-S. Huang, *Curr. Opin. Chem. Biol.* **65**, 136–144 (2021).
51. N. Anand, R. Eguchi, P.-S. Huang, "Fully differentiable full-atom protein backbone generation," Seventh International Conference on Learning Representations (ICLR 2019), New Orleans, Louisiana, 6 to 9 May 2019.
52. R. R. Eguchi, C. A. Choe, P.-S. Huang, *PLOS Comput. Biol.* **18**, e1010271 (2022).
53. Z. Lin, T. Sercu, Y. LeCun, A. Rives, "Deep generative models create new and diverse protein structures," 35th Conference on Neural Information Processing Systems (NeurIPS 2021), 6 to 14 December 2021.
54. M. Jendrusch, J. O. Korbel, S. K. Sadiq, bioRxiv 2021.10.11.463937 [Preprint] (2021); https://doi.org/10.1101/2021.10.11.463937.
55. L. Moffat, J. G. Greener, D. T. Jones, bioRxiv 2021.08.24.457549 [Preprint] (2021); https://doi.org/10.1101/2021.08.24.457549.
56. L. Moffat, S. M. Kandathil, D. T. Jones, bioRxiv 2022.01.27.478087 [Preprint] (2022); https://doi.org/10.1101/2022.01.27.478087.
57. L. Li *et al*., *J. Phys. Chem. C* **112**, 12219–12224 (2008).
58. J. Wang *et al*., RFDesign: Protein hallucination and inpainting with RosettaFold, version 2, Zenodo (2022); https://doi.org/10.5281/zenodo.6808038.

ACKNOWLEDGMENTS

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. **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.). **Author contributions:** Designed the research: J.W., S.L., D.J., D.T., J.L.W., S.O., and D.B. Developed the motif-constrained hallucination method: J.W., D.T., S.L., I.A., and S.O. Contributed code and ideas for hallucination: M.B. and J.D. Generated designs using hallucination: J.W., S.L., D.T., and S.O. Developed the inpainting method: D.J. and J.L.W. Contributed code and ideas for inpainting: M.B., J.W., S.L., and D.T. Generated designs using inpainting: D.J., J.L.W., and A.S. Analyzed data: J.W., S.L., D.J., D.T., J.L.W., and M.E. Trained neural networks: D.J., J.L.W., and M.B. Performed RSV-F experiments: K.M.C., R.R., L.F.M., and J.W. Performed di-iron experiments: J.L.W. and D.J. Performed EF-hand experiments: A.S. and J.L.W.

Performed PD-L1 experiments: W.Y., D.R.H., J.W., S.L., and D.J. Contributed reagents and technical expertise: T.S., J.-H.C., L.F.M., N.B., B.I.M.W., B.C., A.M., and F.D. Wrote the manuscript: J.W., D.J., J.L.W., S.L., D.T., S.O., and D.B. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Code and neural network weights are available at https://github.com/RosettaCommons/RFDesign and https://github.com/sokrypton/ColabDesign and archived at Zenodo (*58*). Plasmids of designed proteins are available upon request. **License information:**

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abn2100
Materials and Methods
Supplementary Text
Figs. S1 to S21
Tables S1 to S5
Algorithm S1
References (*59*–*86*)
MDAR Reproducibility Checklist
Data S1 and S2

View/request a protocol for this paper from *Bio-protocol*.

Submitted 11 November 2021; accepted 24 June 2022
10.1126/science.abn2100

Scaffolding protein functional sites using deep learning

Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L. Watson, Karla M. Castro, Robert Ragotte, Amijai Saragovi, Lukas F. Milles, Minkyung Baek, Ivan Anishchenko, Wei Yang, Derrick R. Hicks, Marc Expòsit, Thomas Schlichthaerle, Jung-Ho Chun, Justas Dauparas, Nathaniel Bennett, Basile I. M. Wicky, Andrew Muenks, Frank DiMaio, Bruno Correia, Sergey Ovchinnikov, and David Baker

Science **377** (6604), . DOI: 10.1126/science.abn2100

Designing around function

Protein design has had success in finding sequences that fold into a desired conformation, but designing functional proteins remains challenging. Wang *et al*. describe two deep-learning methods to design proteins that contain prespecified functional sites. In the first, they found sequences predicted to fold into stable structures that contain the functional site. In the second, they retrained a structure prediction network to recover the sequence and full structure of a protein given only the functional site. The authors demonstrate their methods by designing proteins containing a variety of functional motifs. —VV

Science (ISSN 1095-9203) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.